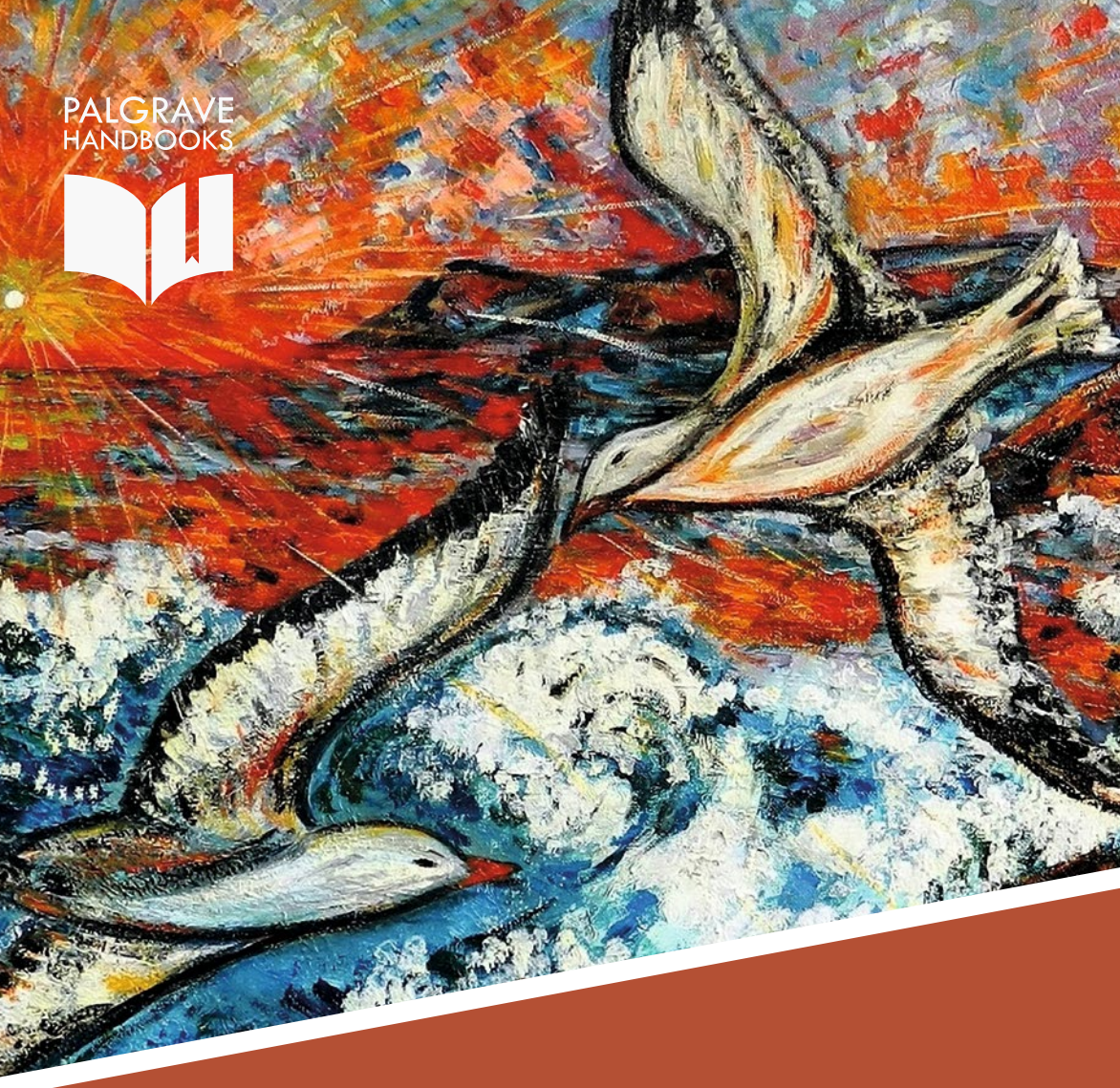


PALGRAVE  
HANDBOOKS



# THE PALGRAVE HANDBOOK OF APPLIED LINGUISTICS RESEARCH METHODOLOGY

Edited by Aek Phakiti, Peter De Costa,  
Luke Plonsky and Sue Starfield



The Palgrave Handbook of Applied Linguistics  
Research Methodology

Aek Phakiti • Peter De Costa  
Luke Plonsky • Sue Starfield  
Editors

# The Palgrave Handbook of Applied Linguistics Research Methodology

palgrave  
macmillan

*Editors*

Aek Phakiti  
Sydney School of Education and Social Work  
University of Sydney  
Sydney, NSW, Australia

Luke Plonsky  
Applied Linguistics  
Northern Arizona University  
Flagstaff, AZ, USA

Peter De Costa  
Department of Linguistics, Germanic, Slavic,  
Asian and African Languages  
Michigan State University  
East Lansing, MI, USA

Sue Starfield  
School of Education  
UNSW Sydney  
Sydney, NSW, Australia

ISBN 978-1-137-59899-8      ISBN 978-1-137-59900-1 (eBook)  
<https://doi.org/10.1057/978-1-137-59900-1>

Library of Congress Control Number: 2018955931

© The Editor(s) (if applicable) and The Author(s) 2018

The author(s) has/have asserted their right(s) to be identified as the author(s) of this work in accordance with the Copyright, Designs and Patents Act 1988.

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover illustration: © Aek Phakiti

This Palgrave Macmillan imprint is published by the registered company Springer Nature Limited.  
The registered company address is: The Campus, 4 Crinan Street, London, N1 9XW, United Kingdom



# Preface

Applied linguistics is a broad, evolving interdisciplinary field of study, which examines language use with relevance to real-world problems across a range of social contexts using a diverse set of methodologies. This *Handbook* aims to provide a comprehensive, yet accessible treatment of basic and more advanced research methodologies in applied linguistics as well as to offer a state-of-the-art view of various substantive domains within the field. The *Handbook* covers a range of research approaches, presents current perspectives, and addresses important considerations in different research methods, such as designing and implementing research instruments and techniques and analyzing different types of applied linguistics data. Innovations, challenges, and trends in applied linguistics research are addressed throughout the *Handbook*.

This *Handbook* has brought together a range of authors in various areas of research into one volume. The authors work with a variety of languages in a host of research contexts, ensuring both breadth and depth. As the *Handbook* editors, we have curated themes and ideas that are aligned with the current research climate as well as areas that help applied linguists better understand social and educational phenomena and the nature of language, language learning, and language use.

## Readership

As we anticipate that many readers of this *Handbook* may be junior scholars seeking guidance on research methods, and taking into account the many options and pathways on offer, we have striven to ensure that the *Handbook* provides an up-to-date entry point into both approaches that have stood the test

of time and approaches that may be less well known, but offer interesting possibilities and perspectives. This *Handbook* is suitable for use by advanced undergraduate and postgraduate students as well as beginning and well-established applied linguists who would like both a broad and in-depth understanding of contemporary applied linguistics research methods and topics. Specifically, this *Handbook* can be used in applied linguistics, second language studies, and TESOL graduate programs around the world. Libraries, university departments, and organizations dealing with applied linguistics issues will also find this *Handbook* to be an invaluable resource.

## Comments or Suggestions

The editors would be grateful to hear comments and suggestions regarding this *Handbook*. Please contact Aek Phakiti at [aek.phakiti@sydney.edu.au](mailto:aek.phakiti@sydney.edu.au), Peter De Costa at [peteridecosta@gmail.com](mailto:peteridecosta@gmail.com), Luke Plonsky at [lukeplonsky@gmail.com](mailto:lukeplonsky@gmail.com), or Sue Starfield at [s.starfield@unsw.edu.au](mailto:s.starfield@unsw.edu.au).

Sydney, NSW, Australia  
East Lansing, MI, USA  
Flagstaff, AZ, USA  
Sydney, NSW, Australia

Aek Phakiti  
Peter De Costa  
Luke Plonsky  
Sue Starfield

# Acknowledgments

We wish to express our heartfelt thanks to the contributors of this *Handbook* who worked hard to produce great chapters and promptly responded to our requests and comments on earlier drafts. You are all truly amazing. We would also like to thank the many researchers, authors, and methodologists who published research articles, book chapters, and books not only in applied linguistics but across various disciplines. Their contributions have helped us deepen our understanding of numerous issues and methods relevant to applied linguistics. Next, we are very grateful to Palgrave for their kind support throughout the completion of this *Handbook*.

We would like to thank our colleagues and friends at Georgetown University, Michigan State University, Northern Arizona University, the University of New South Wales, and the University of Sydney who discussed with us essential ideas and issues to be included in this *Handbook* and read and commented on several chapter drafts, in particular: Janette Bobis, Jesse Egbert, Mia Jun, Amy Kim, Wendy Li, Alison Mackey, Guy Middleton, Lourdes Ortega, Brian Paltridge, Jack C. Richards, Fran Waugh, and Yiran Xu.

Finally, we would like to thank our partners who supported us throughout the process of putting together this *Handbook*, including weekend Skype time when our schedules—sometimes across four distinct time zones—would align.

# Contents

<b>Part I</b>	<b>Research Approaches and Methodology</b>	<b>1</b>
<b>1</b>	<b>Applied Linguistics Research: Current Issues, Methods, and Trends</b>	<b>5</b>
	<i>Aek Phakiti, Peter De Costa, Luke Plonsky, and Sue Starfield</i>	
<b>2</b>	<b>Habits of Mind: How Do We Know What We Know?</b>	<b>31</b>
	<i>Richard F. Young</i>	
<b>3</b>	<b>Quantitative Methodology</b>	<b>55</b>
	<i>Luke K. Fryer, Jenifer Larson-Hall, and Jeffrey Stewart</i>	
<b>4</b>	<b>Qualitative Methodology</b>	<b>79</b>
	<i>Shim Lew, Anna Her Yang, and Linda Harklau</i>	
<b>5</b>	<b>Mixed Methodology</b>	<b>103</b>
	<i>Alison Mackey and Lara Bryfonski</i>	
<b>6</b>	<b>Traditional Literature Review and Research Synthesis</b>	<b>123</b>
	<i>Shaofeng Li and Hong Wang</i>	
<b>7</b>	<b>Research Replication</b>	<b>145</b>
	<i>Rebekha Abbuhl</i>	

<b>8</b>	<b>Ethical Applied Linguistics Research</b> <i>Scott Sterling and Peter De Costa</i>	163
<b>9</b>	<b>Writing a Research Proposal</b> <i>Sue Starfield</i>	183
<b>10</b>	<b>Writing a Research Article</b> <i>Betty Samraj</i>	199
<b>Part II Research Instruments, Techniques, and Data Sources</b>		221
<b>11</b>	<b>Interviews and Focus Groups</b> <i>Matthew T. Prior</i>	225
<b>12</b>	<b>Observation and Fieldnotes</b> <i>Fiona Copland</i>	249
<b>13</b>	<b>Online Questionnaires</b> <i>Jean-Marc Dewaele</i>	269
<b>14</b>	<b>Psycholinguistic Methods</b> <i>Sarah Grey and Kaitlyn M. Tagarelli</i>	287
<b>15</b>	<b>SLA Elicitation Tasks</b> <i>Susan Gass</i>	313
<b>16</b>	<b>Introspective Verbal Reports: Think-Alouds and Stimulated Recall</b> <i>Melissa A. Bowles</i>	339
<b>17</b>	<b>Corpus Research Methods for Language Teaching and Learning</b> <i>Magali Paquot</i>	359
<b>18</b>	<b>Digital Discourses Research and Methods</b> <i>Christoph A. Hafner</i>	375



<b>Part III</b>	<b>Data Analysis</b>	391
<b>19</b>	<b>Correlation and Simple Linear Regression in Applied Linguistics</b> <i>Reza Norouzian and Luke Plonsky</i>	395
<b>20</b>	<b>Exploratory Factor Analysis</b> <i>Aek Phakiti</i>	423
<b>21</b>	<b>Confirmatory Factor Analysis and Structural Equation Modeling</b> <i>Aek Phakiti</i>	459
<b>22</b>	<b>Analyzing Group Differences</b> <i>Luke Wander Amoroso</i>	501
<b>23</b>	<b>Statistics for Categorical, Nonparametric, and Distribution-Free Data</b> <i>Jesse Egbert and Geoffrey T. LaFlair</i>	523
<b>24</b>	<b>Reliability Analysis of Instruments and Data Coding</b> <i>Kirby C. Grabowski and Saerhim Oh</i>	541
<b>25</b>	<b>Analyzing Spoken and Written Discourse: A Role for Natural Language Processing Tools</b> <i>Scott A. Crossley and Kristopher Kyle</i>	567
<b>26</b>	<b>Narrative Analysis</b> <i>Phil Benson</i>	595
<b>27</b>	<b>Interaction Analysis</b> <i>Elizabeth R. Miller</i>	615
<b>28</b>	<b>Multimodal Analysis</b> <i>Jesse Pirini, Tui Matelau-Doherty, and Sigrid Norris</i>	639

<b>Part IV</b>	<b>Selected Research Topics and Areas in Applied Linguistics</b>	659
<b>29</b>	<b>Instructed Second Language Acquisition</b> <i>Shawn Loewen</i>	663
<b>30</b>	<b>Bilingualism and Multilingualism</b> <i>Tej K. Bhatia</i>	681
<b>31</b>	<b>Forensic Linguistics</b> <i>Samuel Lerner</i>	703
<b>32</b>	<b>World Englishes</b> <i>Peter De Costa, Jeffrey Maloney, and Dustin Crowther</i>	719
<b>33</b>	<b>Heritage, Community, and Indigenous Languages</b> <i>Shereen Bhalla and Terrence G. Wiley</i>	741
<b>34</b>	<b>Translation and Interpreting</b> <i>Claudia V. Angelelli</i>	761
<b>35</b>	<b>Identity</b> <i>Ron Darwin</i>	777
<b>36</b>	<b>Gesture Research</b> <i>Gale Stam and Kimberly Buescher</i>	793
<b>37</b>	<b>Language Policy and Planning</b> <i>David Cassels Johnson and Crissa Stephens</i>	811
<b>38</b>	<b>Second Language Pragmatics</b> <i>Soo Jung Youn</i>	829
<b>39</b>	<b>Language Testing and Assessment</b> <i>April Ginther and Kyle McIntosh</i>	845

	Contents	xiii
<b>40 Linguistic Landscape</b> <i>David Malinowski</i>		869
<b>41 Researching Academic Literacies</b> <i>David Bloome, Gilcinei T. Carvalho, and Sanghee Ryu</i>		887
<b>Index</b>		903

## Notes on Contributors

**Rebekha Abbuhl** is Associate Professor of Linguistics at California State University Long Beach, where she teaches courses in language acquisition, research methods, and pedagogy. Her research interests include second language writing and the role of feedback in the development of foreign language proficiency.

**Luke Wander Amoroso** is Assistant Professor of Linguistics at Truman State University. His research interests are in the validity and reliability of L2 tests, speaking proficiency, and second language acquisition (SLA) research methodology. He tries to keep his features in order and works as a language testing consultant with the United States Department of Justice. He works with ESL and EFL teachers in the United States and China to incorporate insights from SLA interaction research into ESL/EFL teaching methods.

**Claudia V. Angelelli** is Chair of Multilingualism and Communication at Heriot-Watt University, UK, and Professor Emerita of Spanish Linguistics at San Diego State University, US. Her research lies at the intersection of sociolinguistics, applied linguistics, and translation and interpreting studies. She has authored *Medical Interpreting and Cross-cultural Communication* (2004) and *Revisiting the Role of the Interpreter* (2004) and co-edited *Researching Translation and Interpreting* (2015) and *Testing and Assessment in Translation and Interpreting Studies* (2009). Her work has appeared in *The Annual Review of Applied Linguistics*, *The Critical Link*, *Cuadernos de ALDEEU*, *International Journal of the Sociology of Language (IJSL)*, *Interpreting, Meta*, *MonTI*, *The Translator*, *Translation and Interpreting Studies (TIS)*, and numerous edited volumes.

**Phil Benson** is Professor of Applied Linguistics at Macquarie University, Sydney, Australia. His research interests are in autonomy and out-of-class language learning, study abroad, and multilingualism. He has a strong preference for qualitative research and has published on both qualitative research methods and narrative inquiry. He is especially interested in oral history as an approach to research on the long-term lan-

guage learning experiences of multilingual individuals. He is co-author of *Second Language Identity in Narratives of Study Abroad* (Palgrave Macmillan, 2012) and *Narrative Inquiry in Language Teaching and Learning Research* (2013).

**Shereen Bhalla** is a research associate and the facilitator of the Language Policy Research Network and serves as Manager of Online Education at the Center for Applied Linguistics (CAL). At CAL, Bhalla conducts research, co-authors papers, and regularly presents at national and international conferences on issues regarding language policy, heritage language learning, and English as an international language. She has experience teaching and working with pre-service and in-service teachers in the areas of culturally responsive teaching, second language acquisition, writing development and oral communication. She received her PhD in Culture, Literacy and Language from the University of Texas at San Antonio.

**Tej K. Bhatia** is Professor of Linguistics and Director of South Asian languages at Syracuse University, Syracuse, New York. He has been Director of the Linguistic Studies Program and Acting Director of Cognitive Sciences. He is also a Faculty Fellow at the Forensic & National Security Sciences Institute. He is Editor-in-Chief of *Brill Research Perspectives on Multilingualism and Second Language Acquisition*. His publications include five handbooks with William C. Ritchie: *Handbook of Bilingualism and Multilingualism* (2013), *A New Handbook of Second Language Acquisition* (2009), *The Handbook of Bilingualism* (2006), *Handbook of Child Language Acquisition* (1999), and *Handbook of Second Language Acquisition* (1996).

**David Bloome** is College of Education and Human Ecology (EHE) Distinguished Professor of Teaching and Learning at The Ohio State University. David's research focuses on how people use spoken and written language for learning, teaching, social relationships, constructing knowledge, and shared histories. He is former president of the National Council of Teachers of English and of the National Conference on Research in Language and Literacy; former co-editor of *Reading Research Quarterly*; and founding editor of *Linguistics and Education*. David was inducted into the Reading Hall of Fame in 2008 and in 2015 he received the John. J. Gumperz Lifetime Achievement Award.

**Melissa A. Bowles** is an associate professor in the Department of Spanish and Portuguese and Director of the Second Language Acquisition and Teacher Education PhD concentration at the University of Illinois at Urbana-Champaign. Her main research interests are classroom second and heritage language acquisition and the ways in which instruction differentially affects the two learner groups. She routinely uses verbal reports in her research and has written about them extensively, most notably in *The Think-Aloud Controversy in Second Language Research* (2010).

**Lara Bryfonski** is a doctoral candidate in applied linguistics at Georgetown University. Her research focuses primarily on interaction and corrective feedback in second language acquisition, as well as task-based language teaching and learning. Lara is also a licensed English as a Second Language (ESL) teacher and has taught ESL in a variety of contexts in the U.S. and abroad.



**Kimberly Buescher** is Assistant Professor of Applied Linguistics at the University of Massachusetts, Boston. Her research interests include L2 learning and teaching, L2 literacy, students and teachers' use of gesture, French prepositions, and teacher education preparation. Her dissertation "Developing Second Language Narrative Literacy Using Concept-Based Instruction and a Division-of-Labor Pedagogy" examined the extent to which concept-based instruction and a division-of-labor pedagogy promoted the development of intermediate learners' narrative literacy abilities in French. She has published book chapters on the learning and teaching of French prepositions and the internalization of talk, gesture, and concepts in the L2 classroom.

**Gilcinei T. Carvalho** is an associate professor at Universidade Federal de Minas Gerais, Brazil. He is a member of the Knowledge and Social Inclusion Graduate Program and a researcher at the Center for Literacy Studies, in the School of Education. He explores sociolinguistic approaches in the study of acquisition and the development of written language, including academic literacies. He is co-editor of *Jornal Letra A*.

**Fiona Copland** is Professor of TESOL at the University of Stirling, Scotland. She has taught English and trained teachers in a number of different countries. Her research interests include post-observation feedback in pre-service teacher education, teaching English to young learners and ethics in qualitative research. Fiona has written a book on *Linguistic Ethnography: Collecting, Analysing and Presenting Data* (2015, SAGE) with Angela Creese, as well as edited a collection entitled *Linguistic Ethnography: Interdisciplinary Explorations* (2015, Palgrave) with Julia Snell and Sara Shaw.

**Scott A. Crossley** is Associate Professor of Applied Linguistics at Georgia State University. Scott's primary research focus is on natural language processing and the application of computational tools and machine learning algorithms in language learning, writing, and text comprehensibility. His main interest area is the development and use of natural language processing tools in assessing writing quality and text difficulty. He is also interested in the development of second language learner lexicons and the potential to examine lexical growth and lexical proficiency using computational algorithms.

**Dustin Crowther** is a visiting Assistant Professor of English at Oklahoma State University, and holds a PhD in Second Language Studies from Michigan State University. He previously completed his MA in Applied Linguistics at Concordia University in Montréal, Canada. His research interests include second language pronunciation, the promotion of mutual intelligibility in multilingualistic and multicultural contact, World Englishes, and research methodologies. His research has been published in a wide range of journals, including *Studies in Second Language Acquisition*, *The Modern Language Journal*, and *TESOL Quarterly*.

**Ron Darvin** is a Vanier Scholar at the Department of Language and Literacy Education of the University of British Columbia. Together with Bonny Norton, he received the 2016 TESOL Award for Distinguished Research for their article

“Identity and a model of investment in applied linguistics” that appeared in the *Annual Review of Applied Linguistics*. Ron has also published in *TESOL Quarterly*, *Journal of Language, Identity, and Education*, and *The Routledge Handbook of Language and Identity*.

**Peter De Costa** is an associate professor in the Department of Linguistics and Languages at Michigan State University. His primary areas of research are identity and ideology in second language acquisition. He is the author of *The Power of Identity and Ideology in Language Learning* (Springer, 2016). He also edited *Ethics in Applied Linguistics Research* (2016). His work has appeared in *AILA Review*, *Applied Linguistics Review*, *International Journal of Applied Linguistics*, *Language Learning*, *Language Policy*, *Language Teaching*, *Linguistics and Education*, *Research in the Teaching of English*, *System*, *TESOL Quarterly*, and *The Modern Language Journal*. He recently guest edited special journal issues on scalar approaches to language learning and teaching (*Linguistics and Education*, 2016, with Suresh Canagarajah), teacher identity (*The Modern Language Journal*, 2017, with Bonny Norton), study abroad research methodologies (*System*, 2017, with Hima Rawal and Irina Zaykovskaya), and World Englishes and second language acquisition (*World Englishes*, 2018, with Kingsley Bolton). He is the co-editor of *TESOL Quarterly*.

**Jean-Marc Dewaele** is Professor of Applied Linguistics and Multilingualism at Birkbeck, University of London. He is interested in individual differences in foreign language acquisition and use. He won the Equality and Diversity Research Award from the British Association for Counselling and Psychotherapy (2013) and the Robert Gardner Award for Excellence in Second Language and Bilingualism Research (2016) from the International Association of Language and Social Psychology. He authored *Emotions in Multiple Languages* (second edition published in 2013 by Palgrave).

**Jesse Egbert** is an assistant professor in the Applied Linguistics program at Northern Arizona University. He specializes in corpus-based research on register variation, particularly academic writing and online language, and methodological issues in quantitative linguistic research. His research has been published in journals such as *Journal of English Linguistics*, *International Journal of Corpus Linguistics*, *Corpus Linguistics and Linguistic Theory*, and *Applied Linguistics* (2018, Routledge). His books include an edited volume titled *Triangulating Methodological Approaches in Corpus Linguistic Research* (2018, Routledge) and a book titled *Register Variation Online* (2018, Cambridge).

**Luke K. Fryer** is an associate professor and head of faculty and research postgraduate student teaching and learning programs at the University of Hong Kong. His main area of research is the role of non-cognitive factors like interest within teaching and learning. His work on interest, related motivations, and learning strategies has been published widely in journals such as *British Journal of Educational Psychology*, *Internet and Higher Education*, and *Computers and Education*. His statistical analyses focus on longitudinal structural equation modeling and person-centered analyses.

**Susan Gass** is University Distinguished Professor at Michigan State University. She has published widely in the field of second language acquisition. She serves as co-editor of *Studies in Second Language Acquisition*. She has lectured in many parts of the world, including South America, North America, Asia, Africa, and Australia. From 2002 to 2008, she was the President of the International Association of Applied Linguistics and prior to that she was the President of the American Association for Applied Linguistics. She is the recipient of numerous awards and serves as the Director of the Second Language Studies Program and the English Language Center, both at Michigan State University.

**April Ginther** is an associate professor in the Department of English at Purdue University, where she directs two language support programs. She has been an invited speaker and workshop provider at institutions and conferences around the world, presenting on her primary scholarly pursuits: the development and validation of second language proficiency assessments, the measurement of second language fluency, and the use and interpretation of language proficiency test scores by diverse groups of stakeholders. She recently stepped down as co-editor of *Language Testing*.

**Kirby C. Grabowski** is adjunct assistant professor in the Applied Linguistics and TESOL Program at Teachers College, Columbia University, where she teaches courses on second language assessment, performance assessment, generalizability theory, pragmatics assessment, research methods, linguistics, and L2 pedagogy. She is on the Editorial Advisory Board of *Language Assessment Quarterly* and formerly served on the Executive Board for ILTA as Member-at-Large. She was a Spaan Fellow for the ELI at the University of Michigan, and she received the 2011 Jacqueline Ross TOEFL Dissertation Award for outstanding doctoral dissertation in second/foreign language testing from Educational Testing Service.

**Sarah Grey** is Assistant Professor of Linguistics and Spanish at Fordham University in New York City, United States of America. She uses psycholinguistic approaches and ERPs to study adult second language acquisition and bilingualism, and her work has appeared in *The Modern Language Journal*, *Studies in Second Language Acquisition*, and the *Journal of Neurolinguistics*. She received her PhD in Applied Spanish Linguistics from Georgetown University and prior to joining Fordham University she worked as a postdoctoral research fellow in Psychology and the Center for Language Science at Pennsylvania State University.

**Christoph A. Hafner** is an associate professor in the Department of English, City University of Hong Kong. He has published widely in the areas of English for specific purposes, digital literacies, and language learning and technology. He is co-author (with Rodney H. Jones) of *Understanding Digital Literacies: A Practical Introduction* (Routledge, 2012).

**Linda Harklau** is a professor in the TESOL and World Language Education and Linguistics Program at the University of Georgia. Her research examines language

learning and academic achievement of immigrant youth in high school and college, schooling structure and educational policy, and teacher education. A recipient of the TESOL Distinguished Research Award, she also teaches and publishes on the subject of qualitative methods, particularly longitudinal case study and ethnography.

**David Cassels Johnson** is Associate Professor of Education at the University of Iowa. He holds a PhD (with distinction) in Educational Linguistics from the University of Pennsylvania. His research, teaching, and service focus on how language policies impact educational opportunities for linguistically diverse students, in both bilingual education and English language education programs. He is the author of *Language Policy* (2013, Palgrave Macmillan) and co-editor of *Research Methods in Language Policy and Planning: A Practical Guide* (2015, Wiley-Blackwell, with Francis M. Hult).

**Kristopher Kyle** is an assistant professor in the Department of Second Language Studies at the University of Hawai'i. His research interests include second language writing and speaking, language assessment, and second language acquisition. He is especially interested in applying natural language processing (NLP) and corpora to the exploration of these areas.

**Geoffrey T. LaFlair** is an assistant professor in the Department of Second Language Studies at the University of Hawai'i at Mānoa. He conducts research on large- and small-scale language assessments and quantitative research methods in the field of second language studies. His research has been published in *Language Testing*, *Applied Linguistics*, and *The Modern Language Journal*.

**Samuel Larner** is a lecturer in Linguistics at Manchester Metropolitan University, UK. His PhD thesis, completed in 2012, explored the socio- and psycholinguistic theory of formulaic sequences and their use by authors when writing short personal narratives, with the goal of identifying individual authorial consistency and distinctiveness for authorship purposes. He has published several journal articles, book chapters, and a monograph, focussing mainly on methods of forensic authorship attribution. In addition to teaching and researching forensic linguistics, Samuel undertakes consultancy in authorship analysis.

**Jenifer Larson-Hall** is an associate professor in the English Department at the University of Kitakyushu in Japan. Her research interests lie mainly in second language acquisition but she believes statistics substantially affects conclusions that are drawn in the field and has published a variety of articles and books geared toward applied researchers in second language acquisition. Her most recent book is *A Guide to Doing Statistics in Second Language Research using SPSS and R* (2016, Routledge). Her 2017 article in *The Modern Language Journal*, "Moving Beyond the Bar Plot and Line Graph to Create Informative and Attractive Graphics", argues for the importance of data-accountable graphics.

**Shim Lew** is a doctoral candidate in the Department of Language and Literacy at the University of Georgia. Her area of research is in teacher education for English learners, particularly developing content-area teachers' hybrid professional development as content and language teachers and integrating disciplinary literacy instruction into K-12 STEM classrooms.

**Shaofeng Li** is an associate professor in Foreign/Second Language Education at Florida State University where he teaches courses in second language acquisition and language pedagogy and supervises masters and PhD students. His main research interests include task-based language teaching and learning, form-focused instruction, individual learner differences (especially language aptitude and working memory), and research methods.

**Shawn Loewen** is an associate professor in Second Language Studies in the Department of Linguistics & Germanic, Slavic, Asian and African Languages at Michigan State University. His research interests include instructed second language acquisition, particularly as it pertains to learner interaction. He is also interested in research methodology and the development of statistical knowledge. He teaches a quantitative analysis class, as well as classes on second language acquisition. In addition to journal articles, he has authored *Introduction to Instructed Second Language Acquisition* (2015) and co-authored, with Luke Plonsky, *An A-Z of Applied Linguistics Research Methods* (2016, Palgrave). His co-edited volume (with Masatoshi Sato) *The Routledge Handbook of Instructed Second Language Acquisition* appeared in 2017.

**Alison Mackey** is Professor of Linguistics at Georgetown University. She is interested in interaction-driven second language (L2) learning, L2 research methodology and the applications of interaction through task-based language teaching, as well as second language dialects and identities. She is the editor of the *Annual Review of Applied Linguistics*, published by Cambridge University Press, an official journal of the *American Association for Applied Linguistics*.

**David Malinowski** is a language technology and research specialist with the Center for Language Study at Yale University. With a background in language and literacy education, multimodal communication, and technology-enhanced learning, he conducts research and supports pedagogical innovation on such technology-related topics as internet-mediated intercultural language learning (telecollaboration) and course-sharing with videoconferencing. At the same time, he maintains a significant interest in linguistic landscape, seeking to find productive intersections between urban sociolinguistics and place-based language learning. David holds a masters in TESOL from San Francisco State University and a PhD in Education from UC Berkeley.

**Jeffrey Maloney** is Assistant Professor of English at Northeastern State University. He holds a PhD in Second Language Studies from Michigan State University and an MA in Applied Linguistics from Ohio University. His research interests include lan-



guage teacher training with technology, computer-assisted language learning, and language teacher and learner identity.

**Tui Matelau-Doherty** is a PhD candidate at Auckland University of Technology in New Zealand. Her research uses Multimodal (Inter)action Analysis to explore the relationship between creative practice and ethnic identity. Her masters research examined the ethnic identity co-constructed within tertiary education environments by Māori female students. The findings of this research were published in *Interactions, Texts and Images: A Reader in Multimodality* (2014, De Gruyter). In addition, poems she wrote as part of her data collection were published in the journal *Multimodal Communication*.

**Kyle McIntosh** is an assistant professor in the Department of English and Writing at The University of Tampa, where he works primarily in the academic writing and TESOL certificate programs. His research focuses on English for Academic Purposes, intercultural rhetoric, and writing assessment. With Carolina Pelaez-Morales and Tony Silva, he co-edited the volume *Graduate Studies in Second Language Writing* (2015).

**Elizabeth R. Miller** is Associate Professor of Applied Linguistics in the English Department at the University of North Carolina at Charlotte. Her research involves adult immigrant learners of English in the U.S. and focuses on issues related to language ideologies and learners' agency and identity. Her work has appeared in a number of journals, and two of her recent publications include *The Language of Adult Immigrants: Agency in the Making* (2014) and the co-edited volume *Theorizing and Analyzing Agency in Second Language Learning: Interdisciplinary Approaches* (2015).

**Reza Norouzian** is a PhD candidate in the English as a Second Language program at Texas A&M University. In addition to his doctoral studies, Reza has also obtained a Graduate Certificate in Advanced Research Methods from Texas A&M University. Reza's research interests include instructed second language acquisition and advanced research methods. Reza has published in a number of journals including *Second Language Research* and *Issues in Applied Linguistics*. Reza is a contributor to *StackExchange* (data science, statistics, and programming forum).

**Sigrid Norris** is Professor of Multimodal (Inter)action and Director of the AUT Multimodal Research Centre at Auckland University of Technology in New Zealand. Born in Feudingingen Germany, she received her BA in Russian Language and Literature from George Washington University, and later received an MS and was conferred her PhD in Linguistics by Georgetown University in the United States. She is the founder of the theoretical/methodological framework Multimodal (Inter)action Analysis, has edited and authored numerous academic books, journal articles and book chapters, written two poetry books, and is the editor of the international journal *Multimodal Communication*.

**Saerhim Oh** is Senior Test Development Manager at Assessment Technology and Engineering at Pearson. Her research interests include linguistic tools in second lan-

guage writing assessment, feedback in second language writing, speech recognition in second language speaking assessment, and English Language Learner assessment. She received her doctorate degree in Applied Linguistics from Teachers College, Columbia University. She was the 2017 Robert Lado Memorial Award recipient in recognition of the best graduate student paper presentation at the annual meeting of Language Testing Research Colloquium (LTRC).

**Magali Paquot** is a permanent Fonds de la Recherche Scientifique (F.R.S.-FNRS) research associate at the Centre for English Corpus Linguistics, Université catholique de Louvain. She is co-editor-in-chief of the *International Journal of Learner Corpus Research* and a founding member of the Learner Corpus Research Association. Her research interests include corpus linguistics, learner corpus research, vocabulary, phraseology, second language acquisition, linguistic complexity, crosslinguistic influence, English for Academic Purposes, pedagogical lexicography and electronic lexicography.

**Aek Phakiti** is an associate professor in TESOL at the University of Sydney. His research focuses on language testing and assessment, second language acquisition, and research methods in language learning. He is the author of *Strategic Competence and EFL Reading Test Performance* (2007), *Experimental Research Methods in Language Learning* (2014), *Language Testing and Assessment: From Theory to Practice* (Bloomsbury, forthcoming), and, with Carsten Roever, of *Quantitative Methods for Second Language Research: A Problem-Solving Approach* (2018). With Brian Paltridge, he edited the *Continuum Companion to Research Methods in Applied Linguistics* (2010) and *Research Methods in Applied Linguistics: A Practical Resource* (2015). He is Associate Editor of *Language Assessment Quarterly*. He was Vice President of ALTAANZ (Association for Language Testing and Assessment of Australia and New Zealand, 2015–2017).

**Jesse Pirini** is a lecturer in the School of Management at the Victoria Business School, Victoria University of Wellington. Jesse received his PhD at the Auckland University of Technology, studying knowledge communication, agency and intersubjectivity in high school tutoring. Jesse develops multimodal theory and methodology. He works with a wide range of data sources, including family interaction, high school tutoring, augmented reality and video conferencing. Along with academic journal articles and chapters, Jesse is also the author of a practical workbook for training tutors and he supports community-based peer tutoring programmes.

**Luke Plonsky** is Associate Professor of Applied Linguistics at Northern Arizona University, where he teaches courses in research methods and second language acquisition. Recent and forthcoming publications in these and other areas can be found in journals such as *Annual Review of Applied Linguistics*, *Applied Linguistics*, *Language Learning*, *The Modern Language Journal*, and *Studies in Second Language Acquisition*, as well as in edited volumes published by Cambridge University Press, Wiley Blackwell, De Gruyter, and others. He is also Associate Editor of *Studies in Second Language Acquisition*, Managing Editor of *Foreign Language Annals*, and Co-Director of IRIS ([iris-database.org](http://iris-database.org)).

**Matthew T. Prior** is Associate Professor of Applied Linguistics/Linguistics/TESOL in the Department of English at Arizona State University, where he teaches courses in qualitative methods, discourse analysis, sociolinguistics, TESOL, and second language acquisition. His interests include narrative, discursive-constructionist approaches to identity, and social-psychological dimensions of multilingualism. He is author of *Emotion and Discourse in L2 Narrative Research* (2016) and co-editor of the volume *Emotion in Multilingual Interaction* (2016).

**Sanghee Ryu** is a research professor in the research center of Korean Language and Literature Education at Korea University, South Korea. Ryu's research focuses on the use of discourse analysis and formative-design experiments to explore and improve the teaching and learning of argumentative writing with an emphasis on underlying definitions of rationality. Ryu has taught pre-service teacher education courses on teaching reading and teaching writing at The Ohio State University. She teaches graduate courses on research methodology at Korea University.

**Betty Samraj** is Professor of Linguistics at San Diego State University. Her main research interests are in academic writing in different disciplines (including interdisciplinary fields) and genre analysis. She has conducted analyses of several different genres such as research article introductions, abstracts, masters theses, graduate student research papers, manuscript reviews, personal statements and, most recently, suicide notes. She teaches teacher preparation courses such as English for Specific Purposes and Teaching ESL Reading and Writing in a masters program in applied linguistics.

**Gale Stam** is Professor of Psychology at National Louis University in Chicago, Illinois. Her research interests include language, culture, and cognition; gesture; and L1 and L2 acquisition. She has published articles on changes in thinking for speaking, the importance of looking at gesture in L2 acquisition, gesture and lexical retrieval in an L2 setting, and language teachers' gestures. She serves on the editorial board of the journals *Gesture* and *Language and Sociocultural Theory* and has co-edited two volumes: *Gesture: Second Language Acquisition and Classroom Research* (2008) and *Integrating Gestures: The Interdisciplinary Nature of Gesture* (2011).

**Sue Starfield** is a professor in the School of Education at UNSW Sydney. With Brian Paltridge, she is co-author of *Thesis and Dissertation Writing in a Second Language: A Handbook for Supervisors* (2007) and of *Getting published in academic journals: Negotiating the publication process* (2016) and co-editor of the *Handbook of English for Specific Purposes* (2013). She co-authored *Ethnographic Perspectives on Academic Writing* with Brian Paltridge and Christine Tardy (2016). With Brian Paltridge, she is co-editor of two new book series: *Routledge Introductions to English for Specific Purposes* and *Routledge Research in English for Specific Purposes*. Her research interests include tertiary academic literacies, advanced academic writing, postgraduate pedagogy, ethnographic methodologies, identity in academic writing, and access and equity in higher education.

**Crissa Stephens** is a doctoral candidate at the University of Iowa. Her work uses a critical sociocultural lens to examine how language policies interact with social identity development and opportunity in education. Her teaching and activism in the US and abroad help to inspire her approach, and her recent publications utilize ethnographic and discourse-analytic methods to explore language policy and educational equity in local contexts.

**Scott Sterling** is Assistant Professor of TESOL and Linguistics in the Department of Languages, Literatures, and Linguistics at Indiana State University. His recent work investigates the level of training, current beliefs and practices that the field of applied linguistics has towards research ethics. His main area of focus is meta-research, particularly research ethics, and he has published work related to these topics in various journals and edited volumes in linguistics. He completed his PhD at Michigan State University in 2015 with a dissertation that focused on the complexity and comprehensibility of consent forms used in ESL research.

**Jeffrey Stewart** is Director of Educational Measurement and a lecturer at Kyushu Sangyo University in Japan. He has published articles in numerous journals such as *TESOL Quarterly* and *Language Assessment Quarterly* regarding vocabulary acquisition and testing using a number of advanced statistical modeling tools, most specifically item response theory.

**Kaitlyn M. Tagarelli** works as a postdoctoral fellow in Psychology and Neuroscience at Dalhousie University in Halifax, Canada. She received her PhD in Applied Linguistics from Georgetown University and her research uses behavioral, Event-related Potential (ERP), and Functional magnetic resonance imaging (fMRI) techniques to examine the neural and cognitive mechanisms involved in language learning and processing. Dr. Tagarelli is particularly interested in the brain structures and memory systems underlying language learning, and how individual differences and learning conditions interact with learning processes and outcomes. Her work has appeared in edited volumes and *Studies in Second Language Acquisition*.

**Hong Wang** is a subject librarian and information specialist at the University of Auckland. She has a masters degree in library and information science, a bachelor's degree in foreign language education, and an associate degree in computer science. She has extensive experience in lecturing on information literacy, and she has also taught ESL and Chinese in various instructional settings in China and the U.S.

**Terrence G. Wiley** is Professor Emeritus at Arizona State University and immediate-past President of the Center for Applied Linguistics, specializing in language education and policy. His recent works include *Handbook of Heritage, Community, and Native American Languages: Research, Policy, and Practice* (co-editor, 2014) and *Review of Research in Education, 2014, 38(1)*. Wiley co-founded the *Journal of Language, Identity and Education* and the *International Multilingual Research Journal*. He is organizer of the International Language Policy Research Network of Association

Internationale de la Linguistique Appliquée and recipient of the American Association for Applied Linguistics Distinguished Scholarship and Service Award (2014).

**Anna Her Yang** is a doctoral student in the Department of Language and Literacy Education at the University of Georgia. She is also the project coordinator of a five-year National Professional Development grant. Her research interest primarily focuses on the pedagogical experiences of mainstreamed ESOL (content-area) teachers of English learners.

**Soo Jung Youn** is Assistant Professor of English at Northern Arizona University, USA. Her academic interests include L2 pragmatic assessment, task-based language teaching, quantitative research methods, and conversation analysis. In particular, her research focuses on assessing L2 learners' ability to accomplish various pragmatic actions in interaction by investigating a wide range of interactional features indicative of a varying degree of pragmatic competence using mixed methods. Her studies have recently been published in *Language Testing, System*, and *Applied Linguistics Review*.

**Richard F. Young** is Emeritus Professor of English at the University of Wisconsin-Madison and Chutian Professor in the School of Foreign Languages at Central China Normal University. His research focuses on the relationship between language and social context and has resulted in four books: *Discursive Practice in Language Learning and Teaching* (2009), *Language and Interaction* (2008), *Talking and Testing*, and *Variation in Interlanguage Morphology* (1998), as well as over 70 articles.

# List of Figures

Fig. 1.1	The five key stages of empirical research	13
Fig. 2.1	“Practicing speaking” in Spanish (Hall, 2004, p. 76)	34
Fig. 2.2	Representing embodied cognition (Goodwin, 2003, Fig. 2.9, p. 35)	39
Fig. 3.1	Beeswarm plot of interest in interacting with Chatbot (Data 1) and human partner (Data 2)	61
Fig. 3.2	Diagram of longitudinal Chatbot experiment design	65
Fig. 3.3	Combination interaction/boxplots of the longitudinal Chatbot data	66
Fig. 3.4	A parallel plot showing interest in human versus Chatbot interlocutors over three testing times	68
Fig. 3.5	Hypothesized model of interest in task and course	71
Fig. 3.6	Final model of interest in task and course	71
Fig. 9.1	The four questions framework	185
Fig. 9.2	Visual prompt for a literature review	188
Fig. 9.3	How is my study contributing?	191
Fig. 12.1	Draft 1 of fieldnotes	254
Fig. 12.2	Coded fieldnotes	259
Fig. 12.3	Screenshot of Transana programme used to collate fieldnotes and recordings (Hall, personal data, 2015)	264
Fig. 14.1	Sample visual world. Note: In this example, “cat” is the target, “caterpillar” is an onset competitor, “bat” is a rhyme competitor, and “hedgehog” is an unrelated distractor. Images are from the Multipic database (Duñabeitia et al., 2017)	291
Fig. 14.2	Sample data from mouse-tracking language experiment. Note: The black line represents a competitor trajectory; the gray line represents a target trajectory. Images are from the Multipic database (Duñabeitia et al., 2017)	292

Fig. 14.3	Sample ERP waves and scalp topography maps of the standard ERP correlate of semantic processing ( $N400$ ). Note: Each tick mark on y-axis represents 100 ms; x-axis represents voltage in microvolts, $\pm 3\mu\text{V}$ ; negative is plotted up. The black line represents brain activity to correct items, such as <i>plane</i> in example 2a. The blue line represents brain activity to a semantic anomaly, such as <i>cactus</i> in example 2b. The topographic scalp maps show the distribution of activity in the anomaly minus correct conditions with a calibration scale of $\pm 4\mu\text{V}$ . From data reported in Grey and Van Hell (2017)	296
Fig. 14.4	Examples of (a) semantic priming using lexical decision, (b) masked semantic priming, and (c) syntactic priming using a picture description task. Note: Drawing credit: Kyle Brimacombe	298
Fig. 14.5	Artificial linguistic systems in language learning paradigms (based on Morgan-Short et al., 2010; Saffran et al., 1996; Tagarelli, 2014). Note: Drawing credit: Kyle Brimacombe	302
Fig. 17.1	<i>Grammar and Beyond 4</i> , “Avoid Common Mistakes” box (p. 75)	365
Fig. 17.2	“Be careful note” on the overuse of modal auxiliaries (MEDAL2, p. 17)	367
Fig. 19.1	Scatterplots of four samples of students’ scores	401
Fig. 19.2	Scatterplots indicating small, medium, and large $r$ in L2 research	402
Fig. 19.3	Crosshatched area representing an $r^2$ of 0.25 (25%)	403
Fig. 19.4	Representation of Pearson’s $r$ as a non-directional measure	404
Fig. 19.5	Representation of regression as a directional measure	404
Fig. 19.6	Scatterplot for predicting OLA from $\text{LR}_{(\text{years})}$	406
Fig. 19.7	Menu for selecting simple regression analysis in SPSS	407
Fig. 19.8	Selections for running regression analysis in SPSS	407
Fig. 19.9	Statistics for running regression in SPSS	408
Fig. 19.10	ANOVA partitioning of total sum of squares (SOS) in OLA ( $R^2 = 50.1\%$ )	410
Fig. 19.11	Scatterplot with for $\text{LR}_{(\text{years})}$ predicting OLA with the regression line	413
Fig. 19.12	Factor shown as the commonly shared area among standardized variables	417
Fig. 20.1	EFA versus PCA	425
Fig. 20.2	12 essential steps in EFA	429
Fig. 20.3	Screenshot of the strategy use in lectures data	429
Fig. 20.4	Descriptive statistics options in SPSS	430
Fig. 20.5	EFA in SPSS	432
Fig. 20.6	Factor analysis menu	433
Fig. 20.7	SPSS Descriptives dialog box	433
Fig. 20.8	SPSS extraction dialog box	435
Fig. 20.9	SPSS extraction dialog box	436



Fig. 20.10	Scree plot (PCA)	438
Fig. 20.11	Creating a parallel analysis syntax	439
Fig. 20.12	Customising a parallel analysis syntax in SPSS	440
Fig. 20.13	Extracting factors using the principal axes factoring method with the fixed factor number = 5	441
Fig. 20.14	Scree plot (PAF)	443
Fig. 20.15	Rotation dialog box (direct Oblimin method)	444
Fig. 20.16	Rotation dialog box (Varimax method)	445
Fig. 20.17	Options dialog box	445
Fig. 20.18	Creating a factor score	452
Fig. 20.19	Factor scores in the SPSS data sheet	452
Fig. 20.20	Creating a composite score for comprehending strategies	453
Fig. 21.1	A third-order factor CFA model	462
Fig. 21.2	CFA model of reading performance (Standardised solution; $N = 651$ )	463
Fig. 21.3	CFA model of reading performance (Unstandardised solution; $N = 651$ )	464
Fig. 21.4	A hypothesised SEM model of the influences of trait cognitive and metacognitive processing on reading performance ( $N = 651$ )	466
Fig. 21.5	Eight essential steps in CFA or SEM	470
Fig. 21.6	Open a data file in EQS	478
Fig. 21.7	EQS spreadsheet	479
Fig. 21.8	EQS diagram drawing tool	480
Fig. 21.9	EQS diagram drawing canvas	481
Fig. 21.10	Factor structure specification	481
Fig. 21.11	A hypothesised CFA of comprehending strategies	482
Fig. 21.12	EQS model specifications	483
Fig. 21.13	EQS model specifications	484
Fig. 21.14	Analysis in EQS	485
Fig. 21.15	Distribution of standardised residuals	486
Fig. 21.16	Parameter estimates options	490
Fig. 21.17	First-order CFA for comprehending strategy use (unedited version)	491
Fig. 21.18	Revised first-order CFA model for comprehending strategy use	492
Fig. 21.19	Distribution of standardised residuals (revised model)	493
Fig. 21.20	Revised CFA model of comprehending strategy use	494
Fig. 22.1	Histogram of vocabulary test scores for Teaching Method 2	507
Fig. 23.1	Bar plot displaying normed frequency of <i>said</i> in news and other CORE registers	527
Fig. 23.2	Resampled mean differences based on the data from Donaldson (2011)	533
Fig. 25.1	TAALES GUI	579
Fig. 25.2	TAALES .csv file for analysis	580



**xxx**      **List of Figures**

Fig. 25.3	WEKA explorer	581
Fig. 25.4	File selection in WEKA	581
Fig. 25.5	Histogram for normally distributed TAALES index	582
Fig. 25.6	Histogram for non-normally distributed TAALES data	583
Fig. 25.7	Selection of model in WEKA	585
Fig. 25.8	Selection of cross-validation type in WEKA	585
Fig. 25.9	Initial linear regression model reported in WEKA with suppression effects	586
Fig. 25.10	Final linear regression model	587
Fig. 33.1	Heritage language dissertations between 2011 and 2016 by country	748
Fig. 33.2	Language(s) studied by articles in the <i>Heritage Language Journal</i> and dissertations. Note: Some of the articles published in the <i>Heritage Language Journal</i> and dissertations published in the ProQuest Database contain the examination of one more heritage language	751

# List of Tables

Table 5.1	Common types of mixed methods designs	107
Table 6.1	A comparison between traditional reviews and research syntheses	140
Table 9.1	Thesis proposals: structure and purpose (based on Paltridge & Starfield, 2007, p. 61)	194
Table 10.1	Moves in empirical research article introductions	203
Table 10.2	Dimensions to consider when constructing a research article	214
Table 10.3	Discovering norms for use of metadiscoursal features	215
Table 14.1	Common language-related ERP effects	295
Table 19.1	Two variables showing a <i>perfectly positive</i> Pearson's <i>r</i>	397
Table 19.2	Two variables showing a <i>perfectly negative</i> Pearson's <i>r</i>	398
Table 19.3	Imperfect <i>r</i> due to differences in <i>ordering</i>	399
Table 19.4	Imperfect <i>r</i> due to differences in <i>scores' shapes</i>	399
Table 19.5	Imperfect <i>r</i> due to differences in <i>scores' shapes</i> and <i>ordering</i>	399
Table 19.6	Data for predicting OLA from LR ( $N = 10$ )	405
Table 19.7	SPSS output of model summary from simple regression analysis	408
Table 19.8	ANOVA output table for simple regression analysis in SPSS	409
Table 19.9	Output for regression coefficients in SPSS	412
Table 19.10	Result of prediction of OLA from LRyears for our ten participants	412
Table 19.11	Coefficients table with modified scale for predictor variable	413
Table 20.1	Descriptive statistics of items one to five	431
Table 20.2	Cronbach's alpha coefficient of the questionnaire	431
Table 20.3	KMO and Bartlett's test based on 37 items	434
Table 20.4	Communalities (initial and extracted)	437
Table 20.5	Total variance explained	438

Table 20.6	A comparison of eigenvalues from the dataset with those from the parallel analysis	440
Table 20.7	KMO and Bartlett's test based on 25 items	441
Table 20.8	Initial and extraction values based on the PAF method	442
Table 20.9	Total variance explained (five-factor extraction)	442
Table 20.10	Factor correlation matrix	444
Table 20.11	Rotated factor matrix based on the Varimax method (based on 25 items)	447
Table 20.12	Rotated factor matrix based on the Varimax method (based on 24 items)	448
Table 20.13	Cronbach's alpha coefficient of each factor (based on 24 items)	448
Table 20.14	Rotated factor matrix based on the Varimax method (based on 23 items; final model)	449
Table 20.15	Rotated factor matrix based on the Varimax method (based on 23 items; final model)	450
Table 20.16	Correlations between factor scores and composite scores ( $N = 275$ , ** = $p < 0.01$ )	453
Table 21.1	Common symbols used in CFA and SEM	461
Table 21.2	Summary of the key goodness-of-fit criteria for CFA and SEM (based on, e.g., Brown, 2015; Byrne, 2006; Kline, 2016; Schumacker & Lomax, 2016)	473
Table 21.3	Model fit indices	487
Table 21.4	Test statistics	489
Table 21.5	The results of the revised first-order CFA model of comprehending strategy use	493
Table 21.6	EFA and CFA factor loadings for the comprehending factor	495
Table 22.1	Teaching method and vocabulary test score data	504
Table 22.2	Descriptive statistics for hypothetical ANOVA data	506
Table 22.3	Standard deviations for each group of vocabulary test scores	508
Table 22.4	ANOVA output for mean differences in vocabulary test score	510
Table 22.5	Tukey HSD post hoc comparisons of mean differences in vocabulary test score	511
Table 23.1	Frequency of <i>said</i> in online news and other CORE registers	526
Table 23.2	Contingency table of frequencies for <i>said</i> in the CORE	527
Table 23.3	Summary of key information for the analyses included in this chapter	537
Table 25.1	Correlation between SST scores and variables entered into regression	584
Table 33.1	Type of data collected in HL-CL dissertations and articles	749
Table 33.2	Methodological trends in dissertations and articles	750
Table 38.1	Summary of data collection methods in L2 pragmatics research as discussed by Kasper and Rose (2002)	836

# Part I

## Research Approaches and Methodology

There are ten chapters in this part of the Handbook. In each chapter, the authors discuss historical development, current and core issues, key research processes, an overview of data analysis, challenges and controversial issues, and limitations and future directions. The authors also provide resources for further reading.

- In Chap. 1 (Applied Linguistics Research: Current Issues, Methods, and Trends), Aek Phakiti, Peter De Costa, Luke Plonsky, and Sue Starfield present a broad contextualization of applied linguistics research, locating its focus within current debates and concerns of relevance to the field of applied linguistics. The editors highlight the field's growing interest in research methodology and offer a rationale for the selection of topics and issues in the *Handbook*, such as methodological reform, transparency, transdisciplinarity, and the impact of technology.
- In Chap. 2 (Habits of Mind: How Do We Know What We Know?), through the sociology of science pioneered by Ludwik Fleck, Richard Young explores the ways that different researchers attend to different aspects of language learning and use are habits of mind grounded in the communities to which they belong. Young uses Fleck's three characteristics of thought collectives—their rhetoric, their epistemology, and incommensurability among thought collectives—to consider different methodologies of applied linguistic research and to describe how different habits of mind constrain how researchers know what they know.
- In Chap. 3 (Quantitative Methodology), Luke K. Fryer, Jenifer Larson-Hall, and Jeffrey Stewart present and justify the methodological choices in a complex, longitudinal, classroom-based study. Fryer, Larson-Hall, and

J. Stewart walk the reader through choices that must be made in a quantitative analysis step by step while also advocating for best practices in quantitative research, such as using technology as a partner in research methodology, strengthening statistical power by repeated testing of the same participants, and strengthening validity of study results by using a longitudinal design.

- In Chap. 4 (Qualitative Methodology), Shim Lew, Anna Her Yang, and Linda Harklau provide an overview of qualitative research (QR) in applied linguistics, with a particular focus on recent research and developments in the field over the past five years. Lew, Yang, and Harklau explore the philosophical and methodological premises of qualitative methods in the field and relate them to broader developments in QR across the social sciences, particularly in regard to issues of validity and quality of QR. They also review current trends and issues and provide an overview of research types including the varieties of qualitative approaches taken, theoretical frameworks used, and types of data collection and analytical methods employed.
- In Chap. 5 (Mixed Methodology), Alison Mackey and Lara Bryfonski present approaches to mixed methods research for social sciences and applied linguistics research. Mackey and Bryfonski demonstrate how different approaches can be used together in applied linguistics research, and provide practical advice on how to conduct a mixed methods study along with some suggestions for design and data analysis.
- In Chap. 6 (Traditional Literature Review and Research Synthesis), Shaofeng Li and Hong Wang discuss the procedures and best practices of traditional reviews of the literature and research syntheses for applied linguistics research. Li and Wang compare the two approaches and propose ways to integrate them in order to inform an empirical research project.
- In Chap. 7 (Research Replication), Rebekha Abbuhl traces the history of replication research in the field of applied linguistics, culminating in a discussion of current views of replication research as a means of evaluating the internal and external validity of a study, illuminating phenomena of interest, and ultimately, driving both theory and pedagogy forward. Abbuhl provides an overview of different types of replication studies (exact, approximate, conceptual) with recent examples from the field and concludes with current recommendations for facilitating replication research, including those pertaining to reporting and data sharing.
- In Chap. 8 (Ethical Applied Linguistics Research), Scott Sterling and Peter De Costa present the historical development of research ethics in applied linguistics and outline core issues that applied linguists are likely to face when conducting research. Sterling and De Costa recommend the best

ethical practices within the field. This chapter can be used to support the teaching of research ethics in graduate-level applied linguistics courses.

- In Chap. 9 (Writing a Research Proposal), Sue Starfield discusses the form and function of a typical research proposal which provides a rationale and motivation for a research study. Starfield provides a range of tools and techniques that can assist doctoral or graduate students in conceptualizing and writing this high-stakes document.
- In Chap. 10 (Writing a Research Article), Betty Samraj discusses key features of the applied linguistics research article. Samraj presents the novice writer with some questions to consider when constructing a research article in applied linguistics. This chapter also focuses on pedagogical issues related to the instruction of writing the research article, such as rhetorical consciousness raising and the use of annotated corpora.



# 1

## Applied Linguistics Research: Current Issues, Methods, and Trends

Aek Phakiti, Peter De Costa, Luke Plonsky,  
and Sue Starfield

### Introduction

Applied linguistics is a broad, evolving, interdisciplinary field of language and language-related study across diverse social contexts (e.g., Cook, 2003; Davies & Elder, 2004; Hall, Smith, & Wicaksono, 2011; Kaplan, 2010; Pawlak & Aronin, 2014; Phakiti & Paltridge, 2015; Schmitt, 2002; Simpson, 2011). This diversity, while a strength of the field, occasions disagreement in how applied linguistics should be defined. Part of the reason for this is that, as de Bot (2015) points out, the word *linguist* as used in the term *applied linguist* often leads to the misunderstanding that applied linguists simply “apply” lin-

---

A. Phakiti (✉)

Sydney School of Education and Social Work, The University of Sydney,  
Sydney, NSW, Australia

e-mail: [aek.phakiti@sydney.edu.au](mailto:aek.phakiti@sydney.edu.au)

P. De Costa

Department of Linguistics and Languages, Michigan State University,  
East Lansing, MI, USA

e-mail: [pdecosta@msu.edu](mailto:pdecosta@msu.edu)

L. Plonsky

Northern Arizona University, Flagstaff, AZ, USA

e-mail: [luke.plonsky@nau.edu](mailto:luke.plonsky@nau.edu)

S. Starfield

School of Education, UNSW Sydney, Sydney, NSW, Australia

e-mail: [s.starfield@unsw.edu.au](mailto:s.starfield@unsw.edu.au)

guistic knowledge but most applied linguists would agree that this is hardly ever the case, as due to its interdisciplinary nature, applied linguistics as a field of study sits at the intersection of a diversity of fields.

Research fields that are related to and influence applied linguistics research include linguistics, psychology, philosophy, education, and sociology. Furthermore, several terms, such as applicable linguistics (Mahboob & Knight, 2010) and applied language studies (Richards, Ross, & Seedhouse, 2012), are used interchangeably with applied linguistics, contributing to the richness of debates over the definition of applied linguistics.

## Evolution of Applied Linguistics Research

Applied linguistics is a relatively youthful field which emerged in the latter of half of the twentieth century; one of the field's flagship journals, *Applied Linguistics*, published its first issue in 1980 and others are of even more recent vintage. There are common terms that underlie applied linguistics research (e.g., language, linguistics, language learning, real-world language use, and language in social contexts), and there are several subdisciplines under the applied linguistics umbrella (e.g., first and/or second language acquisition/learning, language teaching and education, language assessment, and bilingualism/multilingualism). In each of these subdisciplines, researchers ask and address different research problems, employing a variety of philosophies and methodologies.

In his pursuit to address the definition and scope of applied linguistics, de Bot (2015) employed interviews and questionnaires with approximately 100 applied linguists. Key questions included “what is applied linguistics?”, “who are the key players?”, and “who influences them and whom do they influence in turn?”. de Bot also traces the main trends that have driven and impacted on applied linguistics in the past 30 years, including social and dynamic turns and the adoption of various paradigms in the field.

The primary intent of this *Handbook*, however, is not to resolve existing difficulties in defining applied linguistics. Rather, the *Handbook* has been driven by, and builds on, current work in applied linguistics research, and it aims to contribute to the advancement of research methodology in applied linguistics with a particular focus on supporting the access of early career researchers and emerging scholars to the rich diversity of research approaches currently being deployed in the field.



## Methodological Reform in Applied Linguistics

That this volume is being published now, in 2018, is not by sheer chance. Along with many other developments and signs of the field's maturity (see, e.g., Cook, 2015; de Bot, 2015; Gass, 2009; Gass & Mackey, 2012), awareness of and concern over methodological issues has greatly increased in recent years. Recent remarks by the former editor of *The Modern Language Journal* confirm this view, reminding us that "it appears that at this point in the development of applied linguistics, [methodological issues] demand a kind of professional scrutiny that goes directly to the core of what we do and what we know and what we can tell our publics that we know—and not only how we do it" (Byrnes, 2013, p. 825). We believe this volume represents an indicator of and contribution to what Byrnes referred to as the "methodological turn" (p. 825) taking place in the field.

Other indicators can be found in a wide variety of venues and activities that make up this vibrant discipline. At the societal level, for example, the American Association for Applied Linguistics (AAAL) recently introduced a *Research Methods* strand at its annual conference, thus formally recognising that methodological issues are at the heart of our research endeavours and deserve our full attention. In addition, AAAL and other conferences such as the Second Language Research Forum (SLRF) and the Language Testing Research Colloquium (LTRC) have hosted numerous pre-conference workshops in recent years focusing on methodological issues and practices.

As a journal-based social science, it is also natural that signs of methodological reform would be evident in the field's journals. Indeed, the past few years have seen a number of special issues in leading applied linguistics journals (e.g., *Applied Linguistics*, *Language Learning*, and *TESOL Quarterly*) that have aimed to address methodological issues. In 2016, *Applied Linguistics* published a special issue on "Innovation in research methods in applied linguistics," which explored how innovation had the potential to influence methods of applied linguistics research (e.g., digital learning environments, open-source software, and the rise of mixed methods research). And the previous year's special issue in the same journal focused on one particular methodological approach: meta-analysis. Also in 2015, *Language Learning* published a special issue on "Improving and extending quantitative reasoning in second language research." Following this issue, *Language Learning* put forth a new set of guidelines for reports of quantitative research (Norris, Plonsky, Ross, & Schoonen, 2015). Similarly, another one of the field's leading journals, *TESOL Quarterly*, recently revised and updated its research guidelines on experimental

research, survey research, ethnographic research, discourse analysis, and practitioner research (Mahboob et al., 2016), acknowledging the importance given to authors' approaches to research when submitting articles to the journal. More recently, *Language Teaching* has created a new section titled "Thinking Allowed" that has established applied linguists describe and suggest how to conduct research within specific areas and on specific topics (e.g., Norton & De Costa, 2018). Increasingly, journals such as *System* are beginning to explore how methodological innovation can be realised along topical lines such as a study abroad (De Costa, Rawal, & Zaykovskaya, 2017).

Another force behind the move to improve research practices in applied linguistics stems from the growing body of methodological syntheses. Such studies apply meta-analytic techniques to systematically review research and reporting practices. Some have focused on a particular substantive domain, such as second language interaction (Plonsky & Gass, 2011), feedback on second language writing (Liu & Brown, 2015), computer-mediated communication (Ziegler, 2016), and learner corpus research (Paquot & Plonsky, 2017). Others have looked across domains at one or more research techniques such as replication research (Marsden, Morgan-Short, Thompson, & Abugaber, *in press*), self-paced reading (Marsden, Thompson, & Plonsky, *in press*), designs, analyses, and reporting practices (Plonsky, 2013, 2014), and multiple regression analyses (Plonsky & Ghanbar, *in press*). Regardless of their focus, the purpose of these syntheses is not solely to describe and evaluate, but to use their findings to offer empirically grounded recommendations for improving research and reporting practices in the sub-domains they examine.

In addition, applied linguists are increasingly beginning to engage in discussions about conducting ethical research (e.g., De Costa, 2016; Tao, Shao, & Gao, 2017) and ensuring that the next generation of applied linguists is sufficiently prepared to deal with ethical dilemmas that may emerge during the research process. That leading professional organisations, such as the British Association for Applied Linguistics and the American Association for Applied Linguistics, have ethical guidelines listed on their websites is further testament to the growing significance assigned to preserving ethical care when conducting research.

One final indicator of methodological development to highlight are discussions and applications of novel analytical techniques. In most cases, the tools being introduced or tested are not entirely new; they are simply new to applied linguistics. Examples of techniques that have been introduced recently include bootstrapping (Plonsky, Egbert, & LaFlair, 2015; see also Egbert & LaFlair, Chap. 23), mixed effect modelling (Cunnings & Finlayson, 2015; Linck & Cunnings, 2015), and Bayesian inference (Norouzian, de Miranda, & Plonsky, *in press*).

## Published Books in Applied Linguistics

There have been a number of handbooks devoted to issues, topics, and methods in applied linguistics research (e.g., Davies & Elder, 2004; Kaplan, 2010; Simpson, 2011; see discussion and a variety of perspectives on the place and value of handbooks in the *The Modern Language Journal's* "Perspectives" section; Byrnes, 2011). The publication of these handbooks has advanced our understanding of applied linguistics research. The present *Handbook* will add to this collection by providing up-to-date, cutting-edge coverage of research methodologies for applied linguistics research, as well as an update of some current and new research areas in applied linguistics.

In the field of applied linguistics, many research methods books provide broad overviews and detailed methodological procedures for applied linguistics research, but many of these deal with research methodology in general terms, covering ways of conducting applied linguistics research, research assumptions, and research done in specific strands (e.g., SLA, corpus linguistics). Other texts on research methods address a specific approach (e.g., case study, survey, and experimental research designs) within applied linguistics. These books include, for example, Brown and Rogers (2002), Dörnyei (2007), Duff (2008), Fulcher and Davidson (2012), Heller, Pietikäinen, and Pujolar (2018), Holliday (2016), McKay (2006), Mackey and Gass (2012), McKinley and Rose (2017), Nunan and Bailey (2009), Paltridge and Phakiti (2010, 2015b), Perry (2012), and Phakiti (2014).

## The Present Handbook

As mentioned earlier, applied linguists engage in a wide range of fields including *inter alia* language learning and teaching, second language acquisition, bi- and multilingualism, conversation analysis, corpus linguistics, critical discourse analysis, deaf linguistics, discourse analysis and pragmatics, forensic linguistics, language assessment, language planning and policies, language for specific purposes, lexicography, literacies, multimodal communication, and translation. Covering the majority of these topics, the *Handbook* testifies to the maturity of the field, as does the diversity of the methods under discussion.

This *Handbook* is composed of four parts; each of these houses a number of chapters by scholars in applied linguistics: (1) research approaches and methodologies, (2) research instruments, techniques, and data sources, (3) data analysis, and (4) research areas. In the following sections, we provide a broad discussion of the themes in each part of this *Handbook*.

## Research Approaches and Methodologies

Research can be described as a systematic process of inquiry to address a research problem or question of interest. Applied linguists may use a theoretical framework to guide their research inquiry as well as a methodological approach inherent to a particular theory. Research needs to be systematic in the sense that it requires researchers to understand the underlying principles associated with the tools that it employs. These tools include the particular research instrument or technique used and the data analysis techniques the researcher has chosen to use. Such an understanding is needed to ensure the validity or trustworthiness of research findings. In other words, good research practice requires an understanding of what constitutes good research (e.g., transparency of data collection and analysis, ethical conduct, and evidence-based inferences and conclusions).

### Primary and Secondary Research

Brown (2014), Dörnyei (2007), and Phakiti and Paltridge (2015) present some common dimensions of research that may be inherent to applied linguistics research. For example, researchers must distinguish between *primary* and *secondary research*. On the one hand, primary research involves the collection and analysis of empirical data (observed data from a source). Journal articles and theses or dissertations, which collect and analyse new empirical data and analysis to address a research problem, are examples of primary research. On the other hand, secondary research analyses and synthesises existing theories, hypotheses, and/or research findings from published sources to help a researcher understand a research topic or issue. Examples of secondary research include a literature review, a state-of-the-art journal article, a book chapter on a particular research issue or topics, and a research synthesis. It should be noted that research synthesis, including meta-analysis, is a sophisticated secondary research methodology that treats published studies as data (see Li & Wang, Chap. 6; Ortega, 2015; Plonsky & Oswald, 2015).

In reality, primary applied linguistics research is usually built on and informed by secondary research. Secondary research also helps researchers avoid researching a topic that is already well-understood. A review of the relevant literature, for example, helps researchers identify a research gap or a significant problem to be addressed, understand methods (e.g., instruments and types of analysis) other researchers have used, and decide on a research question to be asked and methods to be used (see also Li & Wang, Chap. 6; Starfield, Chap. 9).

## Basic and Applied Research

Primary research may also be further categorised as *basic* or *applied research*. Basic research aims to develop fundamental knowledge about an issue or topic in applied linguistics (what, why, and how something takes place or is present). Basic research helps researchers understand a phenomenon in a real-world context (e.g., how people learn and use a language or how and why a group of politicians use language as a tool to discriminate against immigrants). One outcome of basic research is a theory or a set of hypotheses that explain human behaviours, thoughts, or beliefs about language and language use. Applied research (not to be confused with applied linguistics as a field of inquiry) is related to researchers or practitioners' attempts to solve real-world problems in language learning or use by applying principles or theories from primary research. In language teaching, researchers may examine whether a particular teaching technique or approach helps learners achieve higher proficiency. In language policy and planning, researchers may investigate whether a popular framework or idea can be used in another setting (e.g., whether the CEFR (Common European Framework of Reference for Languages) might be adopted as part of government policy for teaching English in Thailand).

## Cross-Sectional and Longitudinal Research

Another dimension which is useful to understand when conducting research in applied linguistics is the distinction between *cross-sectional* and *longitudinal research*. Cross-sectional research takes place when researchers collect data from a group of research participants at a single point in time using instruments, such as tests, questionnaires, interviews or observations. Longitudinal research requires researchers to collect data over a period of time (e.g., over several years or time points) to understand changes or developments (see e.g., Fryer, Larson-Hall, & Stewart, Chap. 3). Cross-sectional research is more convenient and cost- and time-efficient than longitudinal research, which nevertheless is essential and much needed in applied linguistics, as it can enable us to understand changes, developments, and cause-and-effect phenomena.

## Quantitative, Qualitative, and Mixed Methods Research

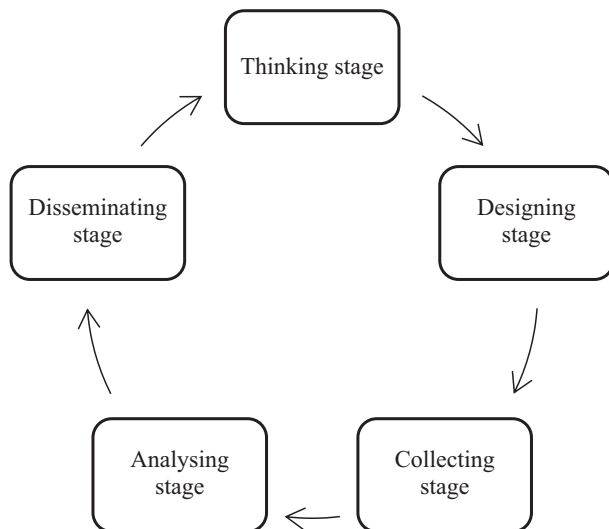
The distinction between quantitative and qualitative research is frequently made when research in applied linguistics is discussed. Moreover, mixed methods research is becoming increasingly popular. The key differences

between quantitative and qualitative research are the nature of data, data analyses, and the underlying philosophies of what constitutes reality (see Riazi, 2017). In quantitative research, on the one hand, data are numerical and some qualitative aspects (e.g., individual differences in levels of motivation, language proficiency, or attitudes) are quantified using a scale. Quantitative research often uses statistical analysis to address research questions. Quantitative researchers also tend to take an objectivist stance when collecting and analysing data and reporting a study (i.e., observing “from the outside” and writing in an impersonal tone). Qualitative research is interested in people’s behaviours, motivations, thoughts, interactions, experiences, perceptions, and meaning-making activities in general. Data may be collected by means of individual or group interviews, observations of behaviours or field notes, and other documents or texts. Qualitative researchers allow subjectivity to help them understand participants in a specific setting, often carrying out what is known as participant observation (Starfield, 2015). There is a range of qualitative methodological approaches and designs in applied linguistics (e.g., case study, ethnographic research, narrative inquiry), many of which are discussed in depth in this volume.

Mixed methods research has recently gained wide recognition in applied linguistics research (e.g., Brown, 2014; Ivankova & Greer, 2015; Mackey & Bryfonski, Chap. 5; Moeller, Creswell, & Saville, 2015; Riazi, 2017). In any study, multiple sources of data are always desirable to help researchers understand a topic of interest. The collection of various types of data in one study is known as data triangulation, which may be solely quantitative or qualitative or a combination of both. The idea behind mixed methods research is to maximise the effectiveness of a study by using the strengths of both quantitative (answering the *what* questions) and qualitative (answering the *how* and *why* questions) research (see Starfield, Chap. 9). Quantitative data may provide generalised findings, whereas qualitative data exemplifies unique cases or exceptions to the norm. It is also important to note that conducting a mixed methods study is not simply about combining quantitative and qualitative data in a single study. For example, mixed methods designs may be concurrent or consequential (see e.g., Ivankova & Greer, 2015; Mackey & Bryfonski, Chap. 5).

## Research Process

As discussed above, research is a systematic inquiry, and to be systematic, one needs to understand some of the key research processes. Paltridge and Phakiti (2015a, p. 260) provides a diagram that outlines specific research stages



**Fig. 1.1** The five key stages of empirical research

including topic selection, literature review, ethical considerations, data analysis, and research communication (e.g., writing a report or conference presentation). In this chapter, five key stages of research are presented (Fig. 1.1), which subsequently inform the organisation of different parts of this *Handbook*.

The *thinking stage* requires researchers to consider which topic should be researched. For example, a topic should be interesting not only to the researcher but also to other scholars in the discipline and be important for the advancement of knowledge. This stage can be enhanced through reflection and a realisation of what constitutes habits of mind as presented in Chap. 2 by Richard Young. Thinking alone is not sufficient for successful research. Researchers need to be able to review existing research to inform their decisions to focus on a particular issue or question. Chapter 6 by Shaofeng Li and Hong Wang helps provide a critical understanding of how to produce useful and informative literature reviews and research syntheses.

In the *designing stage*, researchers consider an approach or method that may be most suitable for the question being asked in the context of the study. As discussed earlier, researchers may decide to conduct quantitative, qualitative, or mixed methods research. In order to help the reader clarify key considerations in choosing a research design, in the first part of this *Handbook*, we have asked Luke K. Fryer, Jenifer Larson-Hall, and Jeffrey Stewart to write on quantitative methodology (Chap. 3); Shim Lew, Anna Her Yang, and Linda Harklau to write on qualitative methodology (Chap. 4); Alison Mackey and



Lara Bryfonski to write on mixed methodology (Chap. 5); and Rebekha Abbuhl to discuss the considerations involved in replicating a study (Chap. 7). Since designing is connected to data collection, it is important to understand the roles of ethics, and therefore a chapter by Scott Sterling and Peter De Costa (Chap. 8) describes what is involved in ethical research. Finally, prior to commencing research, many students are required to produce a research proposal which documents the thinking and designing stages, as well as the collecting and analysing stages; Sue Starfield provides useful and practical guidelines for preparing and writing a thesis or dissertation proposal (Chap. 9).

The *collecting stage* requires researchers to employ research instruments, techniques, and data sources to address their research project. This stage requires researchers to understand the key considerations in developing and using a particular instrument or technique (see research instruments, techniques, and data sources below). The *analysing stage* requires researchers to organise data and perform data analysis suitable for the research question (see research analysis below). Finally, in the *disseminating stage* researchers aim to consolidate their research project by revisiting all the previous stages. Researchers may attend a conference to present their findings, but usually writing a research report (e.g., a thesis or dissertation, a journal article) is expected as this approach enables dissemination to a much wider audience. In the first part of this *Handbook*, Betty Samraj (Chap. 10) discusses how to write a research article. As reflected in Fig. 1.1, the reporting stage is the beginning of the thinking stage, in which researchers consider the next research problem to be addressed. Finally, there are diverse or unique approaches and methods for doing research in a particular area or topic; Part 4 of this *Handbook* provides specific cases as examples of key processes (see research topics and areas below).

## Research Instruments, Techniques, and Data Sources

Applied linguistics researchers have tended to adopt research instruments and techniques traditionally associated with the paradigm and habits of mind (see Young, Chap. 2) within which their study is located: broadly quantitative or qualitative approaches. As explained above, mixed methods approaches have sought to leverage the benefits of drawing on techniques and instruments from both paradigms. In a recent paper, King and Mackey (2016) call for the field to go beyond mixed methods and pursue the greater integration of



research strategies from diverse epistemological perspectives in terms of the benefits this can afford to our understanding of complex, real-world, language-related problems.

In line with the overall approach taken in the *Handbook*, Part 2 highlights the diversity of research methods and data collection techniques adopted in applied linguistics. In addition to providing a state-of-the-art discussion of eight contemporary investigative approaches, each chapter contains profiles of sample studies selected by the author that illuminate the approach under examination. The first two chapters of this section are typically associated with qualitative research designs, while those that follow are more closely linked to quantitative approaches. Mixed method approaches would tend to draw on combinations of the strategies and techniques discussed below. While the data collection methods described in Part 2 are primarily employed in language learning research, several of them could fruitfully be used in other areas of applied linguistics too.

In his chapter on the interview in applied linguistics research (Chap. 11), Matthew Prior frames the interview as both “research instrument” and “social practice,” delving into the complexities of our everyday understandings of the term. While he takes a constructivist approach, focusing on what he calls the “narrative” interview, in his discussion of the evolution of the interview, he notes that “the interview is described as methodology, approach, method, technique, and tool—terminology that may evoke very distinct (and discipline-specific) conceptual and procedural lenses.”

In Chap. 12, Fiona Copland explores the when, where, how, why, and so what of field notes and observation as methodological tools in ethnographically oriented research. While still relatively underutilised in applied linguistics, ethnographic perspectives that draw their data from field notes and observation are becoming more widely adopted as awareness grows as to the value of the qualitative data that these techniques can afford. The chapter includes examples from Fiona’s own work and studies by colleagues.

Chapter 15 discusses techniques for eliciting data that enable applied linguists to study how languages are learned as well as the type and extent of learners’ knowledge about their additional languages. Susan Gass reviews two widely used data elicitation techniques - judgment and elicited imitation - and considers the uses and misuses of judgment data. The major focus of the chapter is on the practicalities involved in collecting acceptability/grammaticality judgment data, although other types of judgment data and an alternative to judgment data, namely, magnitude estimation, are also referred to.

In Chap. 16, Melissa Bowles provides an overview of the use of verbal reports, both think-alouds and stimulated recalls, and data elicitation techniques that are

used in first and second language research by researchers of various theoretical orientations to gather data about learners' thought processes. The chapter synthesises research that has examined their validity, finding them to be valid, if implemented appropriately, and includes a concise guide to their use in language research, from data collection to analysis, concluding with a discussion of the method's limitations and possible triangulation with other data sources.

Four of the chapters in this section highlight the impact of newer technologies on data collection. Online questionnaires that enable the collection of large amounts of data are discussed in Chap. 13 with Jean Marc Dewaele's detailed account of questionnaire development also applicable to questionnaires of diverse types in a range of contexts. The chapter draws on its author's extensive experience of carrying out large-scale online surveys and includes issues of sampling, representativity, researching vulnerable groups, and potential pitfalls as well as considering opportunities for novel research.

In Chap. 14, Sarah Grey and Kaitlyn Tagarelli provide a comprehensive description of how psycholinguistics methods which are experimental and laboratory based can both inform our understandings of language learning and use and reveal the psychological representations and processes that underlie language learning and use. The first part of the chapter examines common psycholinguistics measures and tasks, many of which are enabled by advances in technology, including decision tasks, reaction time, mouse tracking, and eye tracking as well as event-related potentials (ERPs). The second part covers psycholinguistic paradigms that investigate the mental underpinnings of language, reviewing common experimental approaches with relevance for applied linguistics, namely, priming, visual world, and language learning paradigms.

The ability to construct and search large online databases afforded by technological advances has facilitated the growth of corpus linguistics. Its array of methods and tools has contributed significantly to much more detailed and accurate descriptions of a wide variety of languages, in particular with regard to frequency, collocation, grammar and lexis, and situational factors. In Chap. 17, Magali Paquot focuses on corpus research for language learning and teaching, an area on which the methods and principles of corpus linguistics have had a significant practical impact.

The affordances of digital tools in the study of digital discourses can fundamentally alter the nature of context, text, and interaction. In Chap. 18, Christoph Hafner provides a detailed examination of the issues and challenges that arise for applied linguists when studying digital discourses and using digital tools. A key challenge is to develop appropriate approaches with appropriate methods of data collection, processing, and analysis. The chapter considers

four prominent approaches: digital ethnography, computer-mediated discourse analysis, corpus analysis, network analysis as well as ethical issues raised by these relatively new investigative approaches and research sites.

Technology has also enabled the creation of a unique resource, the *IRIS* digital repository <https://www.iris-database.org>, an online, searchable database intended to support methodological transparency in applied linguistics. *IRIS* is a free, open resource that hosts materials for many areas of language-related research, and numerous examples of the research instruments and techniques discussed above have been deposited and are easily accessible (see Marsden, Mackey, & Plonsky, 2016). Researchers are encouraged to participate and share instruments they have developed.

While the *Handbook* does not explicitly espouse the “layering” approach to data collection King and Mackey (2016) argue for, readers may wish to consider their arguments for greater integration when designing their studies. Nevertheless, the *Handbook* offers detailed insight into a range of data collection techniques and research instruments that may be used singly or in combination.

## Research Analysis

One way to understand the place of data analysis is to consider the research cycle discussed above (Fig. 1.1) in parallel to the process of preparing a meal. Just as a cook plans and shops for groceries, a researcher designs the study and collects the data. Likewise, serving the meal might be compared to writing up an article or presenting at a conference. The analysis stage in the metaphor is, of course, what we do with the groceries and other ingredients that will become the meal. As we see throughout Part 3 of this *Handbook*, just as there are many ways of preparing different types of meals, data in applied linguistics can be analysed in myriad fashion. It is also important to note that the analytical decisions we make (and the cooking techniques we apply), though varied, are not generally better or worse than others we might choose; rather, they represent and reflect different theoretical and often epistemological positions, preferences, and “tastes.”

As we examine some of these different stances and practices, we would draw your attention to two broad themes or dimensions across the third section of the *Handbook*. We encourage you to reflect on where your own research might fall in each and on how your own work could be enhanced by expanding your analytical repertoire by taking on new or enhanced techniques.

## Qualitative Versus Quantitative Analysis

One immediate and perhaps obvious way to characterise the different approaches to data analysis presented in section “[Research Instruments, Techniques and Data Sources](#)” concerns whether the techniques are applied to qualitative or quantitative data. For example, in this volume, there is a clearly quantitative orientation in the chapters covering correlation (Norouzian & Plonsky, Chap. 19), factor analysis and structural equation modeling (Phakiti, Chaps. 20 and 21), ANOVA and its variants (Amoroso, Chap. 22), distribution-free analyses (Egbert & LaFlair, Chap. 23), and measurement error and reliability (Grabowski & Oh, Chap. 24). Meanwhile, other chapters address techniques for handling largely qualitative data, such as narrative analysis (Benson, Chap. 26), interaction analysis (Miller, Chap. 27), and multimodal analysis (Pirini, Matelau, & Norris, Chap. 28).

The qualitative-quantitative distinction can be useful, but it can also be misleading, limiting our perspectives and closing potentially fruitful efforts to understand language-related phenomena. For instance, whereas the quantitative approaches described in Part 3 can help us to detect patterns and probabilities across a wide range of cases, they may obscure individual variation. Quantitative analyses also necessarily involve abstractions—often many levels of abstraction, actually—from the objects of our investigation. Techniques applied to qualitative data are, therefore, often considered to be “closer” to phenomena of interest.

For these reasons, there is much to be gained from an approach to data analysis that applies mixed methods (Mackey & Bryfonski, Chap. 5). Analyses of spoken and written discourse, such as those described in Crossley and Kyle (Chap. 25), often lend themselves very naturally to this approach. For other domains, mixed data types and analyses will often require stepping a bit outside of our comfort zone. We may even have to break with the conventions of our sub-domains which are often entrenched in one particular approach to data collection or analysis. Despite these conventions, almost any sub-domain could benefit from the addition of further and especially complementary data. For example, although Miller’s approach to interaction analysis is largely qualitative, conversational interaction certainly has been examined using quantitative techniques (e.g., Mackey, 2007; Plonsky & Gass, 2011).

## Breadth Versus Depth

Another dimension that can be used to characterise the different analytical approaches in Part 3 of this *Handbook* is their scope. Some analyses are designed to provide depth of knowledge on a single variable or a perhaps very

small set of variables. Approaches with a very narrow scope are often described as undertaking “micro-level” analysis. Other analyses aim for breadth of coverage around a particular language user, linguistic structure, context, or process. Analyses in this vein are much broader in scope and are often described as “macro-level” or multivariate.

As with many other decisions made throughout the process of conducting a study in applied linguistics or any other social science, choices on this dimension—between breadth and depth, macro- and micro-levels, univariate and multivariate analyses—are not generally or inherently superior to their alternates. They reflect, rather, a different focus, much like a camera which can be used to capture an entire landscape or the detail on a single flower petal found in that same massive landscape.

In correlational analyses, for example, researchers are generally concerned with pairs of variables or bivariate relationships. Such an approach can be very useful and straightforward. However, it is also limited in that very few variables of interest in applied linguistics can be fully understood in this fashion. Language learning, use, assessment, and virtually everything else applied linguists care about is inherently multivariate (e.g., Brown, 2015; Plonsky & Oswald, 2017). Consider the notion of textual cohesion discussed in Crossley and Kyle’s chapter as measured by tools developed for natural language processing such as Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) or TAACO (Crossley, Kyle, & McNamara, 2016). Any analysis of spoken or written language focusing on a single linguistic feature or even a small set of features or variables associated with cohesion would necessarily overlook others relevant to the construct of interest. At the same time, however, one could also argue that a study examining only cohesion, without any consideration of the broader social and/or educational context in which the production took place is, likewise, limited in perspective (i.e., too narrowly focused). The larger point here in the case of scope, as we find in evaluating decisions made when conducting a study, is that the analyses must be aligned with the goals of the study.

## Research Topics and Areas

Building on Parts 1, 2, and 3 that introduced methodologies, methods, and data analytic approaches, the chapters in Part 4, which further exemplify the research issues discussed in the preceding sections of the *Handbook*, are arranged along topical lines. As noted, applied linguistics is both a diverse and a growing field. Because of its inherent diversity, how one carves up and interprets this

field is also open to some level of interpretation, with professional organisations such as the American Association for Applied Linguistics electing to divide it along strand lines and handbooks (e.g., Simpson, 2011) according to topics. Given the relative subjectivity involved in determining what constitutes topics in applied linguistics, some degree of selection also had to be made in deciding what to be included in the fourth and final part of this *Handbook*. The final decision rested on an informed analysis of contemporary topics published in leading applied journals, monographs, and edited volumes as well as an informal survey of topics presented at major applied linguistics conferences in recent years. At the same time, we fully recognise that the entire breadth of applied linguistics is not represented in this section.

A review of the chapters in this section of the *Handbook* reveal that authors of each chapter commendably attempt to first explain what is examined within each respective topic before moving to describe how the topic can be examined. While they do not make explicit reference to King and Mackey's (2016) notion of *layering*, which calls for "the explicit consideration of research problems from a range of distinct epistemological perspectives" (p. 210), all chapters endorse the view that problems within applied linguistics can be investigated through a variety of methodologies, methods, and data analytic approaches. In addition, the authors in this *Handbook* section advocate that a methodological combination is often needed to best understand, measure, and apprehend an identified concern. Also binding the following chapters is an explication of methodological rigour and methodological innovation (De Costa et al., 2017). Such rigour and innovation are exemplified through four themes that emerged from the chapters.

## Transdisciplinarity

Because applied linguistic problems are best investigated from a range of epistemological perspectives (King & Mackey, 2016), a given topic can be examined from different perspectives. Tej Bhatia (Chap. 30), for example, highlights that bilingualism and multilingualism can be studied from cognitive, linguistic, neurological, psycholinguistic, sociolinguistic, and educational linguistic perspectives. More importantly, he also underscores the value of engaging in intellectual boundary crossings, a point that is reiterated by David Cassels Johnson and Crissa Stephens in their discussion of language planning and policy (Chap. 37). As emphasised by them, insights from adjacent fields, such as linguistic landscape and developments in geosemiotics, as evidenced by the adoption of the constructs of *scales* and *nexus analysis*, have provided language

policy scholars with an expanded theoretical and methodological tool kit. In a similar vein, research on academic literacies (Bloome, Carvalho, & Ryu, Chap. 41) has been enriched by cognitive approaches to researching literacy by focusing on the context of individual-text interaction and quantifying cognitive processes so that they can be statistically related (see also Loewen, Chap. 29). Such insights can be juxtaposed with a conceptualization and investigation of literacy as a social practice as well as a broadened understanding of social context to include social institutions and cultural ideological contexts.

## From Language to Semiotics

Also common across several chapters is an expanded understanding of language, one that moves beyond only focusing on language to include semiotics. As Gale Stam and Kimberly Buescher rightly point out in Chap. 36 with regard to gesture research, being competent in a second language involves the use of the entire body. Thus, an examination of language development from their perspective would require the use of video-recorded data that would be analysed through a discourse analytic approach like conversation analysis. Semiotic resources such as gaze, posture, gestures, and dress are also central to the investigation of identity, as noted by Ron Darvin (Chap. 35). Also underlining the importance of the “semiotic” beyond just the “linguistic,” David Malinowski (Chap. 40) argues for a rigorous consideration of the materiality and embodiment of multilingualism in his discussion of linguistic landscapes. Addressing pedagogical concerns, Soo Jung Youn’s (Chap. 38) exploration of sociopragmatics acknowledges the importance of understanding how semiotic resources are often used in conjunction with language to ensure that intercultural communication is successfully realised.

## Technological Developments

Methodological innovation has also been spurred by strides in technology that have afforded applied linguists’ valuable assistance in advancing their research agendas. Within translation and interpreting (Chap. 34), Claudia Angelelli points out that technology has resulted in multimodal and remote interpreting, for example. In addition, social media such as Google Translate have transformed ways in which translation is carried out and studied. Language testing and assessment (Chap. 39), according to April Ginther and Kyle McKintosh, has also benefited from the use of powerful statistical methods, including item



response theory (IRT) and structural equation modeling (SEM). These methods have been particularly helpful when analysing language testing data. Within linguistic landscape research (Chap. 40), David Malinowski highlights that digital tools such as Google Street View and other technological developments like Geographic Information Systems (GIS), digital mapping, and data visualisation now allow for greater levels of granular analysis. Finally, virtual representations of the self, made possible through digital affordances, have allowed new identities to be studied (Darvin, Chap. 35).

## Ethical Research Practices and Commitment to Social Change

Inroads into the digital realm, alongside other methodological progress, do not come without moral consequences, however. New ethical practices also need to be developed in accordance with changing times, as underlined by Ron Darvin (Chap. 35). Darvin reminds readers of the need to establish and implement ethical online research practices. April Ginther and Kyle McIntosh (Chap. 39) also underscore the need to consider ethical issues surrounding language testing and assessment. On a broader note, ethical research actions can take the form of engaging in research acts that bring social change (Sterling & Gass, 2017). One way to enact such change, as stated by Samuel Lerner (Chap. 31), is to carry out forensic linguistics research whose findings can prove useful when researchers use data in highly publicised trials and police investigations. Another way to contribute to social change is to conduct research on underrepresented populations such as heritage, community, and indigenous language learners (Bhalla & Wiley, Chap. 33). Finally, and in keeping with the emphasis on ethics addressed in Chap. 8 (Sterling & De Costa) in Part 1 of the *Handbook*, David Malinowski (Chap. 40), reiterates the significance of researcher reflexivity.

Indeed, as the chapters that follow will illustrate, applied linguistics is characterised by topical diversity. As researchers, the onus is on us to continually upgrade our methodological skill sets in order to keep abreast of these shifting research sands (Loewen, Chap. 29).

## Conclusion

This *Handbook* covers a range of research approaches, presents current perspectives, and addresses important considerations in different research methods. The topics—both conceptually and, at times, in a more hands-on



manner—run the gamut from designing and implementing research instruments and techniques to analysing data and reporting results across a wide variety of sub-domains. The *Handbook* also covers a wide range of contemporary research topics in applied linguistics and discusses several state-of-the-art issues (e.g., advancement of research methods or topics in applied linguistics).

It is our hope that the chapters in this volume serve the field and promote continued improvement in the ways we go about understanding language-related phenomena. We must also be sure to reflect regularly on our practices and seek to improve our skills, as methodological know-how is absolutely critical to the advancement of the field. To be sure, it is only through sound scientific practice, rigorous research methods, and transparent reporting of those methods that we, as a field, can advance our knowledge.

## Resources for Further Reading

Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University Press.

This book provides useful guidelines on how to conduct research in TESOL (Teaching English to Speakers of Other Languages) through the application of a mixed methods approach. The book outlines important research processes and addresses useful strategies to analyse quantitative, qualitative, and mixed data. The book has wider applications beyond TESOL research.

De Costa, P. I. (Ed.). (2016). *Ethics in applied linguistics research: Language researcher narratives*. New York: Routledge.

This edited volume, with a foreword by Lourdes Ortega and afterword by Jane Zuengler, is comprised of narratives of established applied linguists who reflect on how they negotiated ethical dilemmas that emerged during the research process. By focusing on their personal challenges, this book sheds light on *microethical* issues that are dealt with on a personal level. Such issues stand in contrast to *macroethical* issues that are generally mandated by ethical review boards and professional organisations.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.

This book presents key research methods in applied linguistics and includes a comprehensive discussion of quantitative, qualitative, and mixed methods approaches, criteria for judging research quality, cross-sectional and longitudinal data collection, data analysis, and research reports.

McKinley, J., & Rose, H. (Eds.). (2017). *Doing research in applied linguistics: Realities, dilemmas and solutions*. London and New York: Routledge.

This edited volume focuses on what researchers often face when they conduct research in applied linguistics (e.g., problems in the planning stage, during data collection, during data analysis, and when writing up a report). This volume highlights the distinction between ideal research methods and actual research practice as researchers are faced with real-life research contexts.

Mirhosseini, S.-A. (Ed.). (2017). *Reflections on qualitative research in language and literacy education*. New York: Springer.

This edited volume addresses epistemological perspectives, methodological problems, and practical considerations related to language and literacy research. Design and implementation issues in relation to established qualitative research traditions are also described. Practical aspects of a few actual instances of qualitative research conducted in different contexts exemplify methodologies discussed in the volume.

Paltridge, B., & Phakiti, A. (Eds.). (2015). *Research methods in applied linguistics: A practical resource*. London: Bloomsbury.

Paltridge and Phakiti's updated edited volume provides an introductory overview of research methods and approaches in applied linguistics (especially in the areas of language learning and teaching). New to this edition are chapters on mixed methods research and analysis, narrative inquiry, ethics in applied linguistics research, and developing a research project, as well as chapters in areas of research which include researching language learning strategies, young learners, teachers' beliefs, and language teacher education.

Paltridge, B., & Starfield, S. (2016). *Getting published in academic journals: Navigating the publication process*. Ann Arbor, MI: University of Michigan Press.

Aimed at junior scholars seeking to disseminate their research in academic journals, this book provides a step-by-step pathway through what may at times seem a complex and sometimes frustrating process.

Plonsky, L. (Ed.). (2015). *Advancing quantitative methods in second language research*. New York, NY: Routledge.

Basic descriptive and inferential statistics are appropriate in many if not all studies based on most quantitative data. On other occasions, however, a more sophisticated and often multivariate approach is needed. This volume seeks to guide second language researchers in their ability to both understand and apply such procedures. Analyses with hands-on tutorials include multiple regression, factor analysis, cluster analysis, mixed effects modelling, meta-analysis, bootstrapping, and structural equation modelling, among others.

Riazi, A. M. (2017). *Mixed methods research in language teaching and learning*. South Yorkshire: Equinox.

This book discusses the theoretical and philosophical foundations of mixed methods research in language learning and teaching, covering key research processes in a mixed methods study, and providing several examples of studies using mixed methods research.

Roever, C., & Phakiti, A. (2018). *Quantitative methods for second language research*. London and New York: Routledge.

This book provides a foundation for quantitative research literacy and practice. It includes 15 chapters on various statistical and quantitative analyses commonly used for second language research through the use of IBM SPSS programme (e.g., correlations, analysis of variance, multiple regression, and reliability analysis). A companion website hosts datasets and review exercises.

## References

- Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University Press.
- Brown, J. D. (2015). Why bother learning advanced quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 9–20). New York, NY: Routledge.

- Brown, J. D., & Rogers, T. S. (2002). *Doing second language research*. Oxford: Oxford University Press.
- Byrnes, H. (2011). Perspectives. *Modern Language Journal*, 95, 628–653.
- Byrnes, H. (2013). Notes from the editor. *Modern Language Journal*, 97, 825–827.
- Cook, G. (2003). *Applied linguistics*. Oxford: Oxford University Press.
- Cook, G. (2015). Birds out of dinosaurs: The death and life of applied linguistics. *Applied Linguistics*, 36, 425–433.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1–16.
- Cunnings, I., & Finlayson, I. (2015). Mixed effects modelling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159–181). New York, NY: Routledge.
- Davies, A., & Elder, C. (Eds.). (2004). *Handbook of applied linguistics*. London: Blackwell.
- de Bot, K. (2015). *A history of applied linguistics: From 1980 to the present*. London and New York: Routledge.
- De Costa, P. I. (Ed.). (2016). *Ethics in applied linguistics research: Language researcher narratives*. New York: Routledge.
- De Costa, P. I., Rawal, H., & Zaykovskaya, I. (2017). Study abroad in contemporary times: Toward greater methodological diversity and innovation. Special issue of *System*, 71.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.
- Duff, P. (2008). *Case study research in applied linguistics*. New York: Lawrence Erlbaum Associates.
- Fulcher, G., & Davidson, F. (Eds.). (2012). *The Routledge handbook of language testing*. London and New York: Routledge.
- Gass, S. (2009). A historical survey of SLA research. In T. K. Bhatia & W. C. Ritchie (Eds.), *The new handbook of second language acquisition* (pp. 3–27). Bingley: Emerald.
- Gass, S. M., & Mackey, A. (Eds.). (2012). *The Routledge handbook of second language acquisition*. London and New York: Routledge.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Matrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193–202. <https://doi.org/10.3758/BF03195564>
- Hall, C. J., Smith, P. H., & Wicaksono, R. (2011). *Mapping applied linguistics: A guide for students and practitioners*. London: Routledge.
- Heller, M., Pietikäinen, S., & Pujolar, J. (2018). *Critical sociolinguistic research methods: Studying language issues that matter*. New York: Routledge.
- Holliday, A. (2016). *Doing and writing qualitative research* (3rd ed.). Thousand Oaks: Sage.

- Ivankova, N. V., & Greer, J. L. (2015). Mixed methods research and analysis. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 63–81). London: Bloomsbury.
- Kaplan, R. (Ed.). (2010). *The Oxford handbook of applied linguistics* (2nd ed.). Oxford: Oxford University Press.
- King, K. A., & Mackey, A. (2016). Research methodology in second language studies: Trends, concerns, and new directions. *Modern Language Journal*, *100*, 209–227.
- Linck, J., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(S1), 185–207.
- Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing*, *30*, 66–81.
- Mackey, A. (Ed.). (2007). *Conversational interaction in second language acquisition: A collection of empirical studies*. Oxford: Oxford University Press.
- Mackey, A., & Gass, S. M. (2012). *Second language research: Methodology and design* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Mahboob, A., & Knight, N. K. (Eds.). (2010). *Applicable linguistics*. London: Continuum.
- Mahboob, A., Paltridge, B., Phakiti, A., Wagner, E., Starfield, S., Burns, A., ... De Costa, P. I. (2016). TESOL quarterly research guidelines. *TESOL Quarterly*, *50*, 42–65.
- Marsden, E., Mackey, A., & Plonsky, L. (2016). Breadth and depth: The IRIS repository. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York, NY: Routledge.
- Marsden, E. J., Morgan-Short, K., Thompson, S., & Abugaber, D. (in press). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning*, *68*(2).
- Marsden, E., Thompson, S., & Plonsky, L. (in press). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*.
- Mckay, S. L. (2006). *Researching second language classrooms*. Mahwah, NJ: Lawrence Erlbaum.
- McKinley, J., & Rose, H. (Eds.). (2017). *Doing research in applied linguistics: Realities, dilemmas and solutions*. London and New York: Routledge.
- Moeller, A. J., Creswell, J. W., & Saville, N. (Eds.). (2015). *Language assessment and mixed methods*. Cambridge: Cambridge University Press.
- Norouzian, R., de Miranda, M. A., & Plonsky, L. (in press). The Bayesian revolution in L2 research: An applied approach. *Language Learning*.
- Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*, 470–476.

- Norton, B., & De Costa, P. I. (2018). Research tasks on identity and language education. *Language Teaching*, 51, 90–112.
- Nunan, D., & Bailey, K. (2009). *Exploring second language classroom research: A comprehensive guide*. Boston: Heinle Cengage Learning.
- Ortega, L. (2015). Research synthesis. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 225–244). London: Bloomsbury.
- Paltridge, B., & Phakiti, A. (Eds.). (2010). *Continuum companion to research methods in applied linguistics*. London: Continuum.
- Paltridge, B., & Phakiti, A. (2015a). Developing a research project. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 260–278). London: Bloomsbury.
- Paltridge, B., & Phakiti, A. (Eds.). (2015b). *Research methods in applied linguistics: A practical resource*. London: Bloomsbury.
- Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, 3, 61–94.
- Pawlak, M., & Aronin, L. (2014). *Essential topics in applied linguistics and multilingualism: Studies in honor of David Singleton*. New York: Springer.
- Perry, F. L. (2012). *Research in applied linguistics: Becoming a discerning consumer* (2nd ed.). New York, NY: Routledge.
- Phakiti, A. (2014). *Experimental research methods in language learning*. London: Bloomsbury.
- Phakiti, A., & Paltridge, B. (2015). Approaches and methods in applied linguistics research. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 5–25). London: Bloomsbury.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35, 655–687.
- Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, 98, 450–470.
- Plonsky, L., Egbert, J., & LaFlair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, 36, 591–610.
- Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, 61, 325–366.
- Plonsky, L., & Ghanbar, H. (in press). *Multiple regression in L2 research: A methodological synthesis and meta-analysis of R<sup>2</sup> values*. Manuscript under review.
- Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). New York, NY: Routledge.
- Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, 39, 579–592.

- Richards, K., Ross, S., & Seedhouse, P. (2012). *Research methods for applied language studies: An advanced resource book for students*. New York and London: Routledge.
- Schmitt, N. (2002). *An introduction to applied linguistics*. London: Arnold.
- Simpson, J. (Ed.). (2011). *The Routledge handbook of applied linguistics*. New York: Routledge.
- Starfield, S. (2015). Ethnographic research. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 137–152). London: Bloomsbury.
- Sterling, S., & Gass, S. (2017). Exploring the boundaries of research ethics: Perceptions of ethics and ethical behaviors in applied linguistics research. *System*, 70, 50–62.
- Tao, J., Shao, G., & Gao, X. (2017). Ethics-related practices in internet-based applied linguistics research. *Applied Linguistics Review*, 8, 321–354.
- Ziegler, N. (2016). Methodological practices in interaction in synchronous computer mediated communication: A synthetic approach. In A. Mackey & E. Marsden (Eds.), *Instruments for research into second languages: Empirical studies advancing methodology* (pp. 197–223). New York, NY: Routledge.



# 2

## Habits of Mind: How Do We Know What We Know?

Richard F. Young

As the contributions to this Handbook attest, methods of inquiry that applied linguists employ differ greatly. Learning and using language is a vast and immensely complex undertaking and any method of investigating it inevitably involves researchers attending to some aspects while disattending to others. The ways that different researchers attend to different aspects of language learning and use are habits of mind grounded in the communities to which they belong. One way of understanding these habits is through the sociology of science pioneered by Ludwik Fleck. According to Fleck (1979/1935), researchers are members of *thought-collectives* with certain habits of mind that direct researchers' attention to assimilate what they perceive into what Fleck called a *thought-style*. In this chapter, I use Fleck's three characteristics of thought-collectives—their rhetoric, their epistemology, and incommensurability among thought-collectives—to consider different methodologies of applied linguistic research and to describe how different habits of mind constrain how we know what we know. I conclude that, though incommensurability exists between certain thought-collectives in applied linguistics, several researchers have argued strongly for complementarity and I relate experiences of individual researchers whose thought-styles have been changed by attending to new data.

---

R. F. Young (✉)

School of Foreign Languages, Central China Normal University, Wuhan, China  
e-mail: [ryoung@mail.ccnu.edu.cn](mailto:ryoung@mail.ccnu.edu.cn)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_2](https://doi.org/10.1057/978-1-137-59900-1_2)



## Attending and Disattending

Language learning and use are rich and complex phenomena that all of us have experienced, sometimes as students, sometimes as teachers. Because of our experiences when we come to look at language learning in classrooms, we tend to focus our attention on some feature of that rich complexity that interests us. Some of that complexity is represented in a recent volume (Markee, 2015), in which contributors focus their attention on one or more parts of the complex whole. Among those contributors, critical theorists attend to connections between what goes on in classrooms and social, political, and ideological processes in the world outside, while conversation analysts attend to how social and institutional realities are constructed and realized in classroom talk. In other traditions, researchers choose to focus on language learning that can occur in classrooms but disagree about the nature of learning: Is learning a language socialization to a society of which the classroom is a part? Is it the development of independent abilities that begin as cooperative action? Or is it the development of knowledge through interaction with others? Taken together, these different traditions paint a collage of classroom discourse, an image made up of different representations, in which researchers from one tradition attend to one representation of classroom discourse and inevitably disregard others.

I have framed this chapter with a question: How do we know what we know about research approaches and methodology in applied linguistics? One answer to that question is we know what we attend to, and what we attend to is grounded in what Janesick (2011) called different “habits of mind”—our personal preferences as researchers and the early training we received. These “habits of mind” include observation habits, interview habits, writing habits, the habits of analysis and interpretation, as well as the habit of collaborating with others. In a recent discussion of research methods in applied linguistics (Young, 2014), I argued that the system of historical philosophy and sociology of science developed by Ludwik Fleck is of cardinal importance in understanding habits of mind: methods adopted by different schools of research in applied linguistics.<sup>1</sup> According to Fleck, what researchers attend to is the product of a method of inquiry that directs their attention to assimilate what they perceive into what Fleck (1979/1935) called a *thought-style*. A thought-style directs researchers’ attention to certain features of the complexity of interaction and away from others, a perceptual bias that Goffman (1974) described as attention distributed along three tracks: a main-line, story-line track that is the official focus of attention; a directional track that provides

organization for the main-line track; and a disattend track consisting of events that are officially treated as irrelevant to the activity in progress.

A predominant thought-style in applied linguistic research focuses attention on the main-line, story-line track; that is, observables that can be recorded in field notes, recorded on tape, or written in a transcript. Yet, every moment of language learning and use is not a world to itself and cannot be captured in its entirety in text, sound, or image. In the world beyond classroom conversations, the personal dispositions of students and teachers accumulated over a lifetime and sometimes longer, the power they exert and resist in their most mundane of actions, and the institutional constraints and affordances of school and public policy are contingent interactions of social life that provide organization for the main-line track, yet the observable moments of language learning, use, or assessment, for instance, are the empirical location in which these social patterns and political warrants exist and are transformed. In this chapter, I wish to argue that attending to one aspect of language learning and use while disattending to other aspects that are treated as irrelevant is what principally distinguishes contemporary methods of research in applied linguistics.

A recent description of how researchers' attention is directed to the main-line, story-line track and away from what is happening elsewhere is provided by Hall's (2004) description of *practicing speaking* in Spanish. In her study, Hall observed a teacher and a high school class of students in the United States who were studying Spanish for the first time. Hall termed the observable activity in the classroom community *practicing speaking*, which she described as follows:

[T]he teacher [...] took primary responsibility for beginning the activity, orchestrating its development and ending the activity. She provided most of the talk, decided what counted as a relevant topic and a relevant comment on the topic, and moved the interaction from student to student and contribution to contribution, deciding which communicative behaviors of the students were to count as part of the performance. (pp. 73–74)

Hall made a transcript of the interaction between the teachers and students, part of which is reproduced in Fig. 2.1. According to Hall, the transcript is of an interaction between teacher and students that typifies the practice.

It is common to see the activities represented in Fig. 2.1 as an instance of classroom discourse, but this representation leads us to attend to one slice of the reality of the classroom. By representing classroom discourse in

- T: Ok aquí tenemos Coca Cola tenemos Pepsi Cola tenemos  
Fresca cómo se llaman  
*(ok here we have Coca Cola we have Pepsi Cola we have  
Fresca what are they called)*
- S1: Budweiser
- T: Refresco refresco sí se llaman refrescos sí y aquí hay  
refresco. Rápidamente qué es esto  
*(soft drink soft drink yes they're called soft drinks yes  
and here there is a soft drink. Quickly, what is this?)*
- S2: O::u
- S1: Uhum ice cream.
- T: Helado muy bien señor el helado  
*(ice cream very good sir ice cream)*
- S1: Helado  
*(ice cream)*

**Fig. 2.1** "Practicing speaking" in Spanish (Hall, 2004, p. 76)

a transcript, Hall entextualized one kind of talk and in doing so she constrained what researchers know about what is going in the classroom. The process by which a speaking practice is transcribed and thus becomes represented in text has been studied by anthropological linguists. It was first identified by Ochs (1979) and was described by ethnographers of speaking such as Haviland (1996) and Urban (1996), who argued that entextualization involves selecting and repurposing discourse for a new audience. It is possible, however, to see beyond what an entextualized transcript selects by expanding our focus of attention from the main-line track to what is often disregarded at the margins. In fact, Hall herself expanded her description of the class by telling us what was happening on the disattend track:

As the teacher talked to the large group, the students often interacted with their nearby seat mates. While there was occasional use of Spanish in these whispered side sequences, English was the predominant code. As long as the students interacted with each other quietly, and were not major distractions to the larger official instructional discourse, the teacher did not prevent them from doing so. Thus, there was usually a steady hum of background talk among the students throughout each class meeting. (p. 74)

A transcript of “the larger official instructional discourse” is an entextualization of an ongoing lived social process. A transcript is talk ripped from its original social, historical, and physical environment, yet there is a fount of meaning in interaction that cannot be found in a transcript—a rich social semiotic of historical and political processes extending beyond the time and place of interaction. Some of those processes are indexed by Hall as the “steady hum of background talk among the students.” Applying Fleck’s notion of thought-style to research methods for understanding classroom discourse, I argue that a transcript of “official instructional discourse” directs researchers’ attention to certain observable classroom phenomena in which other observations and interpretations are either ignored or assimilated.

A thought-style characterizes what Fleck called a *thought-collective*: a community of persons mutually exchanging ideas and maintaining intellectual interaction, and within this thought-collective a thought-style is a “picture, which is visible only to anybody who takes part in this social activity, or a thought which is also clear to members of the collective only” (Fleck, 1986a/1935, p. 77). Fleck described differences among research methodologies in terms of three differences in thought-style. First, Fleck recognized that the rhetoric used by members of a thought-collective to represent what appears to them as reality differs from one thought-collective to another. In a second difference among thought-styles, the questions that members of a thought-collective ask arise from the thought-style within the thought-collective; in other words, what counts as knowledge and how it relates to truth, belief, and justification are constructs of their thought-style. The third characteristic of thought-styles, according to Fleck, is that they are incommensurable: What researchers within one thought-collective see, researchers working within a different thought-collective do not see and thus think differently about phenomena. As Fleck (1979/1935) wrote, “The principles of [a different thought-collective] are, if noticed at all, felt to be arbitrary and their possible legitimacy as begging the question. The alien way of thought seems like mysticism. The questions it rejects will often be regarded as the most important ones, its explanations as proving nothing or as missing the point, its problems as often unimportant or meaningless trivialities” (p. 109).

The question of what a researcher will attend to in the complex and multiple phenomena of language learning and use is not an individual decision because the decision is a social one. As Fleck (1986a/1935) wrote,

A truly isolated investigator is impossible [...]. An isolated investigator without bias and tradition, without forces of mental society acting upon him, and without the effect of the evolution of that society, would be blind and thoughtless.

Thinking is a collective activity [...]. Its product is a certain picture, which is visible only to anybody who takes part in this social activity, or a thought which is also clear to the members of the collective only. What we do think and how we do see depends on the thought-collective to which we belong. (p. 77)

The sociology of scientific inquiry developed by Fleck may be unfamiliar to some, so to put his approach on a more recognizable footing, I include here a reflection on my own experience as a beginning researcher to illustrate how one person's research may be reinterpreted. Reflecting on Fleck's interpretation of cognition as a collective activity, I reconsider my own early research in second language acquisition (SLA) in which I claimed to have discovered how interaction among social and linguistic factors influences interlanguage variation. This was research I did for my dissertation at the University of Pennsylvania and eventually published (Young, 1991). However, when describing my research as a social activity, I cannot say that I alone discovered how interaction among social and linguistic factors influences interlanguage variation; instead, I must admit that my research was in fact a collective activity, influenced by the discovery by others of systematic variation in the thought-style of mathematical approaches to language variation during the 1980s. In other words, my own thought-style was influenced by the thought-collective of linguists led by William Labov at the University of Pennsylvania in the 1980s.

A thought-collective is defined by Fleck (1979/1935) as "a community of persons mutually exchanging ideas or maintaining intellectual interaction" (p. 39), and when a thought-collective matures, it breaks into a small esoteric professional circle—a small group of specialists who contribute to development of the thought-collective—and a wider exoteric circle of those members who are influenced by the thought-style but do not play an active role in its development. In the case of quantitative variation theory, the esoteric circle consisted of those members of the collective (including me) who adopted and used the categorical analysis software specifically designed for analysis of language variation known as VARBRUL (Pintzuk, 1988) or GoldVarb (Rand & Sankoff, 1990) to perform quantitative analysis of language variation. While this software was widely known among linguists, colleagues in applied linguistics who were part of the thought-collective (in the sense that they felt that quantitative variation theory could be applied to understand variation in interlanguage) had access to the appropriate thought-style (i.e., how to use VARBRUL or GoldVarb) through pedagogical publications and conference presentations by members of the esoteric circle such as Young and Bayley (1996).

In order to explicate differences among research methodologies in applied linguistic research, in the following pages, I use Fleck's three characteristics of thought-collectives, which I will term *rhetoric*, *epistemology*, and *commensurability*, to characterize different applied linguistic methodologies as thought-styles and consider the ways in which they are commensurable. First, I consider the rhetoric that members of a thought-collective use to describe what appears to them as reality, to discuss ideas, and by publishing their research to persuade others.

## Rhetoric

In Fleck's sociology of knowledge, the use of technical terms and specific technology by members of a thought-collective not only denotes something determined by their definitions and use but also encompasses "a certain specific power, being not only a name but also a slogan," which has "a specific thought-charm" (Fleck, 1986b/1936, pp. 99–100) and is thus a means of persuading others. In order to see the role that the rhetoric of presentation plays in creating and maintaining a thought-style within a thought-collective, consider, for example, the means that are adopted by three popular research methods in applied linguistics: quantitative quasi-experimental research, conversation analysis, and corpus linguistics.

In applied linguistics, the rhetorical power of language is perhaps most readily recognized in the numerical language of quantitative research which, as Henning (1986) described it, is "the kind of research that involves the tallying, manipulation, or systematic aggregation of quantities of data" (p. 702). The process of quantitative research involves quantifying phenomena, representing quantities as numbers, manipulating those numbers in order to test hypotheses, and supporting the results of that process with numbers. The thought-style and language of quantitative research was characterized by Talmy (2014), who recognized that within a thought-collective whose members attend to linguistic and cognitive issues in applied linguistic research, quantitative research is the predominant thought-style (Lazaraton, 2005; Loewen & Gass, 2009; Plonsky, 2013). For beginning researchers, too, I have found that numerical methods involving statistical analyses of data from tests, surveys, experiments, and language corpora appeal to students with a background in quantitative sociology or psychology, while qualitative research methods including case study, ethnography, grounded theory, and narrative inquiry appeal to students with undergraduate majors in the humanities. For

these students, as J. D. Brown (1991, 1992) recognized, statistics is indeed a foreign language.

In quite a different thought-collective, Schegloff (1993) argued eloquently that conversation analysts eschew quantification. In talk-in-interaction, counting requires that what is counted (e.g., laughter, response tokens, or continuers) occur in “an environment of relevant possible occurrence” though “not every place that something may not be found is a place at which it is missing” (p. 106). In addition, when *analysts* count what they perceive to be the same thing in talk-in-interaction, they are not counting instances of what are necessarily the same thing to *participants* in the talk they have chosen to analyze. Schegloff concluded that in the present state of conversation analysis, quantification is premature because “We need to know what the phenomena are, how they are organized, and how they are related to each other as a pre-condition for cogently bringing methods of quantitative analysis to bear on them” (p. 114).

As a thought-collective, therefore, conversation analysts eschew the statistical language of quantitative research prevalent in those communities who attend to language learning and use as a cognitive and linguistic phenomenon. A second major difference between the language of CA and that of other thought-collectives is the way in which talk-in-interaction is represented in written form in transcription. Other applied linguistic research methods look beyond the transcript to analyze talk-in-interaction in a wider context and foreground the social semiotic of historical and political processes extending beyond the time and place of interaction and the means by which dominant power and ideology is reproduced, resisted, or transformed. For, as Wortham (2001) has written, “contingent social interactions are the empirical location in which broader theories and social patterns exist and get transformed” (p. 257). Conversation analysis, however, is radically empirical in limiting analysis to what is observed in talk-in-interaction and arguing for any analysis on the basis of *what* is said and *how* it is said. Kasper and Wagner (2014) detail the ways in which CA transcriptions of speech delivery have evolved since the early conventions summarized in Jefferson (1984) to represent talk-in-interaction in all its ecological situatedness. The level of detail required in representing the situatedness of talk-in-interaction is very high because, as Schegloff (2000) wrote,

The level of granularity at which noticing is done matters not only for the social actors being studied, but for us as investigators as well; so too at what level the observed or noticed world is being described. [...] Knowing how granularity works matters then not just substantively, but methodologically (p. 719)



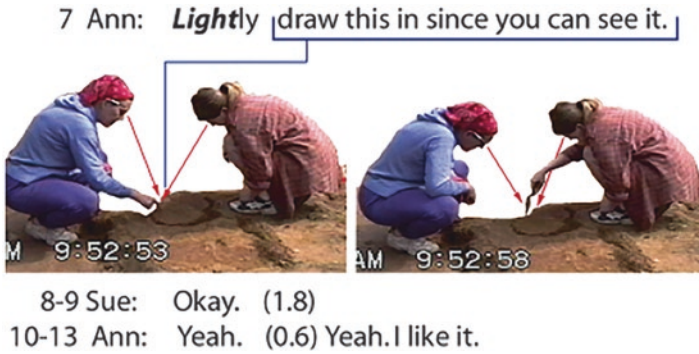


Fig. 2.2 Representing embodied cognition (Goodwin, 2003, Fig. 2.9, p. 35)

The most recent development in representing talk-in-interaction in writing is described by Kasper and Wagner as representation of movement and embodiment—how cognition emerges through the interaction of brain, body, and world. Some early representations of embodiment on the printed page are in the work of Charles Goodwin, which is illustrated in Fig. 2.2.

In Fig. 2.2, two anthropologists, Ann and Sue, are working at a dig and work together using their hands, eyes, and a trowel to investigate a marking of interest in the dirt. Goodwin describes their actions in prose as follows:

In line 7, Ann uses a symbiotic gesture, tracing a circular path over a particular place in the dirt, to show Sue where to draw. In response Sue brings her trowel to exactly where Ann had been pointing (in fact there is almost a collision between Sue's trowel and Ann's finger) and starts to draw [...]. By drawing where she does Sue not only demonstrates that she has seen Ann's gesture, and taken it into account, but that the gesture is in fact the point of departure for her response to Ann (which is done through the action of drawing the outline of the feature rather than talk). Their hands exchange places over the same spot in the dirt. Then as Sue draws, Ann produces talk that is responsive to what she sees Ann doing. (pp. 34–35)

In Fleck's system, the physical representation of embodied cognition in Goodwin's work is a language that is particular to the thought-style that characterizes conversation analysis; in other words, it is a language to describe what appears to members of the thought-collective as reality. In comparison, what appears as reality to some applied linguistic researchers differs from the reality represented in a CA transcript and instead they choose to limit their attention to the words of a text. Researchers in corpus linguistics, for example, first annotate the words in a corpus by structural markup, tagging of words



according to lexical category (POS tagging), and parsing. Although in one approach to corpus linguistics, the words in a corpus should be minimally annotated in order for the corpus data to speak for themselves. Using the annotated corpus, researchers abstract patterns in the data that correspond to areas of interest and further analyze those patterns to glean information about language use in such fields as literary studies, hate speech, political argumentation, discourse within the professions, and contrastive studies of two or more languages. The rhetoric of corpus linguistics is thus grounded in analysis of words within a context of other words and involves either the collection or use of extensive samples of “real-world” text.

The language of numbers in which applied linguists describe linguistic and cognitive phenomena differs from the written representation of embodied cognition employed in conversation analysis, while researchers interested in discovering lexical patterns obtaining over large samples of text use the rhetoric of corpus linguistics, which differs from the rhetoric of CA. And yet, members of all three thought-collectives publish their research in journals within the broad field of applied linguistics; they are all concerned with understanding the process of human language learning and use. The thought-collectives are self-maintaining through their use of technical terms and specific technology; by publishing their research at conferences and in academic journals they establish and maintain a rhetoric that binds members of a thought-collective together and distinguishes them from others. Training in any of these communities involves socialization of novices to ways of speaking and writing about reality specific to a thought-style, what Fleck called the slogans and the thought-charms of the community.

## Epistemology

A second difference among thought-collectives is what counts as knowledge in a thought-collective and how this relates to truth, belief, and justification that build a thought-style characteristic of the community. This is a difference in what I have termed epistemology but, for Fleck, knowledge does not originate in an individual because during every lasting exchange of ideas in conversation or in print ideas and beliefs originate and develop that are not associated with an individual, and the knowledge that an individual has originates from the thought-collective to which the individual belongs. According to Fleck, knowledge has three components: It is a dialectic relation among an individual, the object of that individual’s attention, and the thought-collective within which the individual acts. What an individual researcher attends to and thus

what an individual knows is the product of a thought-style originating within a given community. It does not result from an individual's spontaneous ideas—there is no such thing as a Eureka moment—instead, concepts arise from tacit assumptions within the thought-collective which are not of scientific origin.

In recognizing different approaches to education research, Sfard (1998) took a similar position when she described different tacit assumptions as *metaphors for learning*, and she stressed their importance in guiding scientific investigation:

[O]ne has to reach the most fundamental, primary levels of our thinking and bring to the open the tacit assumptions and beliefs that guide us. This means digging out the metaphors that underlie both our spontaneous everyday conceptions and scientific theorizing. Indeed, metaphors are the most primitive, most elusive, and yet amazingly informative objects of analysis. Their special power stems from the fact that they often cross the borders between the spontaneous and the scientific, between the intuitive and the formal. Conveyed through language from one domain to another, they enable conceptual osmosis between everyday and scientific discourses, letting our primary intuition shape scientific ideas and the formal conceptions feed back into the intuition. (p. 4)

Sfard (1998) proposed two metaphors for learning: an acquisition metaphor and a participation metaphor. In conceiving of learning through the metaphor of acquisition, learning is seen as accumulating basic units of knowledge, which are gradually refined and combined to form ever richer cognitive structures. As she wrote, “This approach, which today seems natural and self-evident, brings to mind the activity of accumulating material goods” (p. 5). In contrast, Sfard wrote that in recent educational theories the permanence of getting and *having* in the acquisition metaphor is replaced by the constant flux of *doing*, and learning is conceived as a process of becoming a member of a community. In other words, the same activity of learning—one with which we are all familiar—is understood in two very different ways when researchers understand learning by considering the process of learning as either the accumulation of material goods or as negotiating membership in a new community. These two metaphors for learning (which Fleck would call *proto-ideas*) influence how researchers establish truth, communicate belief, and justify their claims.

The metaphor of learning as acquisition is clearly represented in the name of one of the branches in our field—second language acquisition—and inspired many of the early studies of SLA, such as the order of morpheme

acquisition studies reviewed in Krashen (1977), which were based on R. Brown's (1973) description of early child first language acquisition. In contrast, the metaphor of learning as participation clearly inspired Young and Miller's (2004) longitudinal study of the changing discourse roles of a student and an instructor in a series of ESL writing conferences.

Turning from metaphors of learning, it is possible to see similar contrasting proto-ideas in ways that different thought-collectives understand the nature of *language*. William Hanks (1996) in his treatise *Language and Communicative Practices* distinguished two kinds of knowledge held by those who study language. The first, which Hanks called the *irreducibility thesis*, is characteristic of theoretical linguistics in which language is considered to be a phenomenon *sui generis* that can be studied and understood independently of how language is used. In Hanks's words, "language cannot be explained by appeals to nonlinguistic behavior, to emotion, desire, psychology, rationality, strategy, social structure, or indeed any other phenomenon outside the linguistic fact itself" (p. 6). The second kind of knowledge of language, which Hanks called the *relationality thesis*, is more typical of research in the humanities and social sciences, in which language is understood as a social phenomenon that is inextricably part of the world of action. Working within the relationality thesis, researchers attend to "the actual forms of talk under historically specific circumstances: not what could be said under all imaginable conditions but what is said under given ones" (p. 7). Both of these concepts are important influences on contemporary applied linguistics, but, as Hanks noted, their proponents "have become increasingly distant and disinterested in talking to one another" (p. 10).

The distinction in applied linguistics between researchers who adopt the irreducibility thesis and those who embrace the relationality thesis is clear, though some researchers have recognized (with Hanks) that you cannot have one without the other. For example, Gee distinguished two ways of knowing about language. In one way, language is primarily a tool for saying things, for giving information, while in another sense language is a tool for accomplishing three activities: saying, doing, and being. Gee (2015) distinguished between these different functions of language by referring to them as two kinds of discourse—Discourse with a big "D" or discourse with a small "d"—which he explained as follows.

The notion of "Big 'D' Discourse" ("Discourse" spelled with a capital "D") is meant to capture the ways in which people enact and recognize socially and historically significant identities or "kinds of people" through well-integrated combinations of language, actions, interactions, objects, tools, technologies, beliefs, and values. The notion stresses how "discourse" (language in use among

people) is always also a “conversation” among different historically formed Discourses (that is, a “conversation” among different socially and historically significant kinds of people or social groups). The notion of “Big ‘D’ Discourse” sets a larger context for the analysis of “discourse” (with a little “d”), that is, the analysis of language in use.

Gee would agree with Hanks that the notion of language as irreducible is the basis for analyzing how language establishes relations among speakers and historically specific circumstances of speech for, in Gee’s words:

Little ‘d’ discourse analysis studies how the flow of language-in-use across time and the patterns and connections across this flow of language make sense and guide in interpretation. Big ‘D’ Discourse analysis embeds little ‘d’ discourse analysis into the ways in which language melds with bodies and things to create society and history. (p. 420)

These epistemological distinctions characterize thought-styles in different communities of language researchers, and they also are associated with distinct ways of obtaining new knowledge. Sfard (1998) applied her acquisition and participation metaphors to describe how researchers understand learning; Hanks (1996) put forward the theses of irreducibility and relationality to represent two general positions on the relation between meaning and context, and Gee (2015) distinguished between big “D” and little “d” discourse also in order to relate meaning to context—the context of language in use and the larger context of interaction among different socially and historically significant kinds of people or social groups. The positions articulated by these three scholars are all ways in which the central concern of applied linguists—what we know about language learning and use—are represented. But in any thought-collective knowledge of language, language use, and language learning does not stand still and the question remains of whether and how methods that researchers use to further their understanding can be related to the epistemological questions that these scholars ask. Ellis (2014) has provided an attempt at an exhaustive list of research methods that applied linguists have used in order to increase their knowledge of language development and language use. He has argued that different research questions require different research methods, thus all the methods listed embody useful techniques which are useful for understanding different phenomenon:

Language learning diaries, functional magnetic resonance imaging (fMRI) scanners, analyses of service interactions using conversational analysis (CA), introspection, visual world eye-tracking, classroom interaction recordings, computer

simulations, artificial grammar learning experiments, billion-word corpora of usage, questionnaires, dynamic assessment, event-related potentials (ERPs), think-alouds, feedback manipulations, *n*-back tasks, psychometric batteries, error analysis, longitudinal corpora, laboratory experiments, classroom field experiments, ethnographic research, agent-based modeling, contrastive analysis, brain connectivity analysis, dynamic systems analysis, behavioral genetics, idiographic and nomothetic approaches, thick and thin descriptions, emic and etic approaches (p. 398)

There is, however, no clear way in which to associate each of the methods in Ellis's (2014) list with a specific epistemology, though it sometimes appears that certain methods are more frequently used to answer research questions that fall on one side or the other of Sford's, Hanks's, and Gee's dichotomies. For example, Toohey (2000) called her investigation of the process of becoming an ESL learner in an early childhood program for immigrant children an ethnography, and she described some children's processes of socialization toward a school identity while others were excluded—a study that better represents Sford's metaphor of participation than the metaphor of acquisition. An example of a research method grounded in Hanks's irreducibility thesis is Steinhauer (2014). Working within the critical period hypothesis, Steinhauer published a review of findings of studies of event-related brain potential (ERP) studies in second language (L2) morpho-syntax in order to relate inferred information about brain activity of second language learners to their age of acquisition and their L2 proficiency. After reviewing decades of studies, Steinhauer concluded that more recent studies “demonstrate quite systematic neurocognitive changes as a function of (individual) L2 proficiency and show that ERPs of late L2 learners at very high levels of L2 proficiency are usually indistinguishable from those of native speakers” (p. 413), thus challenging the critical period hypothesis. Finally, a study based on Gee's distinction between big “D” Discourse and little “d” discourse is Kachru's (1985) comparative discourse analyses of the pragmatics of institutionalized non-native Englishes and second language acquisition. Referring to corpora of L2 English and institutionalized non-native varieties of English, Kachru discussed different approaches to discourse analysis and contrasted the little “d” approach she believed to be prevalent in SLA studies at the time with the discourse of non-native varieties of English in which the relevance of shared sociocultural conventions (i.e., big “D” Discourse) is highlighted in successful linguistic interactions.

In the preceding pages, I have presented Fleck's theory of knowledge and how it relates to truth, belief, and justification arguing that individual knowledge is a dialectic among the individual, the object of attention, and the

thought-collective within which the individual acts. Within a thought-collective, knowledge of the vast and vastly complex undertaking that is language learning and use originates in knowledge of simpler everyday phenomena: Is learning like acquiring stuff or is it like doing things? Is language a unique object in and of itself or is it part and parcel of all human activities past and present? Should we limit our attention to the immediate context of language learning and use, or should we be expanding our attention to consider the many ways in which—through language—society and individuals create, maintain, and challenge power and identity? Different thought-collectives provide different answers to these basic questions, and individual researchers direct their attention to phenomena, ask their research questions, and design their research within the knowledge framework that is accepted within their thought-collective. The thought-style of individuals within the community includes meanings of words and ideas that are thought of as “objective,” they use research instruments which direct their attention on the tracks of the same style, and when incoherent facts are noticed, they are ignored as unimportant. But there are nonetheless moments when facts can no longer be ignored and, when individuals with different thought-styles interact, they may find the others to be fools or heretics but they may just as well find that the interpretation of facts in the other community deserves their attention. The degree to which thought-styles change and evolve depends on the commensurability of the thought-styles in different thought-collectives.

## **(In)commensurability**

In the introduction to their compendium of approaches to applied linguistic research, Hulstijn, Young, and Ortega (2014) expressed concern that two predominant thought-styles in applied linguistics were incommensurable, differing from each other in the nature of the phenomenon being studied and the means used to study it:

Learning a second language (L2) and developing pedagogy on the basis of what is known about learning are complex endeavors, and research into these areas has taken several different paths. One path is that taken by researchers investigating linguistic-cognitive issues, who pursue objectivity with quantitative research methods, often with the help of inferential statistics. Another path is taken by researchers who consider the social context of activity as a cardinal feature of human knowledge and thus of learning and teaching and who employ qualitative research methods such as case study and ethnography. These are only two of many paths that researchers have taken, but they are perhaps the two that

can be most easily discerned and most conveniently contrasted. Do these two paths lie on either side of the stream of knowledge, or is it possible that at one or more bends in the stream there is a bridge by which a researcher may cross to the other side? Pursuing this trope leads to the following question: Is there a gap in research in L2 learning and teaching? And if there is, is the gap ontological (i.e., what is the phenomenon we should be studying?), epistemological (i.e., how should we be studying it?), or both? (p. 362)

Fleck expressed the same concern in his discussion of the incommensurability of thought-styles. Sady (2012) went to some length to infer Fleck's ideas on incommensurability and truth from Fleck's own writings and from Kuhn's (1996) development of Fleck's ideas in *The Structure of Scientific Revolutions*. The incommensurability that can be perceived between researchers working within different thought-collectives is recognized by Fleck as three differences in thought-style. First, echoing Wittgenstein's (2001) concept of the language game, Fleck recognized that the language used by members of a thought-collective to describe what appears to them as reality differs from one thought-collective to another. For example, the term "second language learning" has different meanings according to whether SLA researchers consider learning as acquisition—the incremental accumulation of concepts that are gradually refined and combined with other concepts to form cognitive structures—or, alternatively, whether they consider learning as participation, movement along trajectories of changing engagement in discursive practices. Second, the questions that members of a thought-collective ask arise from the thought-style within the thought-collective; in other words, the phenomena to which members of a thought-collective attend are constructs of their thought-style. The third characteristic of incommensurable thought-styles, according to Fleck, is the difference in perceptions: What researchers within one thought-style see, people working within a different thought-style do not see and thus think differently about phenomena. As Fleck (1979/1935) wrote,

The principles of [a different thought-collective] are, if noticed at all, felt to be arbitrary and their possible legitimacy as begging the question. The alien way of thought seems like mysticism. The questions it rejects will often be regarded as the most important ones, its explanations as proving nothing or as missing the point, its problems as often unimportant or meaningless trivialities. (p. 109)

In investigating L2 learning and use, though researchers in different thought-collectives study the same phenomenon—bilingual cognition—some see it as neural activation at the interface between the cortex and the



limbic system, while others see the genesis of cognition in cultural artifacts, social interaction, and the socioeconomic environment of individuals. Within these different thought-collectives, different thought-styles envisage the nature of bilingual cognition in their own and very different ways.

Nonetheless, the degree to which habits of mind can change depends on the commensurability of thought-styles in different thought-collectives and even the personal experiences of researchers. The way that different researchers attend to different aspects of second language learning and even how the same researcher comes to attend to different aspects have been described by researchers who introspect on their own research process. Mackey (2014), for instance, recounted a shift in her own attention to the nature of interaction involving language learners. From a research posture of disattention to the social factors underlying interaction, in Philp and Mackey (2010), Mackey refocused her attention to consider the relationships among learners and how they impacted what learners were willing and able to listen and attend to and thus what they produced. Mackey's change of focus from the details of classroom interaction to the social context in which interaction occurs parallels the recognition by several authors of the different roles for the human subject in linguistic experiments. Should researchers focus narrowly on the phenomenon of interest or should they expand the focus of their attention to include their subjects' social and political roles with other subjects, with the researcher, and with the society beyond the walls of the classroom? Another example of change in thought-style occurred among researchers in reading and metaphor. In her early studies of the cognitive process of reading in a foreign language, Cavalcanti (1983) found that her subjects not only reported their thoughts about reading but framed them in the social context of their roles as subjects in a research study. In more recent applied linguistic studies of metaphor, which earlier studies had interpreted narrowly as an index of cognition alone, Zanotto, Cameron, and Cavalcanti (2008) argued that metaphor must be understood as social and situated, not merely a reflex of thought. Reflection by these investigators on interaction, reading, and metaphor revealed the importance of attending to the social contexts of human subjects in applied linguistic and SLA studies because they index how subjects position themselves with respect to their social and political roles with other subjects, with the researcher, and with the society beyond the walls of the classroom or laboratory.

The problem of incommensurability among thought-collectives was also addressed by the contributors to Hulstijn et al. (2014), who generally agreed that there is no single, monolithic social-cognitive gap in applied linguistic research. The contributors succeeded in disentangling various independent



contrasts in ontology, epistemology, and methodology, distinguishing a range of views rather than a simple dichotomy. Among those proponents of dichotomies previously reviewed, there is also a sense that differing positions are more often complementary than oppositional. Both Sfard (1998) and Hanks (1996) warn of the dangers of considering their dichotomies as incommensurable. Sfard titled her essay, “On two metaphors for learning and on the dangers of choosing just one,” and in O’Connor’s (2000) review of Hanks’s book, he argued that Hanks developed a position in which the tension between relationality and irreducibility reflects a practical tension in communicative practice, concluding that “although linguistic systems are governed in part by principles unique to language, grammar is neither self-contained nor entirely independent from the social worlds in which individual languages exist. Modes of speaking have an impact on and are influenced by linguistic structure” (p. 229). And Gee (2015), as cited above, also admits that “Big ‘D’ Discourse analysis embeds little ‘d’ discourse analysis into the ways in which language melds with bodies and things to create society and history” (p. 420).

The habits of mind—the phenomena to which researchers attend and how they analyze and interpret them—characterize a thought-style. But neither thought-styles nor thought-collectives are immutable, and the examples of perspectival change reported by Mackey, by Cavalcanti, and by Zanotto et al. show how changes in the thought-styles of individuals index historical development within a particular thought-collective. What this implies is that instead of looking at the differences among social approaches and cognitive approaches to SLA (the thought-collectives), it is more fruitful to consider the thought-styles of individual investigators, their activities, their habits of mind, and how they position themselves within the social processes of investigation and publication. As Janesick (2011) has suggested, habits of mind need to be exercised, they need to be “stretched” on a very frequent basis to overcome incommensurability between thought-styles.

## Conclusion

I framed this chapter with a question: How do we know what we know about research approaches and methodology in applied linguistics? The answer I proposed is we know what we attend to and the habits of mind of researchers—their personal preferences as researchers and the early training they received—to a large extent determine the questions researchers ask, the design and implementation of research studies, and the way data are interpreted. Those same habits of mind leave in the shadows what is not attended to whether by design or, more likely, incidentally. Within the sociology of

science put forward by Fleck, habits of mind do not arise from individual volition or from the cognition of an isolated researcher. Instead they are aspects of a thought-style within a community of like-minded researchers, becoming a member of which involves socialization of the beginning researcher to the rhetoric, the truths, the beliefs, and the justifications that comprise the thought-style of the community. At the same time, membership in one thought-collective implies, either openly or covertly, rejection of the language and knowledge prized by communities that are considered incommensurable with one's own.

Fleck's analysis of the sociology of science is, in its way, pessimistic. If researchers in a given thought-collective do not attend to facts outside their own immediate field of attention and then express their findings in language that is incomprehensible to researchers in other communities, how does the field of applied linguistics as a whole progress? Although, according to Goffman (1974), individuals in interaction maintain a main story line of activity, there are also channels of subordinated activity to which participants disattend, yet disattention does not imply lack of awareness: Not attending to activity does not imply ignorance. As Hall (2004) showed in her description of "Practicing Speaking" in Spanish, attention to the main story line of instructional discourse does not mean that the researcher ignores what is going on at the sidelines. Examples of changing personal thought-styles by Mackey, by Cavalcanti, and by Zanotto et al. demonstrate that there are occasions when the disattend track includes data that may bring about a reorganization of researchers' knowledge of the field and thus expansion of the thought-collective. These are not the scientific revolutions of which Kuhn (1996) wrote but gradual changes which build bridges between thought-collectives. Indeed, if that is to happen, if research methodologies and approaches in applied linguistics are to progress, perhaps what researchers need to do from time to time are the mental stretching exercises that Janesick (2011) advocated.

## Resources for Further Reading

Fleck, L. (1979/1935). *Genesis and development of a scientific fact* (F. Bradley & T. J. Trenn, Trans.). Chicago: University of Chicago Press.

Originally published in German in 1935, this monograph anticipated solutions to problems of scientific progress, the truth of scientific fact, and the role of error in science now associated with the work of Thomas Kuhn and others. Arguing that every scientific concept and theory—including his

own—is culturally conditioned, Fleck was appreciably ahead of his time. And as Kuhn observes in his foreword, “Though much has occurred since its publication, it remains a brilliant and largely unexploited resource.”

Hulstijn, J. H., Young, R. F., & Ortega, L. (2014). Bridging the gap: Cognitive and social approaches to research in second language learning and teaching. *Studies in Second Language Acquisition*, 36(3), 361–421.

In this widely cited article, the coeditors ask whether research in learning and teaching of a second language runs the risk of disintegrating into two irreconcilable approaches: linguistic-cognitive research using quantitative research methods and sociocultural or sociocognitive research using qualitative research methods. The article comprises nine single-authored pieces, with an introduction and a conclusion by the coeditors. The contributors advance the possibility that the approaches are not irreconcilable and that, in fact, cognitive researchers and social researchers will benefit by acknowledging insights and methods from one another.

Sady, W. (2012). Ludwik Fleck. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from <https://plato.stanford.edu/entries/fleck/>

This encyclopedia article is the best introduction in English to Fleck’s philosophy. Sady reviews Fleck’s life and works and presents a clear description and analysis of the central concepts of thought-collectives and thought-style. Sections include (a) genesis and development of a thought-style, (b) how a thought-collective transforms what is socially constructed into “reality,” and (c) incommensurability of thought-styles and the problem of truth.

**Acknowledgments** Thanks to my daughter Jenni Young for reading and correcting an early version of the chapter. Thanks to Luke Plonsky, Sue Starfield, and Aek Phakiti for helping me craft an essay that feels at home in an encyclopedia. Special thanks to Heidi Byrnes for introducing me to the philosophy of Ludwik Fleck in a discussion during Hulstijn and Young’s colloquium at the 2013 conference of the American Association for Applied Linguistics in Dallas, Texas.

## Note

1. Ludwik Fleck was a Polish and Israeli physician and biologist who, in the 1930s, developed the concept of the *thought-collective*, an important concept in the philosophy of science that helps to explain how scientific ideas endure and change over time.

## References

- Brown, J. D. (1991). Statistics as a foreign language-part 1: What to look for in reading statistical language studies. *TESOL Quarterly*, 25(4), 569–586.
- Brown, J. D. (1992). Statistics as a foreign language-part 2: More things to consider in reading statistical language studies. *TESOL Quarterly*, 26(4), 629–664.
- Brown, R. (1973). *A first language: The early stages*. Harmondsworth: Penguin.
- Cavalcanti, M. C. (1983). *The pragmatics of FL reader-text interaction: Key lexical items as source of potential reading problem*. Doctoral thesis, University of Lancaster.
- Ellis, N. C. (2014). Cognitive and social language usage. *Studies in Second Language Acquisition*, 36(3), 397–402.
- Fleck, L. (1979/1935). *Genesis and development of a scientific fact* (F. Bradley & T. J. Trenn, Trans.). Chicago: University of Chicago Press. (Original work published in German: *Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv*. Basel: Benno Schwabe).
- Fleck, L. (1986a/1935). Scientific observation and perception in general. In R. S. Cohen & T. Schnelle (Eds.), *Cognition and fact: Materials on Ludwik Fleck* (pp. 59–78). Dordrecht and Boston, MA: D. Reidel. (Original work published in Polish: O obserwacji naukowej i postrzeganiu w ogóle. *Przegląd Filozoficzny*, 38, 57–76).
- Fleck, L. (1986b/1936). The problem of epistemology. In R. S. Cohen & T. Schnelle (Eds.), *Cognition and fact: Materials on Ludwik Fleck* (pp. 79–112). Dordrecht and Boston, MA: D. Reidel. (Original work published in Polish: Zagadnienie teorii poznawania. *Przegląd Filozoficzny*, 39, 3–37).
- Gee, J. P. (2015). Discourse, small d, big D. In K. Tracy (Ed.), *The international encyclopedia of language and social interaction* (pp. 418–422). Malden, MA: Wiley Blackwell. Abstract. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9781118611463.wbielsi016/abstract>
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. New York: Harper and Row.
- Goodwin, C. (2003). The body in action. In J. Coupland & R. Gwyn (Eds.), *Discourse, the body, and identity* (pp. 19–42). Houndmills and New York: Palgrave Macmillan.
- Hall, J. K. (2004). “Practicing speaking” in Spanish: Lessons from a high school foreign language classroom. In D. Boxer & A. D. Cohen (Eds.), *Studying speaking to inform second language learning* (pp. 68–87). Clevedon: Multilingual Matters.
- Hanks, W. F. (1996). *Language and communicative practices*. Boulder, CO: Westview.
- Haviland, J. B. (1996). Text from talk in Tzotzil. In M. Silverstein & G. Urban (Eds.), *Natural histories of discourse* (pp. 45–78). Chicago: University of Chicago Press.
- Henning, G. (1986). Quantitative methods in language acquisition research. *TESOL Quarterly*, 20(4), 701–717.
- Hulstijn, J. H., Young, R. F., & Ortega, L. (2014). Bridging the gap: Cognitive and social approaches to research in second language learning and teaching. *Studies in Second Language Acquisition*, 36(3), 361–421.

- Janesick, V. J. (2011). *“Stretching” exercises for qualitative researchers* (3rd ed.). Thousand Oaks, CA: Sage.
- Jefferson, G. (1984). Transcription notation. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. ix–xvi). New York: Cambridge University Press.
- Kachru, Y. (1985). Discourse analysis, non-native Englishes and second language acquisition research. *World Englishes*, 4(2), 223–232.
- Kasper, G., & Wagner, J. (2014). Conversation analysis in applied linguistics. *Annual Review of Applied Linguistics*, 34, 171–212.
- Krashen, S. D. (1977). Some issues relating to the monitor model. In H. D. Brown, C. A. Yorio, & R. H. Crymes (Eds.), *On TESOL '77 – Teaching and learning English as a second language: Trends in research and practice* (pp. 144–158). Washington, DC: Tesol.
- Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press.
- Lazaraton, A. (2005). Quantitative research methods. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 209–224). Mahwah, NJ: Erlbaum.
- Loewen, S., & Gass, S. (2009). The use of statistics in L2 acquisition research. *Language Teaching*, 42(2), 181–196.
- Mackey, A. (2014). Exploring questions of balance in interaction research. *Studies in Second Language Acquisition*, 36(3), 380–383.
- Markee, N. (Ed.). (2015). *The handbook of classroom discourse and interaction*. Malden, MA and Oxford, UK: Wiley Blackwell.
- O’Connor, K. (2000). Review of the book *Language and communicative practices* by W. F. Hanks. *Mind, Culture, and Activity*, 7(3), 249–252.
- Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York: Academic Press.
- Philp, J., & Mackey, A. (2010). Interaction research: What can socially informed approaches offer to cognitivists (and vice versa)? In R. Batstone (Ed.), *Sociocognitive perspectives on language use and language learning* (pp. 210–228). Oxford and New York: Oxford University Press.
- Pintzuk, S. (1988). *VARBRUL programs for MS-DOS*. Philadelphia: University of Pennsylvania Department of Linguistics.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35(4), 655–687.
- Rand, D., & Sankoff, D. (1990). GoldVarb version 2: A variable rule application for the Macintosh. Retrieved from GoldVarb Manual website: <http://albuquerque.bioinformatics.uottawa.ca/goldVarb/GoldManual.dir/index.html>
- Sady, W. (2012). Ludwik Fleck. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2012 edition). Retrieved from <http://plato.stanford.edu/archives/sum2012/entries/fleck/>

- Schegloff, E. A. (1993). Reflection on quantification in the study of conversation. *Research on Language and Social Interaction*, 26(1), 99–128.
- Schegloff, E. A. (2000). On granularity. *Annual Review of Sociology*, 26, 715–720.
- Sfard, A. (1998). On two metaphors for learning and on the dangers of choosing just one. *Educational Researcher*, 27(2), 4–13.
- Steinhauer, K. (2014). Event-related potentials (ERPs) in second language research: A brief introduction to the technique, a selected review, and an invitation to reconsider critical periods in L2. *Applied Linguistics*, 35(4), 393–417.
- Talmy, S. (2014). Reflexivity, the cognitive-social divide, and beyond. *Studies in Second Language Acquisition*, 36(3), 383–389.
- Toohy, K. (2000). *Learning English at school: Identity, social relations, and classroom practice*. Clevedon, UK and Buffalo, NY: Multilingual Matters.
- Urban, G. (1996). Entextualization, replication, and power. In M. Silverstein & G. Urban (Eds.), *Natural histories of discourse* (pp. 21–44). Chicago: University of Chicago Press.
- Wittgenstein, L. (2001). *Philosophical investigations: The German text, with a revised English translation* (G. E. M. Anscombe, Trans., 3rd ed.). Oxford, UK and Malden, MA: Blackwell.
- Wortham, S. (2001). Language ideology and educational research. *Linguistics and Education*, 12(3), 253–259.
- Young, R. (1991). *Variation in interlanguage morphology*. New York: Peter Lang.
- Young, R., & Bayley, R. (1996). VARBRUL analysis for second language acquisition research. In R. Bayley & D. R. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 253–306). Amsterdam and Philadelphia: John Benjamins.
- Young, R. F. (2014). Tangled up in blue. *Studies in Second Language Acquisition*, 36(3), 389–397.
- Young, R. F., & Miller, E. R. (2004). Learning as changing participation: Negotiating discourse roles in the ESL writing conference. *Modern Language Journal*, 88(4), 519–535.
- Zanotto, M. S., Cameron, L., & Cavalcanti, M. C. (2008). Applied linguistic approaches to metaphor. In M. S. Zanotto, L. Cameron, & M. C. Cavalcanti (Eds.), *Confronting metaphor in use: An applied linguistic approach* (pp. 1–8). Amsterdam and Philadelphia, PA: John Benjamins.



# 3

## Quantitative Methodology

Luke K. Fryer, Jenifer Larson-Hall, and Jeffrey Stewart

### Introduction

In this chapter, we demonstrate the practice of quantitative methods of experimental design and analysis through the use of an illustrative study. This study presents three questions each addressed with different but related quantitative approaches. The aim is to (1) highlight the connection between research questions and analysis, (2) discuss the importance but limitations of more commonly utilized quantitative approaches, and (3) present a sophisticated but practical research design for effectively assessing new teaching-learning practices. The chapter, therefore, begins with a cross-sectional examination of the data (What is the difference?), followed by longitudinal testing on repeated measurements of the same kind of data (Is there a lasting difference?) and ends with a longitudinal test of a particular statistical model that includes hypothesized variables (What is the overall picture?). We use a number of approaches to answer these questions. We believe an illustrative approach will help our

---

L. K. Fryer

University of Hong Kong, Hong Kong, Hong Kong  
e-mail: [lukefryer@yahoo.com](mailto:lukefryer@yahoo.com)

J. Larson-Hall (✉)

University of Kitakyushu, Kitakyushu, Japan  
e-mail: [drlarsonhall@gmail.com](mailto:drlarsonhall@gmail.com)

J. Stewart

Kyushu Sangyo University, Fukuoka, Japan  
e-mail: [jeffjrstewart@gmail.com](mailto:jeffjrstewart@gmail.com)



readers understand what quantitative methodology is more effectively than simply discussing the theoretical structure of quantitative work. We hope this highlights the flexibility and utility of quantitative approaches, as there is never only one way to look at any given data set.

## Defining Quantitative Research

Most of us can think of times when a language learning technique has worked wonders for us or people we know; on the other hand, we can probably think of contexts where we felt lower motivation to study due to factors in our environment. Personal reflection and formalized qualitative research can be invaluable in helping us systematically organize and understand these thoughts, feelings, realizations, and challenges. Upon reflection and research, we may formulate theories as to how methods of study we have discovered can be used by others in practice, or how we could better address learner needs in situations where they are discouraged. Empirically demonstrating that our hypotheses have the intended effect, however, is harder to do, and frequently quantitative evidence is necessary in order to create changes in policies in larger institutions that could stand to benefit from our findings.

Quantitative research on our informal hypotheses can result in empirical evidence supporting the efficacy of the various solutions we may consider and help with objectivity at the same time. To establish the efficacy of a hypothesized solution, we must show that how we attempt to address an issue—be it with a teaching technique, training session, or use of a medication—has the desired effect not just with a handful of subjects anecdotally, but with a broader sample of the population after accounting for a variety of alternate reasons the phenomena could have occurred. Should our hypothesis survive such scrutiny, the resulting evidence can be essential to persuading not just those who already share our intuitions, but those that do not as well. This practice is, therefore, essential to arriving at sound conclusions and policies in situations where views differ. In addition, quantitative data, constrained by the necessity of measurement, may help researchers to think about how a given research question can be operationalized in an objective manner.

The basic elements of a quantitative approach to research are measuring things that you can count (hence, “quantity”) and gathering enough of this data to perform a statistical analysis. Defining a quantitative approach typically involves contrasting this approach against a qualitative approach, where numbers are not considered the final answer and the investigation does not look for a typical response to a treatment but rather investigates individual



reactions and responses in a thorough and comprehensive manner. One example of qualitative studies in applied linguistics is the exploration of bilingual identities, where ethnographies and interviews dominate the data collection process. In this area, a quantitative measure of identity, as equal to scores on psychometric questionnaires, has been criticized on the grounds of being severely reductive and oversimplified.

Although quantitative research is sometimes sharply contrasted with qualitative approaches, in the second language acquisition (SLA) field, Brown (2004) has posited a continuum where the quantitative end places more value on generalizability, reliability, and validity while the qualitative end emphasizes dependability, credibility, and confirmability. Certain types of research can fall more onto one end or the other of the scale, so an experiment where participants who come to a lab are randomly assigned to a treatment and take a test would fall squarely on the quantitative end of the continuum. On the contrary, a study which observed students in a classroom, recorded their utterances and then looked for patterns of interactions, would be classified as a qualitative study. However, other types of studies could be located between the two ends of this continuum.

Since the 1970s, quantitative research in applied linguistics has grown from a rarity to a near-requirement for publication in many leading journals (Loewen & Gass, 2009). It is worth considering that pressuring researchers and students of applied linguistics to engage almost exclusively in quantitative research risks “wagging the dog” by steering the field toward only the types of questions that can be answered using such methods. As such, it is more important than ever to engage in best practices with regard to quantitative research, including using multiple approaches that can answer more complex types of questions (Duff, 2010), such as longitudinal studies (Ortega & Iberri-Shea, 2005) and mixed methods research that incorporates both quantitative and qualitative methods (Brown, 2014; Jang, Wagner, & Park, 2014; see also Mackey & Bryfonski, Chap. 5). However, we believe the underlying goal of quantitative methods—to adhere to the scientific method through systematic, empirical investigation—remains an important one, even if it is sometimes overshadowed by the complexity of the statistical methods it frequently employs.

This chapter is written for readers who are already familiar with the basic ideas of quantitative research, such as identifying dependent and independent variables in a research design, classifying variables as interval level or categorical, and understanding research hypotheses and associated ideas of statistically testing those hypotheses (although we do not assume readers necessarily feel proficient in conducting those analyses). For readers who may not yet be familiar with these ideas, there are numerous book-length resources to help

novice researchers in the language acquisition field become familiar with these concepts, including Blom and Unsworth (2010), Brown and Rodgers (2003), Larson-Hall (2015), Mackey and Gass (2011), Paltridge and Phakiti (2015), Phakiti (2014), Porte (2010), and Roever and Phakiti (2018).

## Introduction to the Empirical Study

The first step in any study is to consider a real-life question that is interesting to you. Here, we will discuss what we will call the “Chatbot” study, which is based (using simulated data) on a real study conducted by the first author of this chapter (Fryer, Ainley, Thompson, Gibson, & Sherlock, 2017). Chatbots are software programs designed to simulate conversations with human users, which in recent years have become ubiquitous across the Internet acting as web receptionists, salespersons, and conversational partners (through voice to text functions). Our question was whether Chatbots would be useful as a source of language practice for pre-intermediate English as a foreign language (EFL) students. While previous research has suggested they might be motivating and useful for language learning (Fryer & Carpenter, 2006; Fryer & Nakao, 2009; Hill, Ford, & Farreras, 2015; Jia & Chen, 2008; Walker & White, 2013), research investigating this question has not been conducted.

However, we not only wondered about the effects of Chatbots, but also about whether students would find Chatbots interesting enough to sustain their interest over the long term once the novelty of the task had diminished. Our context was a university-level English course which all students at the university were obliged to take; such a compulsory subject necessarily includes students with substantial variation in interest in learning English.

We submit that student interest is a powerful component of a successful classroom. One might argue that the strongest markers of teacher success are that students learn something and leave class just as or more interested in the topic of study (for a philosophical discussion, see Dewey, 1913, a classic work available free online; for a review of the psychological implications, see Schiefele, 1991). Interest (the psychological construct) is a content-specific construct that can be divided into situational (fleeting) and individual (enduring) interest (Hidi & Renninger, 2006). The interest that students bring with them into a classroom has various components: (1) general interest for the subject that students bring with them (domain interest), (2) interest they have for a given class (course interest), and (3) interest for conducted tasks both in and outside of class (task interest) (see Fryer, Ainley, & Thompson, 2016). The more interested students are in something, the more they will persist with

it (Tobias, 1995). It is relevant, therefore, to begin by evaluating the potential of Chatbots for language learning practice. This might be done by comparing students' interest in practicing conversation with Chatbots relative to the natural alternative of student partners (task interest). Since language learning demands sustained practice over months and years, results of this test will lead to the question of the longer-term interest in artificial intelligence (AI) versus human partners. Finally, for teachers considering long-term persistence, students' interest in the course and in learning the target language beyond the context of formal education also arise as important questions.

Once the research question has been determined, one must then determine how to quantify the variables concretely, so that they can be measured. The variables involved were (1) interest in the task, (2) interest in the course, (3) interest in the domain (in this case, English classes at the university), and (4) partner (Chatbot or human). For this study, the "partner" variables were human beings (other students in the class that participants chatted with via text on tablet devices) or Chatbots.

Although applied linguistics does rely heavily on quantitative data, the constructs and processes that we find of interest in this field (e.g., L2 knowledge, identity, or motivation) are inherently qualitative in nature. We often use numbers to facilitate quantitative analyses, but when we do this, we abstract from the true constructs as they exist in their natural state, and of course this is the case when measuring motivation.

Scale development for the current study relied on top-down theory rather than bottom-up qualitative information from students. While both approaches have merit, the gap between a "layman" perspective of interest and the collative (multifaceted) construct we aimed to measure necessitated a theory-driven approach. Rather than asking students how they understood interest in classroom tasks and the course, we developed a pool of items based on current theory regarding the nature of interest. An initial pool of items (ten for each task and course) were piloted twice with a large sample of students. The results from the successive piloting were analyzed with exploratory and then confirmatory factor analysis (CFA), thereby eliminating poorly performing items. This piloting resulted in two replicable scales, each consisting of five items, which approaches the conservative minimum for latent modeling (for a discussion of this, see Brown, 2006).

One must also give serious consideration to research design. Hudson and Llosa (2015) explore the issue of randomization. The first issue is to choose a population, which in this case will be Japanese college students required to take first-year English courses at a particular university. A true experiment, which is required if one would like to make statements about causation,

demands random selection from the population. In this case, we were not able to use random selection, but we did randomly choose which condition (Chatbot or human) to first assign to each student within each class.

Another issue considered was the question of novelty. People may be naturally inclined to like new technology in the classroom, but that does not mean that they will continue to like it. To guard against this possibility, we had all of the students try out the Chatbot technology in a session before the data gathering began.

## Means Comparison Analysis

Once the research question has been decided, the measurement of variables has been operationalized and the groups gathered, researchers should consider how to analyze the data. Considering this ahead of time will help illuminate areas that might need special attention as the research is conducted, so it is worthwhile to think about what the independent and dependent variables will be if the study design is a cause-and-effect type, or what variables could be related if the design is a correlational or regression design (see Norouzian & Plonsky, Chap. 19). For more help in understanding what these types of variables are with illustrations from the field of SLA, see Phakiti (2014), Porte (2010), or Larson-Hall (2015).

In the Chatbot study, our initial question is whether there is a difference in student interest in a task when we use Chatbots versus when the students chat with other humans. This is a basic test of mean differences between two choices, and we would use a statistical test called a *t*-test for this (see Amoroso, Chap. 22, for more details). Because all the students used both the Chatbot and a human partner and rated their interest in both, we must use a paired-samples *t*-test to account for the fact that the two groups of scores are not independent (i.e., correlated, since two scores were given by each participant).

A *t*-test will test the difference in means between two groups, taking into account the variation between them as well. We always want to start by looking at the summary statistics first, meaning the average score ( $X$ ), standard deviation (SD), and number of participants in each group ( $N$ ). These descriptive statistics describe the average performance of the group and are the bare minimum of information for consumers of research to begin to understand the results of studies. They are also necessary to have if the study is to be included in the future in a meta-analysis. Thus, best practices mean that we should never leave these numbers out (Larson-Hall & Plonsky, 2015).

However, a survey of 606 studies in two leading SLA journal by Plonsky (2013) found that only 77% of the studies gave mean scores, and only 60% gave standard deviations, and that these were not necessarily given for all relevant comparisons. For this study the descriptive statistics for participants' interest are: Bots,  $X = 3.8$ ,  $SD = 1.2$ ,  $N = 121$ ; Humans,  $X = 3.9$ ,  $SD = 1.0$ ,  $N = 121$ .

Another tool which can describe the performance of the group is graphics; since they are visual they can appeal to a different part of our brain than numbers and be an invaluable addition to information we receive about experiments (Tufté, 2001). Larson-Hall and Plonsky (2015) recommend including data-rich graphics for the main analyses of a study, and Hudson (2015) contains a number of pertinent guidelines for creating consciously informative graphics.

For this study, the beeswarm plot in Fig. 3.1 is a good choice to visually show how the two groups performed relative to each other (this graphic was generated using Atsushi Mizumoto's langtest.jp website, "Comparing paired samples"). The graphic shows interest as a composite based on five questions measured on a Likert scale ranging from one to six, where six indicates higher interest.

The next step, if the number of participants in the study is large enough, is to look at *inferential statistics*. Before undertaking our inferential test (the *t*-test), there are a number of things we should do. The first is to decide on an

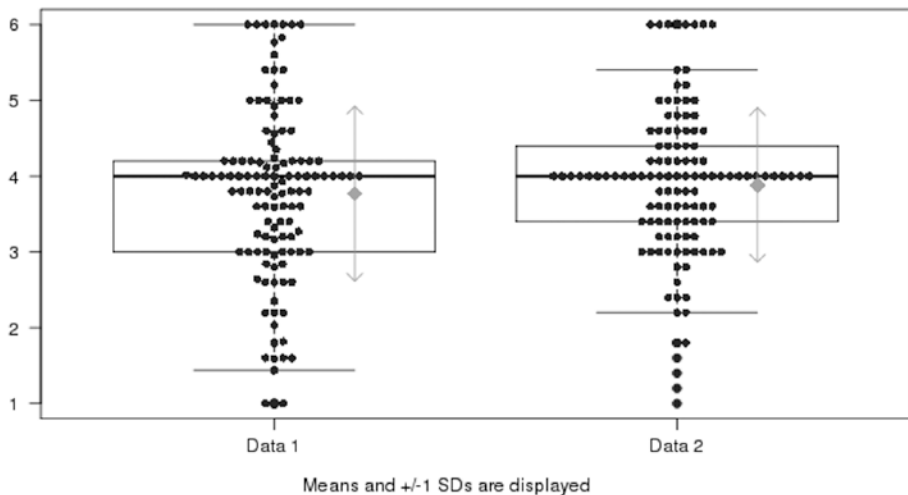


Fig. 3.1 Beeswarm plot of interest in interacting with Chatbot (Data 1) and human partner (Data 2)

*a priori* alpha level. This means we want to decide what amount of uncertainty we are comfortable with, as we establish probabilities when using inferential statistics. This level is often set at  $\alpha = 0.05$ .

However, this approach assumes the use of null hypothesis statistical testing (NHST); many problems have been noted with this approach (Larson-Hall, 2015; Norris, 2015; Norris, Ross, & Schoonen, 2015). A more informative approach which has been strongly recommended by Cumming (2012) is to look at confidence intervals (CIs) and effect sizes.

Next, we should comment on whether the assumptions for the statistical test have been met. The pertinent assumptions are that the data are normally distributed and the variances are equal. The boxplots in Fig. 3.1 showed outliers, indicating a non-normal distribution. The data points for the interest in interacting with Chatbots shown in Fig. 3.1 (Data 1) also seemed to be asymmetric around the median line (the line is not in the middle of the box), indicating a deviation from a normal distribution. Although the requirement that the distribution of data be normal is an exact requirement, in practice it is impossible to know with certainty that data is normally distributed; however, outliers or strong skewing of the data (non-symmetry in the boxplot) would indicate non-normality. Since the boxplots indicate some deviation from a normal distribution, it would be prudent to use robust statistics, a variation of the inferential statistics that most readers are familiar with. These types of statistics rely on computer-intensive procedures which were not available in the first 60 or 70 years of the twentieth century and which improve on the conclusions drawn (Tukey, 1960; Wilcox, 2001). The second author has advocated the use of such procedures as an objective way to deal with outliers (Larson-Hall & Herrington, 2009), and it is certainly preferable to simply deleting what appear to be outliers from a data set, an action which results in non-independence of the data. Such procedures can be performed in SPSS or R with relative ease (see Larson-Hall, 2015, for further information).

Larson-Hall and Plonsky (2015) also recommend that researchers state whether statistical power was considered in the design of the study in deciding how many participants to test. Actually, power can only be calculated when using an NHST approach, since one must decide on an alpha level. For the purposes of testing a hypothesis of difference, the G\*Power application can be used to make an *a priori* calculation of how many participants would be needed to find a medium or larger effect (G\*Power is downloadable from <http://www.gpower.hhu.de>). Doing an *a priori* power calculation in this app for a paired-samples *t*-test of the “difference between two dependent means,” assuming a two-tailed hypothesis, a medium effect size of  $d = 1.0$  (Plonsky & Oswald, 2014), alpha level of 0.05, and power of 0.80, results in a total

sample size of ten persons. Although researchers must work within the constraints that they find on their participant numbers, it is also important to do power analyses so that the effort that goes into a study does not waste time conducting research that is not powerful enough to answer the desired question. In this case, because the same participants are tested twice, the participant numbers needed are much smaller than if groups of different people were to be tested.

A final consideration in reporting is the reliability of our questionnaire. Cronbach's alpha for all scales was high ( $> 0.90$ ). It is generally held that a Cronbach's alpha of over 0.70 is acceptable (Devellis, 2012).

Now let us take a look at the results of a robust paired  $t$ -test on the two groups (it is a percentile bootstrapped 20% trimmed means paired-samples  $t$ -test; R code similar to this procedure can be found on page 307 in Larson-Hall, 2015). We consider confidence intervals to be more informative and appropriate than  $p$ -values, so we will report those first. This test returns a 95% confidence interval which gives a range where 95% of the time we would expect to see the possible difference in means between the groups fall if we repeatedly tested these groups. The 95% confidence interval is  $[-0.01, 0.21]$ , meaning that the difference between the groups might be as much as 0.21 of a point, or it might be zero (since the CI passes through zero). The range for the confidence interval is fairly narrow, so we can be confident that the actual difference between groups is small and there is basically no difference between groups. Another way of thinking about this confidence interval is that it shows the size of the differences between groups, which at the most might be only 0.21 points out of 6 possible points, so the effect size of the difference between groups is quite small.

If we want to look at the results the more traditional way, with  $p$ -values, the robust analysis returns a  $p$ -value of  $p = 0.10$ , which is larger than our alpha level of  $\alpha = 0.05$ . In this way, we would conclude that the probability of the groups being the same is good and we will accept our null hypothesis that the groups are equal.

In both cases we would also like to generate a standardized effect size called a Cohen's  $d$  effect size. Mizumoto's website, which generated Fig. 3.1, also calculates Cohen's effect sizes, and  $d = 0.1$ , 95% CI  $[-0.02, 0.23]$ . This effect size means that the difference between the groups amounts to 0.1 times a standard deviation of difference between groups, which is quite small. All of our results (confidence intervals with effect sizes contained in the size of the original measurement,  $p$ -values, and standardized effect sizes) point to the fact that the differences between groups are not very strong and that we should consider the groups as equivalent.



This is good news for our English program—we have just found out that our students are as happy chatting with an artificial intelligence on their computer notebooks as chatting with other students. Because of this, we might decide to have Chatbot homework so that students can continue to practice “speaking” outside of the classroom, but without the hassle of making a scheduled time to talk to a real person. Our experiment shows that the technology is definitely useful and could possibly be an important factor to help improve our students’ interest in English and motivation to continue. Since our school is planning to make a significant investment in the tablet technology, we’re quite happy to see these results.

## Longitudinal Analysis

Our comparison of students’ interest in the Chatbot versus human raters was encouraging. However, it may not be safe to conclude from the results that this will always hold true, even with the examined population. Until now, our study can be considered to be what is known as *cross-sectional*, comparing our experimental condition (students attempting the Chatbot activities) with a control condition (students speaking with one another, as usual), which is used as a reference point for comparison. Many studies in our field can be described this way; frequently, we only measure the effect of our experimental condition once in a single study, and leave it at that.

A limitation of such studies is that since the dependent variable is only measured once, we receive only what could be considered a “snapshot” of our treatment’s effect on learners. Until now we have only measured students’ engagement in the Chatbot activity at just a single point in time. Of course, what we would really like to know is whether the Chatbot holds students’ interest as much as a human partner does in the long run. In such cases, it may be worthwhile to measure student interest in our treatment not once, but multiple times throughout the semester. Such a study can be considered to be *longitudinal*. We submit that longitudinal studies are almost always to be preferred over cross-sectional studies, and Ortega and Iberri-Shea (2005) and Ortega and Byrnes (2008) concur.

Of course, one must actually design this kind of study ahead of time, and this is indeed what we did. Prior to the beginning of the course, students’ interest in English in general was also estimated with a short survey (five items); subsequently, all students used the Chatbot to reduce the novelty factor. Three weeks later, students were randomly divided into either (1) the Chatbot first, human second, or (2) the human first, Chatbot second group; half of the students tried a conversation that began with structures used in the



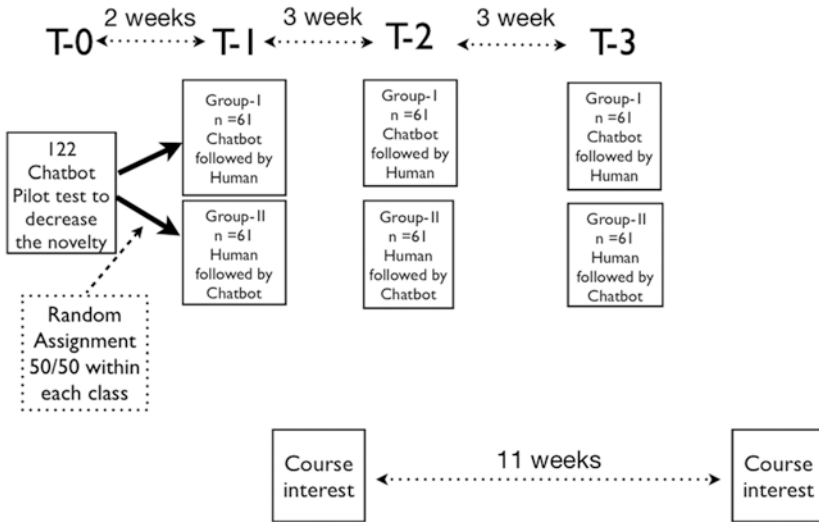


Fig. 3.2 Diagram of longitudinal Chatbot experiment design

course content with a human partner and the second half of the class tried the same conversation with the Chatbot, employing the tablet and voice to speech software (this is labeled as T-1 in Fig. 3.2). After doing the task with their first partner, students then switched and did the task with the other type of partner. Immediately after the tasks with each type of partner at each time period (T-1, T-2, and T-3) were finished, students completed a short five-item survey measuring interest in the task (the dependent variable measured). Finally, after all the tasks were completed, the students took a survey of interest in the *course*, which they had also taken 11 weeks earlier, one week after T-1.

We created a control condition by having a counterbalanced design where each participant could act as their own control. Having a control group is essential if one wants to make any statements about the effect of a variable or condition. In addition, testing the same person multiple times results in a reduction of the error in the statistical component, thus strengthening the results of the study by increasing power and decreasing the number of participants needed.

Another point to make about this illustrative study is that technology makes gathering data, even longitudinally and from large numbers of students as this study does, much less difficult than it has been previously. Technology is changing the playing field of both teaching and research and has major implications for applied linguistics (Duff, 2010). Tablets and clickers make a micro-analysis of the type we are discussing here infinitely more possible,

without interfering with valuable class time. We hope researchers will consider the value of gathering large amounts of data whatever their question may be—student interest, pronunciation accuracy, understanding of grammatical structures, and so on.

In looking at the data, first let us consider the mean scores and standard deviations. In all cases, the number of participants was 121. For Chatbots, the interest at Time 1 (Week 7) was  $X = 3.8$ ,  $SD = 1.2$ ; at Time 2 (Week 10) it was  $X = 3.1$ ,  $SD = 1.4$ ; and at Time 3 (Week 13) interest was  $X = 2.8$ ,  $SD = 1.5$ . Thus we see the interest declining over time, with the standard deviation getting larger each time. For humans, the interest at Time 1 was  $X = 3.9$ ,  $SD = 1.0$ ; Time 2 was  $X = 3.8$ ,  $SD = 1.1$ ; Time 3 was  $X = 3.8$ ,  $SD = 1.2$ . The interest for a human interlocutor stayed essentially the same every time, with similar standard deviations. Figure 3.3 shows an interaction boxplot, giving interaction plots as well as boxplots for the data in two different configurations.

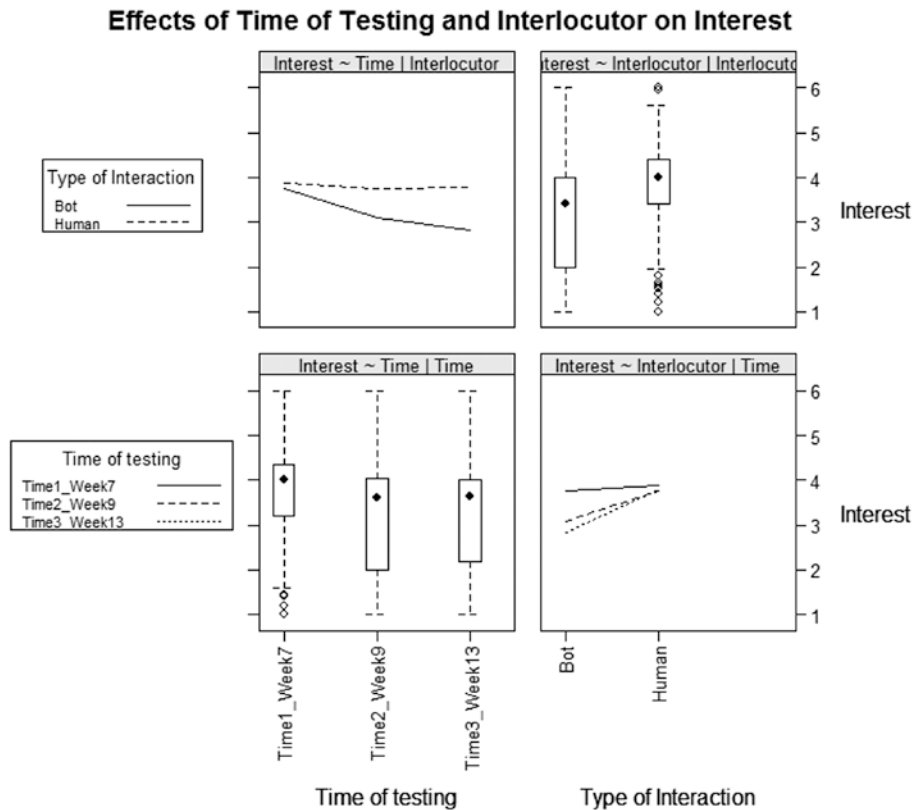


Fig. 3.3 Combination interaction/boxplots of the longitudinal Chatbot data

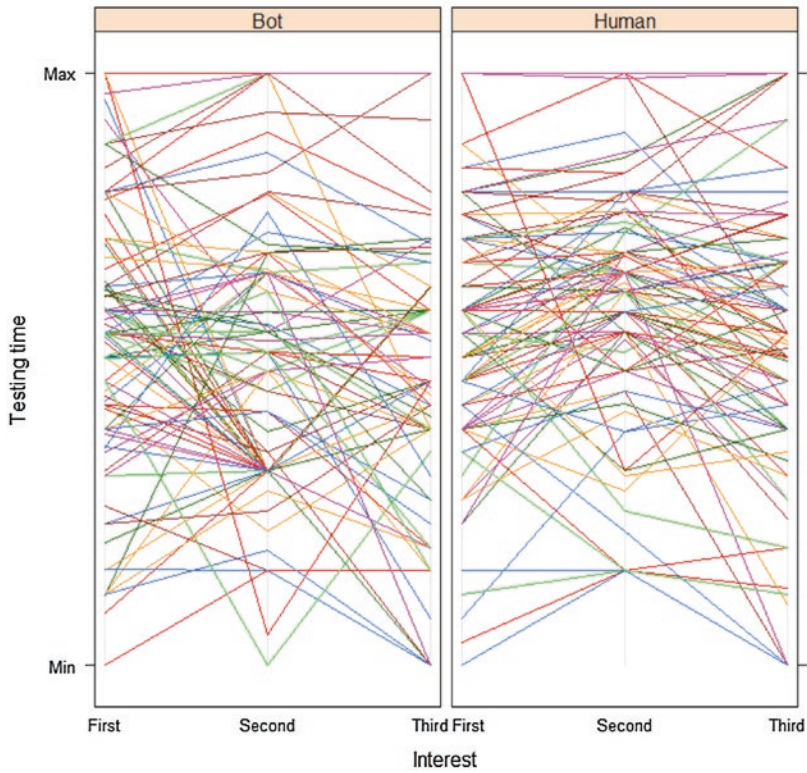
The graphics in Fig. 3.3 illustrate the picture of interest declining in the Chatbot over time, while enthusiasm for human interlocutors remained fairly constant throughout the length of the semester, although at all testing times the standard deviations were fairly wide and the range covered the entire scale, indicating a large array of opinions among the students.

## Repeated-Measures ANOVA

Although the paired-samples  $t$ -test we performed in the first step of the analysis was useful in comparing one condition to the other at one time period, for comparing the scores of the same groups of students over multiple time periods, we need to look to a repeated-measures (RM) ANOVA which can take into account the lack of independence between measurements at more than two points.

Right away, we can see from the descriptive statistics and Fig. 3.3 that two of the assumptions behind the statistical test are violated; the data are not normally distributed because boxplots indicate outliers, and standard deviations are also different when summarizing over humans versus Chatbots, with the Chatbots having a wider standard deviation (and thus variance). We proceeded with a least-squares RM ANOVA anyway; violating assumptions means it will be more difficult to find a statistical result and results in a loss of power. In this case, however, even with the violations of assumptions we were able to find a statistical two-way interaction between interlocutor and testing time,  $F_{2,223} = 25.3$ ,  $p$  Spilt 0.001, generalized eta squared = 0.02 (because sphericity was violated the Huynh-Feldt correction was applied to this result). This means that the participants did not rate their interest equally at all time periods, and this interest was affected by whether they were talking to humans or Chatbots. A formal statistical analysis would investigate the exact differences in the intersection of time and interlocutor by conducting further RM ANOVAs for testing time at the value of interlocutor (Chatbot or human), or, conversely, depending on our interests, a paired-samples  $t$ -test between Chatbots or humans could be conducted at each value of the testing time. We will skip the formal results of such an analysis in this paper, but it is clear from the interaction plots in Fig. 3.3 that the students lost interest in the Chatbots over time. The initial paired-samples  $t$ -test showed that there was no difference between Chatbot and human interlocutors at the first testing, but testing for later times showed that students became less interested in talking with Chatbots over time. Figure 3.4 gives a parallel plot showing individual results for this same data. Although individuals differed widely, the trends of higher interest scores for humans are clear as well.

### Effects of Time of Testing and Type of Interaction on Interest



**Fig. 3.4** A parallel plot showing interest in human versus Chatbot interlocutors over three testing times

We want to point out that doing only a one-time analysis then leads us to a completely erroneous conclusion about the effectiveness of Chatbots. By testing at multiple times we strengthen both our statistical and logical case and realize that the first testing time was not showing a true or complete picture of students' long-term interest in interlocutors.

## Model-Based Analysis

To summarize up to this point, cross-sectional evidence suggested no significant difference in interest between Chatbots and humans, but longitudinal analysis indicated a small yet statistical decrease in interest over time for the Chatbot speaking task but not the human speaking tasks. These results signal

the difference between these two fundamental research designs. By employing the same research design but a different analytical framework, another layer of questions might be asked. Regression analysis is an analytical method which is particularly useful for estimating the predictive connections between variables over time. While the test of such predictive relationships can be conducted with cross-sectional data, Tracz (1992) suggests that time-order (repeated-measures) data is a minimum research design component for any such estimated links to indicate potential causality (X causes Y). Tracz in fact sets out three research design criteria which must be met for causal implications to be drawn: (1) temporal sequencing of variables (X precedes Y), (2) a relationship among variables ( $r_{xy}$  Spigt 0), and (3) control (X is controlled for when predicting Y with Z). The first condition can be met relatively easily through design and the second condition must be present naturally. The third condition can be perceived as a continuum from High control (rigid experimental design) to Low control (accounting for prior variance on pre-measures).

In the current study, Criteria 1 and 2 were addressed by the use of a longitudinal research design and focusing on strongly related variables. The study's design aimed to control (Criteria 3) future interest in two ways. First, prior variance was controlled for by including prior measurements of all variables. Initial condition (Chatbot or human) was controlled for by random assignment to the treatment groups.

The research design described (presented in Fig. 3.2) lends itself to indicating predictive relationships through regression analysis. Basic regression analysis is the prediction of one variable by another variable (see Norouzian & Plonsky, Chap. 19). Multiple regression or hierarchical linear modeling can include multiple predictor variables, thereby potentially explaining more variance in the dependent variable. If the regression analysis seeks to explain or predict multiple outcomes (dependent variables), then a model-based approach is necessary.

There are two related model-based analytical approaches which might be utilized to test the predictive relationships between multiple independent and dependent variables. The first employs observed variables (mean-based) and is typically referred to as path analysis. The second approach employs latent variables (variables based on multiple indicators rather than mean scores) and is generally referred to as structural equation modeling (SEM; see also Phakiti, Chap. 21). It should be noted that path analysis forms the underpinnings of SEM. Both analytical approaches allow for a full model test, which means that each regression is not calculated separately or sequentially but instead they are all estimated simultaneously. This simultaneous analysis

enables researchers to more completely understand the individual contribution of both different predictor variables and the same predictor over multiple time points. This approach also provides model test statistics or fit indices based on the chi-square for the model. These fit indices allow for the comparison of different models and general understanding of how well the model fits the data (Kline, 2011).

In most cases, deciding between a mean or latent variable-based approach comes down to sample size. A latent-based approach provides the benefit of addressing measurement error inherent in mean scores and greater degrees of freedom, but demands a substantially larger sample size (e.g., > 200, 500, 1000; for an in-depth discussion see Fan, Thompson, & Wang, 1999). The measurement error-free latent variables afforded by this approach result in more accurate estimates of prediction and have been suggested as essential for reciprocal modeling over time (Pedhazur, 1982). The additional degrees of freedom allow latent-based models to test a greater number of regression relationships while still providing a measure of model fit (see Loehlin, 1998).

We resolved to utilize SEM. To begin with, confirmatory factor analysis (CFA) of all variables was undertaken to assess their convergent and divergent validity. CFA is a type of SEM which just assesses the latent indicator-based variables (measurement models) for fit to the data. It is an essential first step to assess whether your variables are being appropriately estimated (in other words, reasonable loading of indicator items; convergent validity) and then whether the correlation between your modeled variables is reasonable. For SEM, high correlation between modeled variables is the chief concern. Correlational relationships greater than 0.90 suggest insufficient divergent validity and will likely result in non-convergence for analyses. Acceptable fit and reasonable inter-correlations from the confirmatory factor analysis suggest proceeding with the longitudinal analysis. Figure 3.5 presents the hypothesized model to be tested. The lines in Fig. 3.5 represent regressions tested, while the circles represent the latent (multiple indicator) variables modeled.

Analysis of the hypothesized model resulted in acceptable fit to the data set. The model regression outcomes are presented in Fig. 3.6.

Based on Peterson and Brown's (2005) recommendations for regression benchmarks, in line with Hattie's (2009) guidelines for educational effect sizes, all modeled predictive effects except one were large. The strong predictive connections were expected given the theoretical consistency of the variables and relatively small time between the data points. The insignificant link between Chatbot tasks and students' interest in the course is precisely what this analytical approach is designed to look for. After accounting for prior course interest and students' interest in the human speaking task, the Chatbot

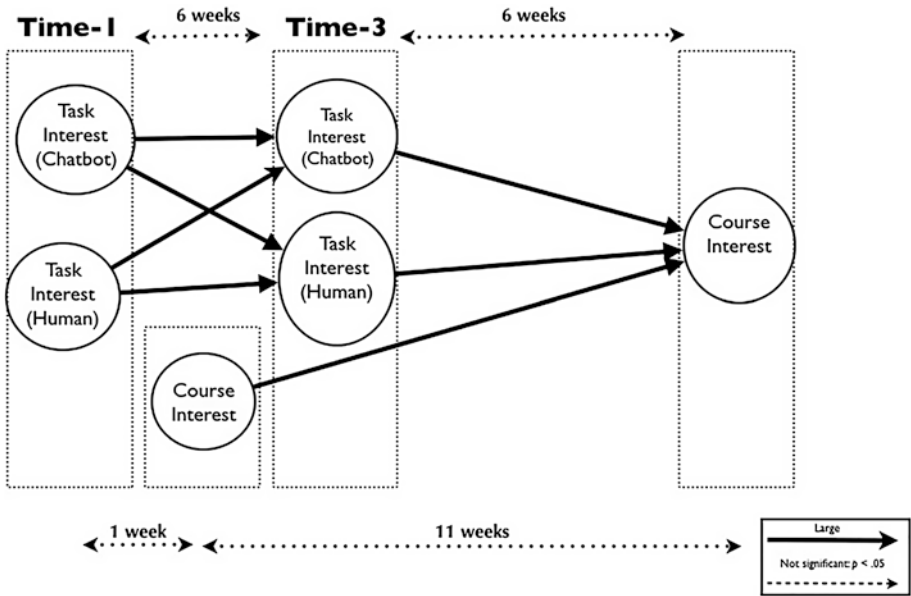


Fig. 3.5 Hypothesized model of interest in task and course

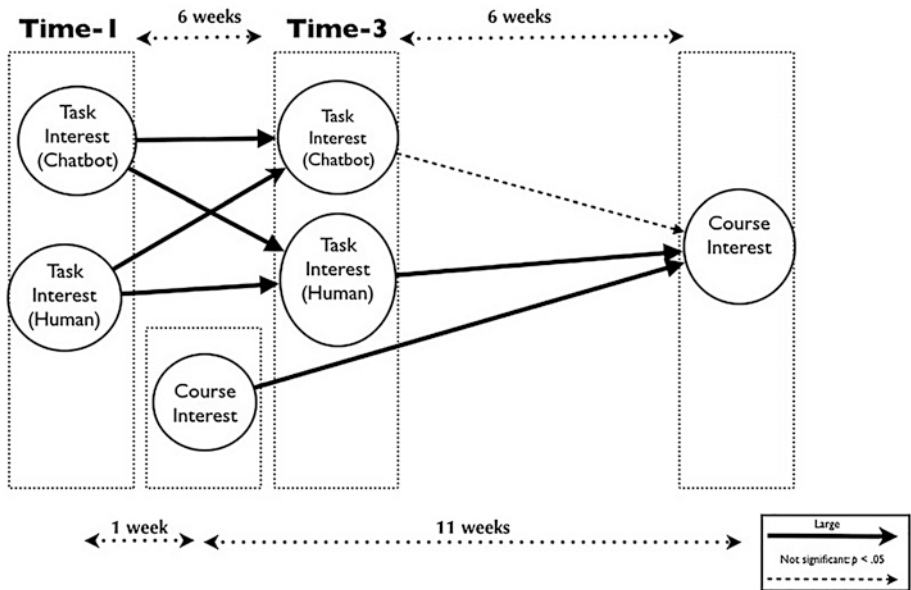


Fig. 3.6 Final model of interest in task and course



task failed to contribute to students' future interest in the course. This is an important finding because technology like Chatbots is generally brought into the class to stimulate interest in the language under study. Also, as noted, course interest is an essential mediator between interest in classroom tasks and interest in the broader domain under study. This finding demonstrates that in addition to interest in Chatbot conversation declining over time, it also fails to enhance student's long-term interest.

## Summary

In this chapter, we illustrated the steps of the quantitative approach to research methodology by walking the reader through three different research questions concerning the same research topic. We have noted that there is never only one way to look at any given data set, but have shown that progressively more sophisticated analysis can perhaps come closer to answering the questions that researchers actually hold; in this case, looking at the big picture of how Chatbots affect not just current interest in the task but interest in the course overall is probably the most informative and relevant of any of our three questions. We urge researchers to pursue knowledge of more sophisticated statistical techniques that are able to address these bigger picture questions (cf. Brown, 2015).

We would also suggest that SLA researchers utilize theoretical frameworks and analytical methods that best suit their research questions. It is common practice in many areas of SLA (not just motivation) to continue using long-standing methods and questionnaires which may not be suited to address some teachers' and students' needs. The broader fields of education and psychology can offer substantial alternatives to current approaches, which may in some cases lack strong theoretical and empirical support. SLA researchers would do well to explore all possible avenues, including those outside SLA. One possible reason for the reluctance of many SLA practitioners to embrace practices in neighboring fields could be enduring beliefs that SLA differs entirely from other social and educational science. While there may be substantial differences for some aspects of language acquisition (e.g., other fields besides linguistics may have little to say about phonemic acquisition), a considerable portion of what is often called acquisition is in actual fact forms of classroom learning. We believe classroom language learning research can be enriched by the broader fields of education, psychology, and educational psychology, most specifically.



We believe the quest to improve our practices in quantitative methodology must embrace many different facets, but that the results will be better answers to our questions and further scientific progress.

## Resources for Further Reading

Brown, J. D. (2004). Research methods for applied linguistics: Scope, characteristics, and standards. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 476–501). Oxford: Blackwell Publishing Ltd.

This chapter contains a thorough overview of types, topics, and purpose of research in the field of applied linguistics. Brown has laid out the steps of quantitative research so clearly that we did not see any point in repeating his excellent points and thus took a somewhat different approach for this chapter.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

This book is an important resource toward fully understanding confirmatory factor analysis, which is an essential building block for structural equation modeling and advanced applications.

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. New York: John Wiley & Sons.

This book is an extremely practical guide to understanding regression, which is made easy to use and understandable through an example-based structure.

Kline, R. B. (2011). *Principles and practices of structural equation modeling* (3rd ed.). New York: Guilford Press.

This book is quickly becoming the most commonly referenced introductory textbook for structural equation modeling. It is a great introductory text but also a useful resource for researchers with basic skills seeking a better understanding or to learn intermediate level applications.

Larson-Hall, J. (2016). *A guide to doing research in second language acquisition with SPSS and R* (2nd ed.). New York: Routledge.

This book walks the reader through the basics of statistical analyses including descriptive statistics, understanding variables, and foundational statistical methods such as *t*-tests, regression, and ANOVA. The SPSS and R programs are used in parallel.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

This book provides a powerful tutorial on path analysis, which is the underlying statistical framework that structural equation modeling is built upon. Path analysis is poorly understood by many applied researchers but is essential to understanding what is possible with SEM.

Phakiti, A. (2014). *Experimental research methods in language learning*. London: Bloomsbury Academic.

This book walks the reader quite thoroughly through types of quantitative research, validity and reliability, techniques and instruments, and basic statistical concepts.

## References

- Blom, E., & Unsworth, S. (Eds.). (2010). *Experimental methods in language acquisition research*. Amsterdam: John Benjamins.
- Brown, J. D. (2004). Research methods for applied linguistics: Scope, characteristics, and standards. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 476–501). Oxford: Blackwell Publishing Ltd.
- Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University Press.
- Brown, J. D. (2015). Why bother learning advanced quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 9–20). New York, NY: Routledge.
- Brown, J. D., & Rodgers, T. S. (2003). *Doing second language research*. Oxford: Oxford University Press.
- Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

- Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. New York: John Wiley & Sons.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Devellis, R. F. (2012). *Scale development: Theory and application* (3rd ed.). Thousand Oaks, CA: Sage.
- Dewey, J. (1913). *Interest and effort in education*. Boston: Houghton Mifflin Company.
- Duff, P. (2010). Research approaches in applied linguistics. In R. B. Kaplan (Ed.), *Oxford handbook of applied linguistics* (2nd ed., pp. 45–59). New York: Oxford University Press.
- Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 56–83.
- Fryer, L. K., Ainley, M., & Thompson, A. (2016). Modelling the links between students' interest in a domain, the tasks they experience and their interest in a course: Isn't interest what university is all about? *Learning and Individual Differences*, 50, 57–165.
- Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of AI and Human task partners. *Computers in Human Behavior*. <https://doi.org/10.1016/j.chb.2017.05.045>
- Fryer, L. K., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, 10(3), 8–14.
- Fryer, L. K., & Nakao, K. (2009). *Online English practice for Japanese University students: Assessing chatbots*. Paper presented at the Japan Association for Language Teaching national conference, Tokyo. Permanent Online Location: <http://jalt-publications.org/proceedings/articles/84-jalt2009-proceedings-contents>
- Hattie, J. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge and Taylor & Francis.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41(2), 111–127.
- Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49, 245–250.
- Hudson, T. (2015). Presenting quantitative data visually. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 78–105). New York: Routledge.
- Hudson, T., & Llosa, L. (2015). Design issues and inference in experimental L2 research. *Language Learning*, 65(S1), 76–96.
- Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics*, 34, 123–153.
- Jia, J., & Chen, W. (2008). Motivate the learners to practice English through playing with Chatbot CSIEC. In Z. Pan, X. Zhang, A. Rhalib, W. Woo, & Y. Li (Eds.),

- Technologies for E-learning and digital entertainment* (pp. 180–191). New York: Springer.
- Kline, R. B. (2011). *Principles and practices of structural equation modeling* (3rd ed.). New York: Guilford Press.
- Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York: Routledge.
- Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, 31(3), 368–390.
- Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(S1), 127–159.
- Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Loewen, S., & Gass, S. M. (2009). Research timeline: The use of statistics in L2 acquisition research. *Language Teaching*, 42(2), 181–196.
- Mackey, A., & Gass, S. M. (Eds.). (2011). *Research methods in second language acquisition: A practical guide*. Oxford: Wiley-Blackwell.
- Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, 65(S1), 97–126.
- Norris, J. M., Ross, S. J., & Schoonen, R. (2015). Improving second language quantitative research. *Language Learning*, 65(S1), 1–8.
- Ortega, L., & Byrnes, H. (2008). The longitudinal study of advanced L2 capacities: An introduction. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 3–20). New York: Routledge.
- Ortega, L., & Ibarra-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends. *Annual Review of Applied Linguistics*, 25, 26–45.
- Paltridge, B., & Phakiti, A. (Eds.). (2015). *Research methods in applied linguistics: A practical resource* (2nd ed.). London: Bloomsbury Academic.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston.
- Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, 90(1), 175–181.
- Phakiti, A. (2014). *Experimental research methods in language learning*. London: Bloomsbury Academic.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35(4), 655–687.
- Plonsky, L., & Oswald, F. L. (2014). Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912.
- Porte, G. K. (2010). *Appraising research in second language learning: A practical approach to critical analysis of quantitative research* (2nd ed.). Philadelphia: John Benjamins.

- Roever, C., & Phakiti, A. (2018). *Quantitative methods for second language research: A problem-solving approach*. London and New York: Routledge.
- Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist*, 26(3&4), 299–323.
- Tobias, S. (1995). Interest and metacognitive word knowledge. *Journal of Educational Psychology*, 87(3), 399–405.
- Tracz, S. M. (1992). The interpretation of beta weights in path analysis. *Multiple Linear Regression Viewpoints*, 19, 7–15.
- Tufte, E. (2001). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honour of Harold Hotelling* (pp. 448–485). Stanford, CA: Stanford University Press.
- Walker, A., & White, G. (2013). *Oxford handbooks for language teachers: Technology enhanced language learning*. Oxford: Oxford University Press.
- Wilcox, R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer-Verlag.



# 4

## Qualitative Methodology

Shim Lew, Anna Her Yang, and Linda Harklau

### Introduction

This chapter provides an overview of qualitative research (QR) in applied linguistics, with a particular focus on recent research and developments in the field over the past six years. The chapter briefly reviews the historical development of qualitative methods in applied linguistics. It explores the philosophical and methodological premises of qualitative methods in the field and relates them to broader developments in QR across the social sciences, particularly in regard to issues of validity and quality of QR. The chapter then surveys current trends and issues based upon a review of over 100 QR journal articles in applied linguistics since 2011. It characterizes the varieties of qualitative approaches taken, theoretical frameworks used, and types of data collection and analytical methods employed. It points out challenges and controversies and concludes with limitations and future directions in QR in applied linguistics.

---

S. Lew (✉)

Teacher Education and Educational Leadership, University of West Florida,  
Pensacola, FL, USA

e-mail: [slew@uwf.edu](mailto:slew@uwf.edu)

A. H. Yang • L. Harklau

TESOL & World Language Education Program & Linguistics Program,  
University of Georgia, Athens, GA, USA

e-mail: [ayang@uga.edu](mailto:ayang@uga.edu); [lharklau@uga.edu](mailto:lharklau@uga.edu)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_4](https://doi.org/10.1057/978-1-137-59900-1_4)

## Historical Development

QR in applied linguistics has been defined as “research that relies mainly on the reduction of data to words (codes, labels, categorization systems, narratives, etc.) and interpretative argument” (Benson, 2013, p. 1). Compared to quantitative research, qualitative findings generally depend on particular ways of collecting, analyzing, and interpreting data in specific contexts (e.g., Creswell & Poth, 2017; Denzin & Lincoln, 2017).

QR can be differentiated from quantitative research primarily by researchers’ contrasting views of social construction and mental models about the use of numbers (e.g., Creswell & Poth, 2017). Maxwell (2010) suggests that while quantitative research uses variables and correlations to conceptualize the world, QR uses events and processes instead. Moreover, while quantitative approaches seek regularities and causal relations by evidencing a regular association between a change in one entity and a change in another, qualitative researchers tend to view causality as fundamental association with the causal mechanisms and processes of changes across time actually involved in particular events (e.g., Denzin & Lincoln, 2017). Richards (2009) argues that in recent years applied linguistics scholars are increasingly going beyond such dichotomous views of quantitative and qualitative research, moving toward pragmatic approaches that focus more on practical and contextual issues than conceptual debates.

Benson (2013) observes that applied linguistics adopted QR methods considerably later than most other social sciences. It was not until the early 1990s that major journals (e.g., *Applied Linguistics*, *Language Learning*, *Studies in Second Language Acquisition*, and *TESOL Quarterly*) began publishing articles utilizing qualitative methods, and it took a decade more for qualitative methods to appear in research manuals in the field (e.g., Croker & Heigham, 2009; Richards, 2003). *TESOL Quarterly*’s 2003 publication of QR research guidelines (Chapelle & Duff, 2003) represented a further landmark in the field. Since then, QR has appeared regularly in applied linguistics research and has achieved mainstream status (see, e.g., the most recent *TESOL Quarterly* QR research guidelines; Mahboob et al., 2016). Nevertheless, our review concurs with previous work (e.g., Benson, Chik, Gao, Huang, & Wang, 2009; Richards, 2009) suggesting that QR appears less frequently than quantitative research in most applied linguistics journals. Our analysis also finds significant differences in QR publication trends among journals in the past six years, with some (e.g., *International Review of Applied Linguistics*) still publishing very little while the percentage has gradually increased in others (e.g., *The Modern Language Journal*, *Applied Linguistics*, and *TESOL Quarterly*).

## Philosophical and Methodological Premises

Given the status of QR as a standard, if somewhat underrepresented, research approach in the field of applied linguistics, it is important to unpack the varying philosophical and methodological premises behind varying forms of QR.

### Epistemology and Ontology Underlying QR in Applied Linguistics

Every research methodology is associated with a theoretical perspective that embodies “a certain way of understanding *what is* (ontology) as well as a certain way of understanding *what it means to know* (epistemology)” (Crotty, 2009, p. 10). These theoretical perspectives on the nature of being and knowing tend to remain tacit in most studies using quantitative methodologies since they tend to take a realist view, seeking knowledge of an objective reality existing outside of human perception (Sealey, 2013) and an objectivist epistemological perspective assuming that “truth and meaning reside in their objects independently of any consciousness” (Crotty, 2009, p. 42). Epistemological and ontological stances play a more prominent role in QR, however, because of widely varying theoretical perspectives on what kinds of knowledge researchers believe can be produced by their research and how they believe readers should interpret their research findings and outcomes.

Qualitative researchers in applied linguistics may, like quantitatively oriented colleagues, hold realist ontological views. Nevertheless, within the realm of the social sciences, realist philosophers and methodologists hold differing epistemological stances on the extent to which researchers have access to that reality, acknowledging the partiality and fallibility of human knowledge mediated through semiotic representations (e.g., Sealey, 2013). Qualitative researchers, particularly those aligning with post-structuralism, might also hold antirealist ontological views seeking “to obtain knowledge of entities that are conceived as not ‘given’, “that is, not independent of human action or of embeddedness in human culture” (Crookes, 2013, p. 1).

These varying positions have implications for the research methodology one chooses and the claims one makes. For example, even within the single qualitative methodology of grounded theory, some researchers take a realist ontology and positivist epistemology, focusing on “a method of discovery,” “a method of verification,” “an objective external reality,” “a passive, neutral observer,” “categories as emergent from the data,” and “a direct and, often, narrow empiricism” (Strauss & Corbin, 1998). If, on the other hand, research-



ers hold an antirealist ontology with constructivist epistemology, they may use grounded theory highlighting “the flexibility of the method,” “a multiple, processual, and constructed” social reality, and “researchers’ reflexivity” about their “position, privileges, perspective, and interaction” (Charmaz, 2014, p. 13).

## Validity and Quality of QR

While the notion of standards for QR in applied linguistics has been critiqued as potentially simplistic or limiting, issues of quality and rigor of QR research continue to be debated in applied linguistics and across the social sciences (Richards, 2009). Unlike objectivist-oriented research in which validity lies in eliminating or controlling personal biases, many feel that qualitative methods inevitably involve subjectivity when researchers interact with social worlds with a host of assumptions about human knowledge and realities (e.g., Crotty, 2009). Therefore, validity lies in the manner in which researchers manage their subjectivity and in detailed justification for the appropriateness of researchers’ choices of research design for a particular social setting and researcher-subject relations in order to establish transparency and accountability (Holliday, 2013).

Accordingly, instead of validity per se, qualitative researchers may use other terms. For example, in a classic essay, Lincoln and Guba (1985) proposed alternate criteria. *Credibility* refers to whether a researcher has taken necessary steps to ensure their interpretations are trustworthy, such as employing constant comparison, searching for negative evidence, and using member validation. This is related to transparency, whether research reports contain enough information about methods to help readers evaluate the trustworthiness of methods. *Transferability* refers to providing a “thick” or rich enough description of the research so that readers can assess the applicability of the study to their own situations. *Dependability* of the study might be addressed by explicitly noting issues of researcher subjectivity, particularly as it concerns methodology. Finally, researchers might seek *confirmability* by showing as much data to readers as possible in order to demonstrate the relationship between data and claims. Likewise, Tracy (2010) proposes eight key markers to gauge QR quality: (1) worthiness of topic, (2) rich rigor, (3) sincerity, (4) credibility, (5) resonance, (6) significance of contribution, (7) ethics, and (8) meaningful coherence.

Validity in qualitative methodology also often entails researcher reflexivity. Depending upon their approach, researchers might critically reflect on their

personal and theoretical biases, consider their presence in the research, view themselves as tools for inspecting the entire research process (Schwandt, 2001), and acknowledge their voice as part of text (Starfield, 2013).

In all, within the common premise in QR that the researcher is an instrument, there are varying epistemological and ontological stances holding differing views about the nature of what researchers are doing such as whether they are documenting reality; whether they even can do so; if they can, whose reality it is; what kinds of knowledge they aim to attain; what kinds of knowledge are possible; and how they know that the kinds of knowledge are adequate and legitimate. These differing beliefs are important since they influence the selection and use of methodology and methods, whether and how validity is considered, and the extent to which researcher reflexivity is considered.

## Current Trends and Issues

In line with recent reviews (e.g., Harklau, 2011), we also found that English is still the most commonly studied target language in applied linguistics research. More research is needed regarding other target languages and cultures. For example, a small but growing body of QR on heritage language learning has focused primarily on Spanish (e.g., Burke, 2012) and Korean (e.g., Choi & Yi, 2012) in the US context. French is often the focus of QR studies in Canada (e.g., Macintyre, Burns, & Jessome, 2011). Some QR studies involving multiple target languages have taken place in the context of international schools and college foreign language classes (e.g., Jonsson, 2013).

Our review of recent work shows a continuing focus on language learners at the college level (e.g., Bidabadi & Yamat, 2014) and English as a second/foreign language teachers/teacher candidates (e.g., Cheng, 2016; Nuske, 2015). Secondary school learners and teachers (e.g., Brooks, 2015) and college instructors (e.g., Kayi-Aydar, 2011) have also been frequent foci. Studies of young learners and their teachers have been relatively rare in recent work. Also present but uncommon in recent work are QR studies on adult learners (e.g., Judge, 2012) and teacher-learner, child-parent, or immigrant family interactions (e.g., de Fina, 2012).

The US remains the most frequently studied national context for QR in applied linguistics. Other Anglophone countries including the UK, Canada, Australia, New Zealand, and Ireland are also well represented. Less frequently the focus has been on European, or East or South Asian countries. Studies situated in Central and South America, Africa, and Middle Eastern contexts have been few. Likewise, research spanning multiple national contexts, which

usually examines telecollaborative and online language learning, and study abroad, has been relatively rare.

Teaching and learning English as a second/foreign language or other foreign languages continues to be the dominant topic of QR in applied linguistics (e.g., Kim, 2015). Teacher education, professional development, and teachers' experience and reflection of their teaching or learning (e.g., Shirvan, Rahmani, & Sorayyaee, 2016) have also been frequent topics in recent work. Other present topics of QR in applied linguistics include identity and agency in learners, teachers, and immigrants (e.g., Cheng, 2016; Kim & Duff, 2012); language ideologies in multilingual contexts (e.g., Rudwick & Parmegiani, 2013); heritage language learning and teaching (e.g., Choi & Yi, 2012); community and service learning (e.g., Leeman, Rabin, & Román-Mendoza, 2011); refugee experiences (e.g., Warriner, 2013); multilingualism and multiliteracies (e.g., Dorner & Layton, 2014; Takeuchi, 2015); language policies (e.g., Nero, 2015; Tamim, 2014); text/speech analysis (e.g., Britt, 2011); sign language and learners with disabilities (e.g., Cramér-Wolrath, 2015); parent involvement and perceptions of education (e.g., Kavanagh & Hickey, 2013); technology, online gaming, and telecollaboration (e.g., Shin, 2014; Whyte, 2011; Zhang, 2013); linguistic landscape analysis (e.g., Zabrodskaia, 2014); media discourse analysis (e.g., Chamberlin-Quinlisk, 2012); and research ethics (e.g., Perry, 2011). In all, QR in applied linguistics has thus far had an uneven and limited reach, leaving much potential for future exploration.

## Overview of Research Designs/Types

Major types of QR in applied linguistics can be classified in varying ways. However, most scholars (e.g., Benson et al., 2009; Harklau, 2011) identify two main points of departure or approaches for QR methodologies. One is the analyses of "sociocultural and ecological contexts of language learning and teaching" (Harklau, 2011, p. 178) or "the people, situations, and social processes involved in language learning and teaching" (Benson et al., 2009, p. 84). The other is the analyses of "spoken and written texts" (Benson et al., 2009, p. 84), or "the construction of social realities through discourse" (Harklau, 2011, p. 178). The former approach relies primarily upon analysis of participant observation and interviews, while the latter relies primarily on the analysis of audio or video recordings and texts. The former tends to include case study, ethnography, longitudinal studies, think-aloud studies, narrative, self-study, stimulated recall, action research, diary study, and phenomenology, while the latter tends to prioritize discourse analytic methods, analysis of

classroom interaction, conversation analysis (CA), corpus study, genre analysis, and systemic functional analysis (Benson et al., 2009).

## Varieties of Qualitative Approaches

Several methodologies tend to predominate in recent QR in applied linguistics.

### Case Study

While most quantitative approaches seek large samples, a case study explores “a ‘how’ and ‘why’ question” about “a contemporary set of events”—a case, “over which a research has little or no control” (Yin, 2014, p. 14). In applied linguistics, the case usually has been “a person (e.g., a teacher, learner, speaker, writer, or interlocutor) or a small number of individuals on their own or in a group (e.g., a family, a class, a work team, or a community of practice)” (Duff, 2014, p. 233). Qualitative case studies in applied linguistics are often longitudinal (Hood, 2009) since they usually document the process, rather than the outcome, of language learning and teaching, the interaction between individuals and contexts mediated by various factors over time (Harklau, 2008), and the development of learner/teacher identity and cognition. Case studies are often used to explore “previously ill-understood populations, situations, and processes” (Duff, 2014, p. 242). Recent examples include linguistic minority students, heritage language learners, indigenous language learners, and transnational language learners and sojourners. Case studies have not only contributed significantly to theories and models in those areas in applied linguistics but also have often influenced educational policies and practices (Duff, 2014).

Case study is also one of the oldest and most common forms of inquiry in applied linguistics. Recent case study work in applied linguistics may or may not be further specified as *ethnographic*, *multiple* or *multi-case*, *collective*, *naturalistic*, or *narrative* (e.g., Brooks, 2015; Wyatt, 2013). Many studies align methods with well-known case study texts including Merriam (2009), Stake (2000), Duff (2008), Richards (2011), and Yin (2014). Case study is frequently employed in current applied linguistics scholarship to document the implementation of specific instructional tools, programs, or policies. Other recent case study foci have included challenges of language learners and teachers in particular contexts; interactions among teachers/learners, family members, or multiple parties in particular settings; learners’ or teachers’ identity formation or emotions; learning processes and changes in learners’ use of linguistic concepts;

and individual reactions to experiences such as study abroad or teacher reflection.

## Ethnography

Ethnography, originating in anthropology and sociology, refers to a variety of research approaches distinguished by “[involving] some degree of direct participation and observation” and “[constituting] a radically distinctive way of understanding social activity *in situ*” (Atkinson, 2015, pp. 3–4). While ethnography is usually associated with an emic approach in terms of “developing empathy with, and giving voice to, participants” (Markee, 2013, p. 3), in applied linguistics it varies in whether it focuses on “the traditional anthropological emically-oriented *what* of participant understandings and experiences” or on “the interactional sociolinguist’s etically-oriented *how* participants structure social realities through interaction” (Harklau, 2005, p. 189).

Studies in the field may be explicitly identified as ethnography (e.g., Dorner, 2011; Jonsson, 2013; Malsbary, 2014) or ethnographic case study (e.g., Huang, 2014; Marianne, 2011). Alternatively, researchers may note the use of ethnographic data collection and analysis methods including observation, field notes, microethnography, and in-depth interviews (De Costa, 2016). Like the majority of QR methods, data sets are likely to contain interviews, documents or other textual and photographic archival data, and observation and field notes. Some studies may combine ethnography with other methodological strands of QR such as grounded theory and/or critical discourse analysis. Widely cited authorities for this methodology include Glaser and Strauss (1967), Spradley (1980), Strauss and Corbin (1998), Erickson (1992), LeCompte and Schensul (1999), Denzin and Lincoln (2003), Fetterman (1998), DeWalt and DeWalt (2002), Patton (2014), and Emerson, Fretz, and Shaw (2011). Much of the current work using ethnography takes place in multilingual educational settings (e.g., Hornberger & Link, 2012) and investigates issues including the experience of learners or teachers in particular contexts, the development of particular competence, dual language education, multilingual literacies, linguistic landscapes, and language ideologies (see Blommaert, 2013; McCarty, 2014).

## Conversation Analysis

CA, as developed by Harvey Sacks, rose out of the tradition of ethnomethodology to examine language as social action, such as turn-taking (Sacks, Schegloff, & Jefferson, 1974). CA aims to account for the practices and pref-

erence organization of turn-taking, repair, and conversational sequencing during naturalistic or instructed talk-in-interaction (e.g., Hutchby & Wooffitt, 2008). Over the years, CA has broadened greatly to become an influential methodology for examining social interaction and its sequential organization across the social sciences, including applied linguistics (e.g., Markee, 2000; Seedhouse, 2004). Scholars see CA as both informing and informed by applied linguistics. For example, they point out that their research offers significant additions to “classic CA’s entrenched monolingualism” (Kasper & Wagner, 2014, p. 200) through studies of practices of multilingual interaction and multimodality.

Studies in this tradition may be self-identified as CA (e.g., Britt, 2011; Dings, 2014), or they may be distinguished by their use of CA transcription conventions and the systems developed by Sacks et al. (1974). Ethnomethodology, founded by Garfinkel (1967), is often cited as well since CA is derived from ethnomethodology. Topics of work using CA approaches vary from specific aspects of interaction such as alignment activity, membership categories, epistemic change, or code-switching to broader aspects of language such as pronunciation, humor, gesture, or narrative construction. Even broader are CA’s investigations of language learner development and interactional competence over time. Some recent work uses CA as part of a mixed methods design that also incorporates quantitative analysis of discourse features (e.g., Siegel, 2015; Taguchi, 2014).

## Grounded Theory

Originating in the sociology of medicine in the mid-1960s (see Glaser & Strauss, 1967), grounded theory has become a widespread QR methodology across the applied social sciences. It offers “a set of general principles, guidelines, strategies, and heuristic devices” (Charmaz, 2014, p. 3) “for collecting and analyzing qualitative data” to “construct a theory ‘grounded’ in [researchers’] data” (p. 1). Grounded theorists share common strategies including open-ended, inductive inquiry; “[conducting] data collection and analysis simultaneously in an iterative process”; “[analyzing] actions and processes rather than themes and structure”; “[using] comparative methods”; “[developing] inductive abstract analytic categories”; “[engaging] in theoretical sampling”; and “[emphasizing] theory construction rather than description or application of current theories” (p. 26).

Applied linguistics scholarship using grounded theory (e.g., Choi & Yi, 2012; Dorner & Layton, 2014; Valmori & De Costa, 2016) can often be

distinguished by the use or mention of associated techniques or tenets such as open/thematic, axial, and selective coding, or cyclical process. Key methodological texts in this area include Glaser and Strauss (1967), Charmaz (2014), and Strauss and Corbin (1998).

## Narrative Inquiry

Originating in a philosophical perspective that sees story as human beings' most fundamental means of making sense of the world (see Connelly & Clandinin, 2006), narrative inquiry can refer to "research in which narratives, or stories, play a significant role" (Benson, 2014, p. 155). In applied linguistics, studies using narrative inquiry may go by a number of different names including life histories, language learning histories, language learning experience, diary studies, memoirs, language biographies, autobiographies, and autoethnography (Barkhuizen, 2014). Frequently used data sources for this method include "autobiographical records or reflection, published memoirs, written language learning histories, or interviews" (p. 156). Narrative inquiry is most commonly used to document language learners' and teachers' development, practices, identity, agency, beliefs, emotion, positionality, and motivation (e.g., Baynham, 2011; Canagarajah, 2012; Casanave, 2012; Liu & Xu, 2011). It has been applied in a variety of contexts including informal and out-of-class learning, foreign language classrooms, study abroad, and refugee experiences. Ideally, narrative inquiry can capture what Benson (2014) calls "*language learning careers* or learners' retrospective conceptions of how their language learning has developed over the longer term" (p. 165). Narrative inquiries in applied linguistics highlight that learning a language is to acquire and develop new identities beyond learning its forms and functions. Narrative research has been a major methodology for exploring language teachers' and teacher candidates' perceptions and experiences (e.g., Andrew, 2011; Zheng & Borg, 2014). The work of Connelly and Clandinin (2006) has been highly influential in this school of inquiry and is frequently cited, along with Kramp (2004), Benson (2014), and Lieblich, Tuval-Mashiach, and Zilber (1998).

## Critical Discourse Analysis (CDA)

CDA seeks to discover ways in which "social structures of inequality are produced in and through language and discourse" (Lin, 2014, p. 214). In CDA, language is seen as "intrinsically ideological" (p. 215). Theorists posit that it "plays a key (albeit often invisible) role in naturalizing, normalizing, and thus



masking, producing, and reproducing inequalities in society” (p. 215). CDA frameworks seek to theorize, to varying degrees, the relationship between microstructures of language and macrostructure of society, as mediated by social practices. CDA has often been used in applied linguistics to analyze public and media discourse and educational discourse. Studies typically do not use CDA as the sole methodology but in combination with other methodologies such as case study (e.g., Dorner & Layton, 2014). Frequently cited figures in CDA include Fairclough (2003) and Rogers (2011).

## Action Research

Action research in applied linguistics is not distinguishable from other research methods’ methodological features per se. Rather, what makes it distinctive is that research is conducted by teacher-researchers as a form of self-critical inquiry about their educational practice (Cochran-Smith & Lytle, 1999). Situated in local educational practices, the quality and rigor of action research has been called into question. Nevertheless, Burns (2009) argues that action research develops reflective teachers who are committed to thinking professional and, as a result, can identify conditions in their own classrooms that do not coincide with the theories presented to them in their teacher training (Burns, 2009). Thus, the research findings they discover are “far more attractive” and “immediate to [their] teaching situations” (p. 7). Action research is of benefit not only to teacher-researchers by empowering them as agents rather than receivers of theory, research, and policy but also to the profession and field of language teaching (Burns, 2009). Recent published work, for example, has explored teachers’ experience with professional development, documented their implementation of new teaching methods and curricula, and investigated learners’ development (e.g., Burke, 2012; Calvert & Sheen, 2015; Castellano, Mynard, & Rubesch, 2011). Methodological work frequently cited in this area includes Coghlan and Brannick (2010) and Burns (2009).

## Other Qualitative Methodologies

A number of studies in applied linguistics do not specify a particular methodology but instead use the term “qualitative” more generically. Other studies are not explicitly qualitative but nonetheless use data gathering and reporting methods associated with qualitative methodologies such as interviews, discourse analysis, or inductive thematic or content analysis. Moreover, a grow-



ing number of studies combine multiple QR methodologies. Since various qualitative methodologies often share similar methodological features and a common purpose to come to a holistic understanding of a particular phenomenon, this can be a pragmatic and productive combination. Nevertheless, underspecifying or combining QR methodologies should be approached with caution, since there are important differences and potential points of conflict in the epistemological and ontological assumptions underlying various methodologies.

## Theoretical Frameworks

Some view QR in the social sciences as part of a move from personal issues toward public and social issues, such as social policies and social justice (Denzin & Lincoln, 2017). While QR in applied linguistics is often applied with no “clear sociopolitical agenda” (Lazaraton, 2003, p. 3), Benson (2013) notes that applied linguists who adopt sociocultural perspectives and theoretical frameworks “such as sociocultural theory, communities of practice, social realism, and ecological approaches” tend to gravitate toward qualitative methods (p. 3).

A review of current QR in the field indicates that sociocultural theory and critical theory are two of the most frequent self-identified theoretical frameworks. Other widely used theoretical frameworks include language socialization (e.g., Kim & Duff, 2012), Deleuzian and Foucaultian post-structuralism (e.g., Waterhouse, 2012), constructivism (e.g., Wyatt, 2013), activity theory (e.g., Zhu & Mitchell, 2012), systemic functional linguistics (e.g., Dafouz & Hibler, 2013), Bakhtinian dialogism (e.g., Huang, 2014), and positioning theory (e.g., Pinnow & Chval, 2015). Nonetheless, it is important to note that many studies adopt particular constructs from theories without adopting associated theoretical frameworks wholesale.

## Data Collection Methods

Commonly used methods for collecting data in QR in applied linguistics include interviews, observations, questionnaires, audio and video recordings of interaction, and collection of textual artifacts (Benson, 2013; Harklau, 2011). Each may be used alone or in combination (Benson, 2013). Our

review found that over two-thirds used interviews. Interviews without any further specification or “semi-structured” interviews were most commonly mentioned. Interviews were also variously described as “cognitive,” “qualitative,” “open-ended,” “diary-based,” “informal,” “in-depth,” “extended,” “lengthy,” or “ethnographic,” or as “video-stimulated recall interviews” or “interview conversations.” Ten percent of these studies relied exclusively on interviews. Most studies used face-to-face interviews; phone interviews or other types of interviews were rare. Few studies utilized focus group interviews or conversations.

Observations were another common method, used in one-third of the studies we reviewed. The majority of studies used observation in combination with interviews. “Observation” without further specification, “participant observation,” and “classroom observation” were the most common descriptions used. “Ethnographic” and “video” observations were also used. Some studies mentioned accompanying field notes with observation, while others did not. Very few studies mentioned observation protocols or rubrics.

Approximately one quarter of the QR studies we reviewed used questionnaires and a few used surveys. Surveys frequently contained at least some open-ended questions. Audio and/or video recordings were also a common method, used in about ten percent of the studies reviewed. Recordings encompassed activities including interaction, reading, feedback sessions, group discussion, presentations, classes, think-aloud and stimulated recall, and segments of speech from various speakers. Duration of recordings is usually not specified.

We also found a myriad of written artifact data sources used in QR in applied linguistics. These included journals or reflections of language teachers or student teachers, language learner diaries/blogs/journals, oral and written language production samples, presentations and writing projects, essays, online interactions, dialogs, screen capture of text chat, email discussion, logs, emails, self-evaluation reports, evaluations of instruction, narratives, and drawings. Furthermore, data also included samples of classroom curriculum and instructional materials including worksheets, lesson plans, course syllabi and unit outlines, textbooks, newspaper and magazine articles, and web pages. School documents and websites were another type of written artifact, as were public documents including newspaper articles, public comments, websites, reports, and policies. Linguistic landscape studies drew on multimedia artifacts including photos and multilingual street signs. Lastly, researchers noted collecting their own texts as well such as research logs/notes, research journals, analytical memos, and annotations of interviews.

## Data Analysis Methods

Methods of data analysis and data reduction in recent QR reports in applied linguistics are often underspecified. As in Benson's (2013) review, we found that only a handful of studies we reviewed gave full accounts of the analysis process. The majority did not describe data analysis methods and procedures in detail, referring instead to established analytical methods. Most frequently these included techniques from grounded theory (e.g., open, axial, and selective coding and the constant comparative method), phenomenology, ethnography, qualitative content analysis, and ethically generated typological analysis. Researchers often used general descriptive terms such as *interactive*, *iterative*, *inductive*, and *recursive*. An increasing number indicated using software such as QSR NVivo, Transana, and ATLAS.ti.

A lack of specificity in describing data analysis has been a common weakness across the social sciences, with qualitative methodologists calling data analysis "the 'black hole' of qualitative research" (Lather, 1991, p. 149; St. Pierre & Jackson, 2014). They suggest that post-coding data analysis "occurs everywhere and all the time" (St. Pierre & Jackson, 2014, p. 717) rather than linearly and sequentially from data collection to analysis and to representation. They call for increased training for qualitative researchers in data analysis as thinking with theory (Jackson & Mazzei, 2012) rather than as a simple mechanical coding process that "continues to be mired in positivism" (St. Pierre & Jackson, 2014, p. 717). On the other hand, applied linguistics as a field may be at somewhat of an advantage methodologically compared to other branches of the social sciences when transcribing and analyzing verbatim texts because of its rigorous analysis of language and text (Benson, 2013).

## Gaps, Challenges, and Controversial Issues

Despite the growth of QR in applied linguistics, there continue to be gaps, challenges, and controversies. For one thing, the predominance of studies of English acquisition in university settings in Western nations continues. While studies of adolescents have increased, QR studies of other populations including young learners, those who struggle with fragmented schooling, the learning disabled, and LGBT individuals remain rare. Studies of language teachers continue to focus primarily on pre-service educators rather than those already in the classroom. In spite of growing theoretical interest in topics such as community linguistic diversity and superdiversity, multilingualism, and lan-

guage in the professions, studies on settings outside of formal education continue to be relatively rare and are greatly needed.

Moreover, while QR in applied linguistics has made incursions internationally, QR methodologies tend not to address multilingual and multicultural scholarship and rely almost exclusively on Western philosophical and research traditions. Logistics of doing QR in cultural and sociopolitical contexts outside of mainstream Western contexts are not often considered (Flick, 2014), and few studies consider the ethics of doing QR in a particular socio-cultural context (but see De Costa, 2014).

## Limitations and Future Directions

While this review offers a portrait of the field, because of the sheer diversity of QR approaches taken as well as the tendency for applied linguistics researchers to underspecify QR methodologies (Benson et al., 2009), no single review can be exhaustive. While we have shown that QR methods have become a routine and well-accepted mode of inquiry in contemporary applied linguistics research, appearing regularly in peer reviewed journals, books, and book chapters, we have also noted that QR remains somewhat limited both in topic coverage and publication venues. Applied linguistics is not always about second language acquisition (SLA) or multilingualism in Western contexts. Despite the diversity and prosperity of QR in the field, there are still significant gaps and unexamined languages, populations, and regions. Turning to these areas could in turn lead future applied linguistics researchers to develop innovative QR methods. Moreover, there remain clear differences among journals in how receptive they are to publishing QR reports, and thus, a great deal of work needs to be done to promote understanding about the value of QR to publishers as well as policy makers and research funders in the field, who still place a higher value on quantitative research methods. Finally, this review has noted a lack of specificity when describing methodology or data analysis. In order to ensure greater understanding and acceptance of QR in applied linguistics, research reports need to provide fuller accounts of research design.

## Resources for Further Reading

Benson, P. (2013). Qualitative methods: Overview. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Chichester: Wiley-Blackwell.

This chapter provides a brief overview of qualitative methods in applied linguistics encompassing historical background, qualitative research in applied linguistics journals, approaches to qualitative research, data collection methods, data analysis methods, issues of quality, research areas, frameworks, and themes.

De Costa, P. I., Valmori, L., & Choi, I. (2017). Qualitative research methods. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 522–540). New York: Routledge.

This chapter discusses current issues in qualitative research methods in contemporary instructed second language acquisition (ISLA) research. By drawing on sample studies, it explores five issues: understanding and establishing rigor in qualitative research, articulating the discourse analytic approach, exploring researcher reflexivity and ethics, mixing methods, and unconventional blending of theories with different methodologies.

Harklau, L. (2011). Approaches and methods in recent qualitative research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 175–189). New York: Routledge.

This chapter offers a profile of recent trends in qualitative research on second language teaching and learning since 2003 and discusses methods of data collection, methodological frameworks, and future direction.

Richards, K. (2009). Trends in qualitative research in language teaching since 2000. *Language Teaching*, 42(2), 147–180.

This article provides a review of the developments in qualitative research in language teaching since the year 2000 with a particular focus on its contributions to the field and emerging issues.

## References

- Andrew, M. (2011). “Like a newborn baby”: Using journals to record changing identities beyond the classroom. *TESL Canada Journal*, 29(1), 57–76.
- Atkinson, P. (2015). *For ethnography*. Los Angeles, CA: SAGE.

- Barkhuizen, G. (2014). Narrative research in language teaching and learning. *Language Teaching*, 47(4), 450–466.
- Baynham, M. (2011). Stance, positioning, and alignment in narratives of professional experience. *Language in Society*, 40, 63–74.
- Benson, P. (2013). Qualitative methods: Overview. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Chichester: Wiley-Blackwell.
- Benson, P. (2014). Narrative inquiry in applied linguistics research. *Annual Review of Applied Linguistics*, 34, 154–170.
- Benson, P., Chik, A., Gao, X., Huang, J., & Wang, W. (2009). Qualitative research in language teaching and learning journals, 1997–2006. *Modern Language Journal*, 93(1), 79–90.
- Bidabadi, F. S., & Yamat, H. (2014). Strategies employed by Iranian EFL freshman university students in extensive listening: A qualitative research. *International Journal of Qualitative Studies in Education*, 27(1), 23–41.
- Blommaert, J. (2013). *Ethnography, superdiversity and linguistic landscapes: Chronicles of complexity*. Bristol: Multilingual Matters.
- Britt, E. (2011). “Can the church say amen”: Strategic uses of black preaching style at the State of the Black Union. *Language in Society*, 40, 211–233.
- Brooks, M. D. (2015). Notes and talk: An examination of a long-term English learner reading-to-learn in a high school biology classroom. *Language and Education*, 30(3), 235–251.
- Burke, B. M. (2012). Experiential professional development: Improving world language pedagogy inside Spanish classrooms. *Hispania*, 95(4), 714–733.
- Burns, A. (2009). *Doing action research in English language teaching: A guide for practitioners*. New York, NY: Routledge.
- Calvert, M., & Sheen, Y. (2015). Task-based language learning and teaching: An action-research study. *Language Teaching Research*, 19(2), 226–244.
- Canagarajah, A. S. (2012). Teacher development in a global profession: An autoethnography. *TESOL Quarterly*, 46(2), 258–279.
- Casanave, C. P. (2012). Diary of a dabbler: Ecological influences on an EFL teacher’s efforts to study Japanese informally. *TESOL Quarterly*, 46(4), 642–670.
- Castellano, J., Mynard, J., & Rubesch, T. (2011). Student technology use in a self-access center. *Language Learning & Technology*, 15(3), 12–27.
- Chamberlin-Quinlisk, C. (2012). Critical media analysis in teacher education: Exploring language-learners’ identity through mediated images of a non-native speaker of English. *TESL Canada Journal*, 29(2), 42–57.
- Chapelle, C. A., & Duff, P. A. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly*, 37(1), 147–178.
- Charmaz, K. (2014). *Constructing grounded theory*. Thousand Oaks, CA: SAGE.
- Cheng, X. (2016). A narrative inquiry of identity formation of EFL university teachers. *Journal of Education and Training Studies*, 4(5), 1–7.
- Choi, J., & Yi, Y. (2012). The use and role of pop culture in heritage language learning: A study of advanced learners of Korean. *Foreign Language Annals*, 45(1), 110–129.

- Cochran-Smith, M., & Lytle, S. L. (1999). The teacher research movement: A decade later. *Educational Researcher*, 28(7), 15–25.
- Coghlan, D., & Brannick, T. (2010). *Doing action research in your own organization* (3rd ed.). London: SAGE.
- Connelly, F. M., & Clandinin, D. J. (2006). Narrative inquiry. In J. L. Green, G. Camilli, & P. Elmore (Eds.), *Handbook of complementary methods in education research* (3rd ed., pp. 477–487). Mahwah, NJ: Lawrence Erlbaum.
- Cramér-Wolrath, E. (2015). Mediating native Swedish sign language: First language in gestural modality interactions at storytime. *Sign Language Studies*, 15(3), 266–295.
- Creswell, J. W., & Poth, C. N. (2017). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). Thousand Oaks, CA: SAGE.
- Croker, R. A., & Heigham, J. (2009). *Qualitative research in applied linguistics: A practical introduction*. New York: Palgrave Macmillan.
- Crookes, G. (2013). Epistemology and ontology. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–8). Chichester: Wiley-Blackwell.
- Crotty, M. (2009). *The foundations of social research: Meaning and perspective in the research process*. Thousand Oaks, CA: SAGE.
- Dafouz, E., & Hibler, A. (2013). ‘Zip your lips’ or ‘Keep quiet’: Main teachers’ and language assistants’ classroom discourse in CLIL settings. *Modern Language Journal*, 97(3), 655–669.
- De Costa, P. I. (2014). Making ethical decisions in an ethnographic study. *TESOL Quarterly*, 48(2), 413–422.
- De Costa, P. I. (2016). *The power of identity and ideology in language learning: Designer immigrants learning English in Singapore*. Dordrecht: Springer.
- De Fina, A. (2012). Family interaction and engagement with the heritage language: A case study. *Multilingual*, 31(4), 349–379.
- Denzin, N. K., & Lincoln, Y. S. (2003). *The landscape of qualitative research*. Thousand Oaks, CA: Sage Publications.
- Denzin, N. K., & Lincoln, Y. S. (2017). *The SAGE Handbook of qualitative research* (5th ed.). Thousand Oaks, CA: SAGE.
- DeWalt, K. M., & DeWalt, B. R. (2002). *Participant observation: A guide for fieldworkers*. Walnut Creek, CA: AltaMira Press.
- Dorner, L. M. (2011). Contested communities in a debate over dual-language education: The import of “Public” values on public policies. *Educational Policy*, 25(4), 577–613.
- Dings, A. (2014). Interactional competence and the development of alignment activity. *Modern Language Journal*, 98(3), 742–756.
- Dorner, L. M., & Layton, A. (2014). “¿Cómo se dice?” Children’s multilingual discourses (or interacting, representing, and being) in a first-grade Spanish immersion classroom. *Linguistics and Education*, 25, 24–39.
- Duff, P. A. (2008). *Case study research in applied linguistics*. New York, NY: Lawrence Erlbaum.



- Duff, P. A. (2014). Case study research on language learning and use. *Annual Review of Applied Linguistics*, 34, 233–255.
- Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing ethnographic fieldnotes* (2nd ed.). Chicago: University of Chicago Press.
- Erickson, F. (1992). Ethnographic microanalysis of interaction. In M. D. LeCompte, W. L. Millroy, & J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 202–225). New York: Academic Press.
- Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. New York: Routledge.
- Fetterman, D. (1998). *Ethnography: Step by step* (2nd ed.). Thousand Oaks, CA: SAGE.
- Flick, U. (2014). Challenges for qualitative inquiry as a global endeavor: Introduction to the special issue introduction. *Qualitative Inquiry*, 20(9), 1059–1063.
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine de Gruyter.
- Harklau, L. (2005). Ethnography and ethnographic research on second language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 179–194). Mahwah, NJ: Lawrence Erlbaum Associates.
- Harklau, L. (2008). Developing qualitative longitudinal case studies of advanced language learners. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 23–35). New York: Routledge.
- Harklau, L. (2011). Approaches and methods in recent qualitative research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 175–189). New York: Routledge.
- Holliday, A. R. (2013). Validity in qualitative research. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Chichester: Wiley-Blackwell.
- Hood, M. (2009). Case study. In R. A. Croker & J. Heigham (Eds.), *Qualitative research in applied linguistics: A practical introduction* (pp. 66–90). New York: Palgrave Macmillan.
- Hornberger, N., & Link, H. (2012). Translanguaging and transnational literacies in multilingual classrooms: A biliteracy lens. *International Journal of Bilingual Education and Bilingualism*, 15(3), 261–278.
- Huang, I.-C. (2014). Contextualizing teacher identity of non-native-English speakers in U.S. secondary ESL classrooms: A Bakhtinian perspective. *Linguistics and Education*, 25, 119–128.
- Hutchby, I., & Wooffitt, R. (2008). *Conversation analysis* (2nd ed.). Cambridge: Polity.
- Jackson, A. Y., & Mazzei, L. A. (2012). *Thinking with theory in qualitative research: Viewing data across multiple perspectives*. London: Routledge.
- Jonsson, C. (2013). Translanguaging and multilingual literacies: Diary-based case studies of adolescents in an international school. *International Journal of the Sociology of Language*, 224, 85–117.



- Judge, J. W. (2012). Use of language learning strategies by Spanish adults for Business English. *International Journal of English Studies*, 12(1), 37–54.
- Kasper, G., & Wagner, J. (2014). Conversation analysis in applied linguistics. *Annual Review of Applied Linguistics*, 34, 171–212.
- Kavanagh, L., & Hickey, T. M. (2013). ‘You’re looking at this different language and it freezes you out straight away’: Identifying challenges to parental involvement among immersion parents. *Language and Education*, 27(5), 432–450.
- Kayi-Aydar, H. (2011). Re-exploring the knowledge base of language teaching: Four ESL Teachers’ classroom practices and perspectives. *TESL Reporter*, 44(1–2), 25–41.
- Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, 12(3), 239–261.
- Kim, J., & Duff, P. A. (2012). The language socialization and identity negotiations of generation 1.5 Korean-Canadian university students. *TESL Canada Journal*, 29, 81.
- Kramp, M. K. (2004). Exploring life and experience through narrative inquiry. In K. deMarras & S. D. Lapan (Eds.), *Foundations for research: Methods of inquiry in education and the social sciences* (pp. 103–121). Mahwah, NJ: Erlbaum.
- Lather, P. (1991). *Getting smart: Feminist research and pedagogy within the postmodern*. New York: Routledge.
- Lazaraton, A. (2003). Evaluative criteria for qualitative research in applied linguistics: Whose criteria and whose research? *Modern Language Journal*, 87(1), 1–12.
- LeCompte, M. D., & Schensul, J. J. (1999). *Designing & conducting ethnographic research*. Walnut Creek, CA: AltaMira.
- Leeman, J., Rabin, L., & Román-Mendoza, E. (2011). Identity and activism in heritage language education. *Modern Language Journal*, 95(4), 481–495.
- Lieblich, A., Tuval-Mashiach, R., & Zilber, T. (1998). *Narrative research: Reading, analysis and interpretation*. Thousand Oaks, CA: Sage.
- Lin, A. (2014). Critical discourse analysis in applied linguistics: A methodological review. *Annual Review of Applied Linguistics*, 34, 213–232.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.
- Liu, Y., & Xu, Y. (2011). Inclusion or exclusion? A narrative inquiry of a language teacher’s identity experience in the ‘new work order’ of competing pedagogies. *Teaching and Teacher Education*, 27(3), 589–597.
- Macintyre, P. D., Burns, C., & Jessome, A. (2011). Ambivalence about communicating in a second language: A qualitative study of French immersion students’ willingness to communicate. *Modern Language Journal*, 95(1), 81–96.
- Mahboob, A., Paltridge, B., Phakiti, A., Wagner, E., Starfield, S., Burns, A., ... De Costa, P. I. (2016). TESOL Quarterly research guidelines. *TESOL Quarterly*, 50(1), 42–65.
- Malsbary, C. B. (2014). “It’s not just learning English, it’s learning other cultures”: Belonging, power, and possibility in an immigrant contact zone. *International Journal of Qualitative Studies in Education*, 27(10), 1312–1336.

- Marianne. (2011). Reading books in class: What “just to read a book” can mean. *Reading in a Foreign Language*, 23(1), 17–41.
- Markee, N. (2000). *Conversation analysis*. Mahwah, NJ: Erlbaum.
- Markee, N. (2013). Emic and ethic in qualitative research. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–4). Chichester: Wiley-Blackwell.
- Maxwell, J. A. (2010). Using numbers in qualitative research. *Qualitative Inquiry*, 16(6), 475–482.
- McCarty, T. L. (Ed.). (2014). *Ethnography and language policy*. New York: Routledge.
- Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. San Francisco: Jossey-Bass.
- Nero, S. (2015). Language, identity, and insider/outsider positionality in Caribbean Creole English research. *Applied Linguistics Review*, 6(3), 341–368.
- Nuske, K. (2015). Transformation and stasis: Two case studies of critical teacher education in TESOL. *Critical Inquiry in Language Studies*, 12(4), 283–312.
- Patton, M. Q. (2014). *Qualitative research and evaluation methods* (4th ed.). Thousand Oaks, CA: SAGE.
- Perry, K. H. (2011). Ethics, vulnerability, and speakers of other languages: How university IRBs (do not) speak to research involving refugee participants. *Qualitative Inquiry*, 17(10), 899–912.
- Pinnow, R. J., & Chval, K. B. (2015). “How much You wanna bet?”: Examining the role of positioning in the development of L2 learner interactional competencies in the content classroom. *Linguistics and Education*, 30, 1–11.
- Richards, K. (2003). *Qualitative inquiry in TESOL*. New York: Palgrave Macmillan.
- Richards, K. (2009). Trends in qualitative research in language teaching since 2000. *Language Teaching*, 42(2), 147–180.
- Richards, K. (2011). Case studies. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 207–221). London: Routledge.
- Rogers, R. (Ed.). (2011). *An introduction to critical discourse analysis in education*. New York: Routledge.
- Rudwick, S., & Parmegiani, A. (2013). Divided loyalties: Zulu vis-à-vis English at the University of KwaZulu-Natal. *Language Matters*, 44(3), 89–107.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50, 696–735.
- Schwandt, T. (2001). *Dictionary of qualitative inquiry*. Thousand Oaks, CA: Sage.
- Sealey, A. (2013). Realism. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–6). Chichester: Wiley-Blackwell.
- Seedhouse, P. (2004). *The interactional architecture of the language classroom: A conversation analysis perspective*. Malden, MA: Blackwell.
- Shin, D. S. (2014). Web 2.0 tools and academic literacy development in a US urban school: A case study of a second-grade English language learner. *Language and Education*, 28(1), 68–85.

- Shirvan, M., Rahmani, S., & Sorayyae, L. (2016). On the exploration of the ecology of English language teachers' personal styles in Iran. *Asian-Pacific Journal of Second & Foreign Language Education*, 1(1), 12–28.
- Siegel, A. (2015). Social epistemics for analyzing longitudinal language learner development. *International Journal of Applied Linguistics*, 25(1), 83–104.
- Spradley, J. P. (1980). *Participant observation*. New York: Holt, Rinehart and Winston.
- St. Pierre, E. A., & Jackson, A. Y. (2014). Qualitative data analysis after coding. *Qualitative Inquiry*, 20(6), 715–719.
- Stake, R. E. (2000). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 435–454). Thousand Oaks, CA: SAGE.
- Starfield, S. (2013). Researcher reflexivity. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Chichester: Wiley-Blackwell.
- Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: SAGE.
- Taguchi, N. (2014). Development of interactional competence in Japanese as a second language: Use of incomplete sentences as interactional resources. *Modern Language Journal*, 98(2), 518–535.
- Takeuchi, M. (2015). The situated multiliteracies approach to classroom participation: English language learners' participation in classroom mathematics practices. *Journal of Language, Identity & Education*, 14(3), 159–178.
- Tamim, T. (2014). The politics of languages in education: Issues of access, social participation and inequality in the multilingual context of Pakistan. *British Educational Research Journal*, 40(2), 280–299.
- Tracy, S. (2010). Qualitative quality: Eight “Big-Tent” criteria for excellent qualitative research. *Qualitative Inquiry*, 16(10), 837–851.
- Valmori, L., & De Costa, P. I. (2016). How do foreign language teachers maintain their proficiency?: A grounded theory approach. *System*, 57(1), 98–108.
- Warriner, D. S. (2013). “It’s better life here than there”: Elasticity and ambivalence in narratives of personal experience. *International Multilingual Research Journal*, 7(1), 15–32.
- Waterhouse, M. (2012). ‘We don’t believe media anymore’: Mapping critical literacies in an adult immigrant language classroom. *Discourse: Studies in the Cultural Politics of Education*, 33(1), 129–146.
- Whyte, S. (2011). Learning to teach with videoconferencing in primary foreign language classrooms. *ReCALL*, 23(3), 271–293.
- Wyatt, M. (2013). Overcoming low self-efficacy beliefs in teaching English to young learners. *International Journal of Qualitative Studies in Education*, 26(2), 238–255.
- Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: Sage.
- Zabrodskaia, A. (2014). Tallinn: Monolingual from above and multilingual from below. *International Journal of the Sociology of Language*, 2014(228), 105–130.

- Zhang, H. (2013). Pedagogical challenges of spoken English learning in the Second Life virtual world: A case study. *British Journal of Educational Technology*, 44(2), 243–254.
- Zheng, X., & Borg, S. (2014). Task-based learning and teaching in China: Secondary school teachers' beliefs and practices. *Language Teaching Research*, 18(2), 205–221.
- Zhu, W., & Mitchell, D. A. (2012). Participation in peer response as activity: An examination of peer response stances from an activity theory perspective. *TESOL Quarterly*, 46(2), 362–386.



# 5

## Mixed Methodology

Alison Mackey and Lara Bryfonski

### Introduction

The field of applied linguistics research has recently benefited from a surge of methodological innovations that have been accompanied by significant discussions about the productive use of combining methods (Dörnyei, 2007; Hashemi & Babaii, 2013; Mackey, 2015; Moeller, Creswell, & Saville, 2015). While the field of applied linguistics has unquestionably advanced through both tightly controlled laboratory studies (e.g., Johnson & Newport, 1989) and deep ethnographic work (e.g., Cekaite, 2007), it is clear that neither experimental methods nor qualitative research alone can account for many of the questions we explore in our field. The addition of qualitative methods to traditionally quantitative domains can provide quantitative research with a more authentic lens to view language processes in context. We argue, like several others (e.g., Hashemi & Babaii, 2013), that applied linguistics research that approaches questions of language with a range of methods in mind is important as we advance our understanding of how languages are learned, taught, and used in a variety of contexts and domains. Mixed methods research (MMR), also known as multi-method or multi-methodology, employs aspects of both quantitative and qualitative methods and designs to better understand a given phenomenon. In the current chapter, we aim to explain how to utilize quantitative and qualitative data to complement one another to shed light on

---

A. Mackey (✉) • L. Bryfonski  
Georgetown University, Washington, DC, USA  
e-mail: [mackeya@georgetown.edu](mailto:mackeya@georgetown.edu); [lbryfo@gmail.com](mailto:lbryfo@gmail.com)

important questions in our field. We first define and provide a historical overview of mixed methods research in field of applied linguistics. Next, we describe a variety of designs of mixed methods research along with practical advice on three main stages of mixed methods research: planning, implementation, and analysis. The third section explores some challenging ideas in this innovative methodological domain. We finish with an examination of limitations along with some proposals for future directions in mixed methodology research.

## Definitions

Mixed methods research—also known as multi-method, combined methods, mixed research, or triangulation (Mackey & Gass, 2015)—is a strategy of inquiry that allows the researcher to explore a research question from multiple angles potentially avoiding the limitations inherent in using one approach, quantitative or qualitative, independently. Quantitative methodology is traditionally characterized by carefully controlled experimental design and random assignment, while qualitative methodology is characterized by grounded theory, case studies, and detailed description. Quantitative research aims for reliable and replicable design with outcomes that can be generalized across a population. Qualitative data focuses on processes rather than outcomes in order to understand a problem deeply and thoroughly (Mackey & Gass, 2015). Researchers implementing mixed methods approaches generally acknowledge there can be biases inherent in utilizing only one of these methods and attempt to address these biases by combining both strategies.

Mixed methods research has been a prevalent methodology in the sciences for many years (see, e.g., Creswell, 2015), yet it has only been clearly defined as mixed methods in the applied linguistics literature in the past decade (as noted by Hashemi & Babaii, 2013 among others). Over this time, mixed methods research has been defined and described in many ways. Mixed methodology is sometimes referred to as the integration of both quantitative and qualitative methods within a single study. In applied linguistics research, the most common way mixed methods are described is “triangulation” of multiple sources or methods (Mackey & Gass, 2015); however, some researchers (e.g., Creswell, 2003) define triangulation as a specific subtype of mixed methods design (see mixed methods research types below). Mixed methodology has also been defined as “the concept of mixing different methods” and “collecting and analyzing both forms of data in a single study”

(Creswell, 2003, p. 15). According to Tashakkori and Creswell (2007), mixed methodology can be described as “research in which the investigator collects and analyzes data integrates the findings, and draws inferences using both qualitative and quantitative approaches in a single study or program of inquiry” (p. 4).

## Current Issues

### The Qualitative-Quantitative Dichotomy

The idea of a dichotomy between quantitative and qualitative research methods has been routinely debated in the field, often dividing researchers across epistemological stances as well as methodological approaches (Riazi, 2016; Trochim, 2006). More recently, however the field has been adopting and adapting mixed methodologies as one of the approaches for research into how languages are learned (King & Mackey, 2016; Riazi & Candlin, 2014) including entire volumes dedicated to mixed methods approaches in applied linguistic research (Riazi, 2017).

Recent issues of the journal *Studies in Second Language Acquisition* (35(3), 2014) on bridging approaches to SLA research, as well as the 100th anniversary edition of *The Modern Language Journal* (100(S1), 2016) which included an introduction to “layered” approaches to theories and methods of SLA have addressed the issue of methodology in applied linguistic research. In the introduction to the 100th anniversary issue, King and Mackey (2016) argue that “taking such a cross-field, collaborative perspective is essential for us to fully and adequately address pressing, unresolved problems on both academic and practical fronts” (p. 210) citing mixed methods approaches as one means in which to address challenges facing applied linguistics researchers and language practitioners and arguing that not only should researchers blend methodologies but also *layer* epistemological perspectives. In this way, researchers can utilize multiple methods and stances to facilitate the involvement of both cognitive and social factors present in a variety of language processes. Introspective measures such as think-aloud and stimulated recall protocols (see Gass & Mackey, 2016) when used in combination with in-depth interviews and naturalistic observations, supplemented by quantitative data from new technologies such as eye trackers (see Smith, 2012) are one of many ways to layer perspectives and methods. Alternatively, quantitative data can be utilized to supplement qualitative data for example when little is known about a given research area



prior to investigation. This is common in classroom action research (see Loewen & Philp, 2012 for examples of classroom action research), where an instructor first observes students prior to collecting quantitative data such as exam scores.

## Why Choose a Mixed Methodology Design?

Mixed methods designs enable researchers to investigate the same phenomenon from multiple angles. When deciding whether or not to implement a mixed methods design, a researcher must first consider the research question(s) and then ask what kinds of data collection, analyses, and interpretations will best enable them to answer those questions. There is frequently no single “right” answer to the question of which method, and mixing them allows the researcher the freedom not to choose. Depending on the type of design chosen, the blending of multiple methods can enable researchers to explore and validate under-researched phenomena, shed light on difficult-to-interpret findings, address a problem from multiple theoretic perspectives, or to simply conceptualize a problem more deeply and thoroughly.

## Types of Mixed Methods Studies

Although mixed methods designs are typically categorized as either concurrent or sequential depending on the order in which data collection takes place (Creswell, 2003), there is no universally accepted way of designing and implementing a mixed methods study. Many varieties of mixed methods utilize multiple ways of incorporating quantitative and qualitative data. In mixed methods designs, the results from one type of inquiry (more qualitative or more quantitative or primarily descriptive) can be used to inform or understand the results from another method of inquiry. Methods might be integrated at every stage of the design process or might follow each other in a predetermined sequence (see Table 5.1 for an overview).

In any case, it is helpful for applied linguistics researchers to familiarize themselves with the idea that there a wide range of mixed methodology designs are possible before selecting the type that will be most appropriate for their research project since, clearly, some design types lend themselves more readily to applied linguistics research questions than others. In the section, we define and describe several common designs utilized in mixed methods

**Table 5.1** Common types of mixed methods designs

Design types	Description
Concurrent designs	Quantitative and qualitative data collected simultaneously. Collective interpretation of findings
<ul style="list-style-type: none"> <li>• <i>Triangulation</i></li> <li>• <i>Concurrent embedded</i></li> </ul>	<p>All data collected in one phase and weighted equally in analysis</p> <p>All data collected in one phase, but not weighted equally in analysis</p>
Sequential designs	Quantitative and qualitative data collected via multiple phases of research implementation
<ul style="list-style-type: none"> <li>• <i>Explanatory</i></li> <li>• <i>Exploratory</i></li> <li>• <i>Sequential embedded</i></li> </ul>	<p>Quantitative data are collected first. Qualitative data are collected as a follow-up</p> <p>Qualitative data are collected first. Quantitative data are collected as a follow-up</p> <p>Quantitative and qualitative data are collected in multiple iterations to inform and follow-up on one another</p>

research utilizing examples from studies in the field of applied linguistics to illustrate each design type.

## Concurrent Designs

In concurrent designs, the researcher collects both quantitative and qualitative data at the same time and uses the data collectively as a means to interpret the findings of the investigation (Creswell, 2003). Concurrent designs can be further broken down into triangulation designs and embedded concurrent designs.

Triangulation designs were one of the first strategies utilized in mixed methodology research (Creswell, 2003; Jick, 1979). In this type of design, all the data are collected in one phase and receive equal weight during analysis. This allows the researcher to “compare and contrast quantitative statistical results with qualitative findings or validate or explain quantitative results with qualitative data” (Creswell & Plano Clark, 2011, p. 62). For example, a questionnaire that aims to measure a learner’s contact with the target language might include both quantitative Likert-type questions asking them how many hours per week they speak the target language, as well as open-ended responses that elicit more qualitative data such as the domains or tasks in which learner interacts in the language, such as “who do you use the language with outside of class?” Since both types of data are collected within the same survey (i.e., concurrently), the results are triangulated from both qualitative and quantitative data (see Sample Study 5.1 for an example).

### Sample Study 5.1

Lee, E. (2016). Reducing international graduate students' language anxiety through oral pronunciation corrections. *System*, 56(1), 78–95.

#### Research Background

This mixed methodology study investigates the link between the type of oral corrective feedback used and language anxiety level.

#### Research Problems

The research addressed two main questions: What patterns of corrective feedback and learner repair occur in advanced-level adult ESL classrooms? How does oral corrective feedback affect students' anxiety about speaking English?

#### Research Method

- *Type of research*: Concurrent triangulation
- *Setting and participants*: Sixty advanced ESL university students from a variety of language backgrounds participating in an international teaching assistant training program took part in the study.
- *Instruments/techniques*: Both quantitative data obtained from a pre- and post-survey and classroom observations as well as qualitative data from follow-up interviews and open-ended survey questions were collected.
- *Data analysis*: Classroom instruction was recorded for one month and corrective feedback moves were coded by type and by the presence or absence of learner repairs. Pre- and post-surveys collected data on learners' self-perceptions of anxiety, attitude, motivation, and self-confidence. Post-instruction interviews asked follow-up questions about their affective stances toward corrective feedback.

#### Key Results

Findings from this study drew upon data collected by both quantitative and qualitative means. Quantitatively, learners' reports of anxiety decreased from pre- to posttest. Qualitative data revealed that learners' reactions varied based on the types of feedback they received. Specifically, clarification requests were associated with increases in students' language learning anxiety.

#### Comments

The concurrent triangulation of quantitative and qualitative methods shed light on how corrective feedback can affect learners' emotional states even at advanced levels of acquisition.

Concurrent embedded designs are similar to triangulated designs in that both qualitative and quantitative data are collected simultaneously. However, one type of data are embedded, generally meaning that it plays a secondary role to the other data, rather than both being weighed equally. In most research of this type, the focus of the design is quantitative with qualitative data collection added to support or better interpret the quantitative findings (Creswell & Plano Clark, 2011). Embedded designs are primarily utilized when the researcher is interested in measuring the impact of an intervention as well as

understanding the experience of the intervention (Mackey & Gass, 2015). For example, an SLA researcher interested in the effects of task type on learners noticing of corrective feedback provided by their instructor might have students fill out uptake sheets while performing various tasks. These uptake sheets (see Mackey, 2006, for an example) elicit quantitative data on the number of times the learner noticed the provision of feedback, but also more qualitative data on what stage of the task they were in, what kinds of feedback they were provided or even introspectively, how the feedback affected their emotional state. In this example, the primary data would be a comparison of the quantitative data with task type; however this data would be supported and strengthened when combined with the qualitative data giving insights into the learners' responses to the intervention.

## Sequential Designs

In contrast to concurrent designs, sequentially designs involve the collection of data in multiple phases or in a set sequence. Creswell and Plano Clark (2011) break down sequential designs into three categories: explanatory, exploratory, and sequential embedded.

### Explanatory Design

Explanatory designs focus on quantitative data but use qualitative follow-up data to explain quantitative results. Researchers might use an explanatory design to better understand quantitative results, especially if those results were not expected. For example, if the outcomes of a task-based intervention showed no changes in the fluency or complexity of learners' speech in the target language, a stimulated recall interview conducted after the intervention might elucidate some of the reasons for why no changes occurred. For example, learners might not have understood the task instructions or might not have been motivated to complete the task. Without exploring what goes on during a treatment, researchers may lose some critical context. Qualitative data collection can fill this gap (see Sample Study 5.2 for an example of qualitative analyses used to follow-up on quantitative findings). Sometimes researchers will follow-up with qualitative data collection on selected participants whose results are statistical outliers to better understand why those learners did very well or very poorly in comparison to the other participants. However, in all cases of explanatory designs, the results from the quantitative phase are typically emphasized over the supplementary qualitative data (Mackey & Gass, 2015).

### Sample Study 5.2

Demmen, J., Semino, E., Demjen, Z., Koller, V., Hardie, A., Rayson, P., & Payne, S. (2015). A computer-assisted study of the use of violence metaphors for cancer and end of life by patients, family carers and health professionals. *International Journal of Corpus Linguistics*, 20(2), 205–231.

#### Research Background

The authors of this study utilize a combination of corpus linguistics techniques and qualitative thematic analyses in their in-depth investigation of how metaphors of violence are utilized by patients, healthcare professionals, and family carers in discussing cancer and end of life.

#### Research Gaps

The study uses corpus linguistics in order to shed light on the use of violence metaphors in communication about illness and healthcare in a more rigorous and systematic manner than was the case in previous research.

#### Research Method

- *Type of research*: Sequential explanatory
- *Setting and participants*: The study consisted of a 1.5-million word corpus constructed from semi-structured interviews with hospice healthcare professionals, patients diagnosed with terminal cancer, and unpaid family carers looking after family members with terminal cancer. The corpus was primarily made up of data from online forums where healthcare professionals, patients, and family members participated.
- *Instruments/techniques*: Data were compiled into a corpus for further analysis.
- *Data analysis*: First a manual analysis of a sample from the corpus was used to identify and tag all metaphorical expressions utilized to refer to cancer by participants. The results were used to inform a computer-aided quantitative analysis of the whole corpus to examine the distribution of violence metaphors by stakeholders and by context (interview or online forum) and to create lists of the most frequently used violence metaphors by the various stakeholders. The resulting lists were then used in a qualitative exploration of the different ways in which members from the stakeholder groups used violence metaphors, as well as what aspects of the stakeholders' experiences were highlighted through the use of violence metaphors. This qualitative analysis in addition to the quantitative descriptive corpus analysis provided deeper understanding as to when and why these violence metaphors were used in the data collected.

#### Key Results

Findings indicated that patients, carers, and professionals utilize a wider array of violence metaphors than previously identified. Results also showed how metaphor use varied between contexts of interaction (online forums vs. interviews) and by stakeholder groups.

#### Comments

This study provides a unique example of how multiple methods (quantitative followed by qualitative) as well as multiple linguistic analysis techniques (corpus linguistics and qualitative thematic analysis) can shed light on complex questions in applied linguistics.

## Exploratory Design

Exploratory designs are the other side of the coin to explanatory designs. Here, qualitative data are collected first in order to guide subsequent collection of quantitative data. This type of mixed methodology allows researchers to begin to ask questions which might result in answers that can generalize the findings of results from qualitative data collection. Researchers might use an exploratory design to investigate an under-explored phenomenon and define and describe variables before embarking on a quantitative design. Classroom action research (see Mackey & Gass, 2015) might utilize an exploratory design. For example, a classroom language teacher might notice their students seem to be more engaged during different aspects of a lesson. The instructor might then decide to collect some qualitative data on learners' perceptions about their own levels of engagement during different tasks. A helpful quantitative follow-up might be designed to investigate quantitatively whether learners perform better on tasks they report they are more engaged in. This two-phase model allows the researcher or language teacher to develop an understanding of the research question through thematic analyses or other qualitative means and then connect those analyses to quantitative findings. An example of an exploratory design in a classroom context is illustrated in Sample Study 5.3.

### Sample Study 5.3

Mroz, A. (2015). The development of second language critical thinking in a virtual language learning environment: A process-oriented mixed-method study. *CALICO Journal*, 32(3), 528–553.

#### Research Background

The researcher in this study utilized a mixed methods design to investigate the emergence of L2 strategies in collaborative discourse produced by university French learners interacting in a computer-mediated setting.

#### Research Gaps

The research design was motivated by a call for more qualitative data informing the results from CALL (computer-assisted language learning) studies and included a perspective layered by both sociocultural theory and ecological perspectives.

#### Research Method

- *Type of research:* The author defines the study as an embedded design utilizing a data transformation procedure. The design of the study is primarily qualitative with an embedded quantitative element.
- *Setting and participants:* Two undergraduate intermediate French classes consisting of 27 students. Five students volunteered to act as case-study students.

- *Instruments/techniques*: Qualitative data included bi-weekly observations of French classes, a focus group with five case-study students, and retrospective perception data. Quantitative data were elicited from computer-generated logs of students' interactions in an online Second Life community.
- *Data analysis*: The author integrated the findings from both quantitative and qualitative data to examine student perceptions of the impact of interaction in the online community on their critical thinking skills in their second language.

#### **Key Results**

Results indicated an increase in the use of higher-order critical thinking abilities during interaction in the computer-mediated setting. Qualitative themes shed light on the students' reactions to the collaborative learning environment and the use of technology for language learning.

#### **Comments**

This article clearly outlines its design and rationale for the use of mixed methodology. The integration of quantitative data into stages of a primarily qualitative, case-study design makes this a clear example of a sequential exploratory design.

## **Sequential Embedded Design**

In the sequential embedded design, quantitative and qualitative data are collected in multiple phases to both inform and provide insights into potential explanations for one another. As in concurrent embedded designs, one data type (again, in the applied linguistics field, this has been typically quantitative data) is primary. Qualitative data collected before an intervention can aid in participant selection, instrument verification, or to shape the subsequent intervention. Qualitative data collected after an intervention can help to illuminate patterns in quantitative intervention data. The integration of both quantitative and qualitative methods at each stage of a research design is shown in Sample Study 5.4.

### **Sample Study 5.4**

Préfontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *Modern Language Journal*, 99(1), 96–112.

#### **Research Background**

The authors of this study utilized a mixed methods approach to shed light on the relationship between task difficulty and fluency.

**Research Problem**

The researchers asked “What quantitative and qualitative differences exist in the perceived difficulty of narrative tasks in French as an L2?” (p. 99) as well as several other questions regarding the variation in L2 fluency across the tasks and the relationship between perceived difficulty and fluency performance.

**Research Method**

- *Type of methodology*: Sequential embedded.
- *Setting and participants*: Forty adult learners of French studying in an immersion context participated in one of three task conditions.
- *Instruments/techniques*: The three conditions all included a narrative task that varied in terms of the amount of demands it posed on learners’ creativity, content knowledge, and linguistic resources.
- *Data analysis*: Task difficulty was measured both quantitatively with questionnaire data as well as qualitatively with retrospective interview data. Fluency was measured quantitatively using a variety of fluency measures analyzed in Praat.

**Key Results**

The use of these multiple methods allowed researchers to capture a holistic account of the phenomenon under consideration; quantitative questionnaire data demonstrated a link between lexical retrieval difficulty and fluency difficulty that was related to perceived task difficulty, while qualitative interview data explained how the learners evaluated the difficulty of the tasks and also which task features affected their fluency. Quantitative fluency data showed that articulation rate and average pause time were related to task difficulty.

**Comments**

The seamless integration of both quantitative and qualitative methods is clear from the researchers’ research question stated above—the hallmark of an embedded mixed methods design. Since the authors collected data in several phases, this is considered a sequential embedded method.

## New Trends in Mixed Methods Designs

In addition to the mixed methodology models described above, several novel models have been successfully applied to research in the field of applied linguistics. For example, Hashemi and Babaii (2013) examined 205 SLA research articles via a content-based analytic approach to identify current trends in mixed methodology SLA research. They found that in general, concurrent designs are more prevalent than sequential designs and that while many researchers used mixed methodology designs they did not achieve “high degrees of integration at various stages of the study as a quality standard for mixed research” (p. 828). Another finding was the existence of mixed methods research designs that did not fit the typology proposed by Creswell, Plano



Clark, and Garrett (2008). For example, Mackey (2006) utilized a primarily quantitative design with qualitative elements embedded and with follow-up qualitative data collection. Specifically, a pre-/post-test design was utilized to measure noticing and learning of new forms, with a qualitative learning journal method embedded. Stimulated recall interviews and follow-up questionnaires represented explanatory qualitative data. This study, therefore, included both concurrent and sequential features. Another example presented by Hashemi and Babaii (2013) was a needs analysis of Egyptian engineering students conducted by Pritchard and Nasr (2004). The aim of this study was to use the results of the needs analysis to design a reading improvement program for the engineering students. The study began as an exploratory study with qualitative data collected on student needs. The researchers then followed this qualitative information with a quantitative establishment of the validity and reliability of the development materials. The final phase involved embedded quantitative and qualitative data collection to obtain feedback from teachers and students.

## Doing Mixed Methods in Applied Linguistics Studies

After selecting a mixed methods strategy that suits the research questions under investigation, there are several considerations that should be attended to during the planning, implementation, and analysis stages of the mixed methods research.

### Planning

When in the planning stages of the study design, the researcher should be able to answer and justify several decisions including:

1. Will the design give equal weight to both quantitative and qualitative findings or theoretical perspectives?
2. Will the same participants contribute to all phases of the project?
3. Is there a clear justification for the design choice (i.e., concurrent or embedded) stemming directly from the research questions?

Mackey and Gass (2015) suggest sketching a visual diagram or representation of the research design as a useful strategy for visualizing each stage of the project. Within the sketch the research team can specify procedures and

expected outcomes for each phase of the study. This can be useful both for the researchers as they implement the study as well as for the readers of the final report of the project if the design is incorporated into the write-up.

## Implementation

Implementation is the stage where the design is applied and the data are collected. The order in which data are collected will, of course, depend on whether or not the researcher has chosen a concurrent or sequential design. The issue of multiple sources should also be considered, alongside multiple methods. For example, in task-based language teaching studies, researchers are urged to not only collect data on language program needs from teachers and administrators but also from students themselves. Decisions about methods and sources can have bearing on the ultimate validity and reliability of the study outcomes (as further discussed in the section on challenges, below).

## Analysis

The analytic techniques and methods chosen depend heavily on the research questions and mixed methods strategy used to investigate the research questions. When conducting and reporting results and outcome, researchers should verify they can answer the following questions: (a) how will my analyses complement or justify one another? And, (b) will the analyses be integrated or kept separate? It is likely that in the mixed methods research design that data will be collected from multiple sources as well as via multiple data collection techniques, for example, a researcher might triangulate data from interviews of both teachers and students as well as from surveys or questionnaires. In order to best represent the wealth of data that is obtained in mixed methods studies, researchers might need to consolidate data, meaning refine, combine, or boil data down to a more concise format. They may also need to transform the data, which means to modify data so that it can be compared across sources. While we often take it for granted that quantitative data can be represented visually through the use of line plots, box plots, and histograms, qualitative results can also be represented visually. Descriptive matrices (see Miles, Huberman, & Saldana, 2014) or data visualization techniques made possible by qualitative data software like QSR International's NVivo software (see Richards, 1999) can help readers better understand and compare the results from qualitative stages to make sense of the overall findings.

## Challenges in Mixed Methods Research

### Pros and Cons to Mixed Methods Designs

An obvious advantage to mixed methods designs is the way that triangulation of sources helps to shed greater light on a problem. However, occasionally, there are challenges to integrating data from both types of approach into one report. This is particularly the case when the findings turn out to be contradictory. For example, a pretest-posttest design explanatory design may show a significant result in terms of an advantage for one type of treatment, let's say an instructional intervention. In order to provide more details and better understand the result, the researchers build into the design a provision to randomly select some of the participants for focal case studies. If the original experimental group consisted of twenty-five participants (with a similar number in the control group), three participants might be selected as focal case-study participants. However, it is possible that none, or not all of them, follow the same patterns as in the group data. So the data would seem to be contradictory. This situation illustrates how helpful it can be to mix methods, by showing that quantification and aggregation of data can often obscure interesting patterns at the individual level. To mitigate against a confusing report, researchers need to dig deeply into the case studies and use them to explain how improvements are taking place in the group as a whole, but this is not always true for every individual in the group. Case-study data can help explain why, to some extent. For example, a study that provides an in-depth ethnographic investigation of a select group may seek to expand the study to include quantitative findings of group trends, to promote generalizability to a wider population. Case-study findings may contradict findings from quantitative results. These issues must be thoughtfully and carefully described and explained in any research report so that research consumers can understand the nuances to the results from each element in the mixed methods study.

These examples illustrate the importance of making initial decisions about which type of data are used to explain or complement the other. As Abbuhl and Mackey (2015) noted, "Some researchers claim that the two approaches are epistemologically and ontologically incompatible, making split methods little more than unprincipled methodological opportunism" (p. 8). The idea that quantitative and qualitative methods are inherently incompatible is not the position we take, although it is important to guard against it by designing research carefully, and not falling into the trap of presenting a primarily

quantitative study with a token case study at the end or of presenting a statistical analysis of some descriptive data at the beginning of an ethnography—neither of these approaches would qualify as genuinely mixed methods, but rather primarily one or the other, while borrowing a technique from a different paradigm. It is important for true mixed methods study to be designed as such from the outset, recognizing the need for both approaches to fully address the research questions or problem area. While there is no “ideal” mixed methodology study since all research must step from a particular problem, we can say that a sequential embedded design type using both quantitative and qualitative data collection at all stages of the process is an important goal. For example, a tightly controlled experiment that uses integrated introspective and interview methods, where the introspective and interview methods are carefully intertwined.

### **Limitations of Mixed Methods Designs**

Mixed methods studies, while an important and useful approach to investigating questions in our field, are, as is the case for all research, not without limitations. Mixed methods research typically requires more resources in the sense of time, both from the participants and from the researchers. Essentially, researchers need the best of both worlds for a mixed methods study to work, so both the extensive experimental methods, using careful development of procedures and materials, and controlling and counterbalancing variables, together with the intensive integrated and grounded approach to individual participation and understanding that comes with qualitative research and time investment. It is also the case that domain expertise needs to be considered. Many researchers are trained in one paradigm or the other and indeed, naturally gravitate to one or the other in terms of their preference and comfort level for data collection and analysis. Team approaches are helpful in this regard, as well as further training and education. Even when we consider the final outcomes, primarily quantitative research is usually written up in journal articles or book chapters, and qualitative research is more often written up in more lengthy publications, such as monographs. This disparity in outcomes also needs to be considered. Having said all this, it seems logical that using multiple sources and data types to address questions in the field and not being wedded to one paradigm or another will advance knowledge, and the recent increase in interest in mixed methods research supports this.

## Resources for Further Reading

Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University Press.

This textbook provides a useful and practical overview of the basics of research methodology with a focus on combining quantitative and qualitative methods judiciously in TESOL research projects. Chapters provide depth on the key stages in the development of a mixed methods study including: planning a project, gathering and analyzing data, interpreting results, and reporting results. The book focuses in on the domains of classroom research, action research, conversation, and discourse analysis as well as survey-based research and language program evaluations.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Los Angeles, CA: SAGE.

For those interested at examining methodology from a broader social science-oriented perspective, this textbook provides a step-by-step look at all aspects of the mixed methodology research process. This text is especially useful for the many diagrams of the various types of mixed methods designs described in the current chapter and provides greater depth for each of the main mixed methods designs.

Ivankova, N. V., & Greer, J. L. (2015). Mixed methods research and analysis. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource*. London: Bloomsbury.

This practical book chapter defines and describes mixed methods research in manner that is accessible to both undergraduate and postgraduate students studying applied linguistics. The chapter provides an overview of when and why mixed methods research is utilized, philosophical assumptions and issues associated with mixed methods research as well as a breakdown of the steps of implementing a mixed methods study.

*Journal of Mixed Methods Research* (JMMR; <http://mmr.sagepub.com/>).

This peer-reviewed journal focuses on mixed methods research in fields of social, behavioral, health, and human sciences from an international

perspective. The journal is supported and contributed to by leading mixed methodology researchers including John Creswell, Abbas Tashakkori, Charles Teddlie, and Michael Patton among many others.

King, K. A., & Mackey, A. (2016). Research methodology in second language studies: Trends, concerns, and new directions. *Modern Language Journal*, 100(S1), 209–227.

This introduction to the 100th anniversary edition of *The Modern Language Journal* offers an in-depth discussion of the importance of mixing and layering methods and perspectives in order to further the field of applied linguistics. The article summarizes some new advances in methodology and points to future directions for second language and applied linguistics researchers.

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design* (2nd ed.). London: Routledge.

The second edition of this handbook for second language acquisition researchers includes practical advice for quantitative, qualitative, and mixed methods studies. Chapter 9 on mixed methods provides overviews of the main types of mixed methods designs used in second language acquisition research and additionally provides activities and discussion questions that can be used to further deepen understandings about mixed methods research.

Riazi, A. M. (2017). *Mixed methods research in language teaching and learning*. London: Equinox publication Company.

This recent publication is the first in the field to bring together the current body of mixed methodology research in language teaching and learning. The book takes a practical approach in order to promote the use of mixed methods by applied linguistics researchers. The challenges of designing and implementing mixed methods research are also addressed. This is a useful resource for doctoral students, postgraduates, or others developing research proposals that integrate quantitative and qualitative methods.

## References

- Abbuhl, R., & Mackey, A. (2015). Second language acquisition research methods. In K. King & Y. Lai (Eds.), *Encyclopedia of language and education: Vol. 10. Research methods in language and education* (3rd ed.). Dordrecht: Springer.
- Cekaite, A. (2007). A child's development of interactional competence in a Swedish L2 classroom. *Modern Language Journal*, 91(1), 45–62.
- Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches*. Thousand Oaks: SAGE.
- Creswell, J. W. (2015). *A concise introduction to mixed methods research*. Los Angeles, CA: SAGE.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Los Angeles, CA: SAGE.
- Creswell, J. W., Plano Clark, V. L., & Garrett, A. L. (2008). Methodological issues in conducting mixed methods research designs. In M. M. Bergman (Ed.), *Advances in mixed methods research* (pp. 66–83). London: SAGE.
- Demmen, J., Semino, E., Demjen, Z., Koller, V., Hardie, A., Rayson, P., & Payne, S. (2015). A computer-assisted study of the use of violence metaphors for cancer and end of life by patients, family carers and health professionals. *International Journal of Corpus Linguistics*, 20(2), 205–231.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.
- Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in second language research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hashemi, M., & Babaii, E. (2013). Mixed methods research: Toward new research designs in applied linguistics. *Modern Language Journal*, 97(4), 828–852.
- Jick, T. (1979). Mixing qualitative and quantitative methods: Triangulation in action. In J. Van Maanen (Ed.), *Qualitative methodology* (pp. 135–148). Beverly Hills: SAGE.
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1), 60–99.
- King, K. A., & Mackey, A. (2016). Research methodology in second language studies: Trends, concerns, and new directions. *Modern Language Journal*, 100(S1), 209–227.
- Lee, E. (2016). Reducing international graduate students' language anxiety through oral pronunciation corrections. *System*, 56(1), 78–95.
- Loewen, S., & Philp, J. (2012). Instructed second language acquisition. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (1st ed., pp. 55–73). Oxford: Blackwell Publishing Ltd.
- Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, 27(3), 405–430.

- Mackey, A. (2015). Methodological practice and progression in second language research. *AILA Review*, 27, 80–97.
- Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design* (2nd ed.). London: Routledge.
- Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook*. Thousand Oaks, CA: SAGE.
- Moeller, A. J., Creswell, J. W., & Saville, N. (Eds.). (2015). *Language assessment and mixed methods*. Cambridge: University of Cambridge Press.
- Mroz, A. (2015). The development of second language critical thinking in a virtual language learning environment: A process-oriented mixed-method study. *CALICO Journal*, 32(3), 528–553.
- Préfontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *Modern Language Journal*, 99(1), 96–112.
- Pritchard, R. M. O., & Nasr, A. (2004). Improving reading performance among Egyptian engineering students: Principles and practice. *English for Specific Purposes*, 23(4), 425–445.
- Riazi, A. M. (2016). Innovative Mixed-methods Research: Moving Beyond Design Technicalities to Epistemological and Methodological Realizations. *Applied Linguistics*, 37(1), 33–49.
- Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching*, 47(2), 135–173.
- Richards, L. (1999). *Using NVivo in qualitative research*. Thousand Oaks, CA: SAGE.
- Smith, B. (2012). Eye tracking as a measure of noticing: A study of explicit recasts in SCMC. *Language Learning & Technology*, 16(3), 53–81.
- Tashakkori, A., & Creswell, J. W. (2007). Editorial: The new era of mixed methods. *Journal of Mixed Methods Research*, 1(1), 3–7.
- Trochim, W. M. (2006). *The research methods knowledge base* (2nd ed.). Retrieved from <http://www.socialresearchmethods.net/kb/>





# 6

## Traditional Literature Review and Research Synthesis

Shaofeng Li and Hong Wang

### Introduction

A literature review is a retrospective account of previous research on a certain topic, and it may achieve various purposes. For researchers, a literature review may serve to contextualize and inform further research. Specifically, prior to carrying out a new study, the researcher needs to find a *niche* by identifying what has been done on the topic under investigation and to mine through existing methodology with a view to developing instruments and materials that can best answer his/her own research questions. Also, through a literature review, a researcher may draw on existing evidence to verify a theory or build a new theory. For practitioners and policy makers, the conclusions reached in a literature review based on aggregated research findings may serve as a basis for decision-making in terms of how to meet the needs of different stakeholders. Such integration of research and practice is

---

S. Li (✉)

Foreign and Second Language Education Program, School of Teacher Education,  
Florida State University, Tallahassee, Florida, USA

Center for Linguistics and Applied Linguistics, Guangdong University of Foreign  
Studies, Guangzhou, China

e-mail: [sli9@fsu.edu](mailto:sli9@fsu.edu)

H. Wang

University of Auckland Library and Learning Services, The University of Auckland,  
Auckland, New Zealand

e-mail: [wanghong19@gmail.com](mailto:wanghong19@gmail.com)

called evidence-based practice (Bronson & Davis, 2012), which is of particular importance in a heavily practice-oriented discipline such as applied linguistics.

Because of the importance of literature review, there has been a call to treat it as a research method in its own right (Cooper, 2016). In fact, in the *literature* on literature review, a distinction has been made between traditional literature reviews, such as the type that appears in the literature review section of a journal article, and more systematic approaches of previous research, such as meta-analysis, which is conducted following a set of well-defined procedures and protocols. Although there has been much discussion about the differences between traditional reviews and systematic reviews, systematic comparisons have been rare, and the terminology relating to the two approaches and to literature review as a genre has been ambiguous and confusing.

To clarify any potential ambiguities, and for the purposes of this chapter, the term *literature review* is used to refer to the genre as a whole as well as to traditional literature reviews. *Research synthesis* is reserved for systematic reviews such as meta-analysis. Proponents of research synthesis have generally been negative about traditional reviews, criticizing its unscientific nature. However, we argue that both traditional reviews and research syntheses have merits, that they serve different purposes, and that they do not have to be mutually exclusive. The following sections discuss the procedures and best practices of each of the two approaches and conclude by making a comparison between the two approaches and proposing ways to integrate them.

## Traditional Literature Review

Although anyone getting an academic degree or pursuing an academic career may have to write a literature review at some point, there has been surprisingly little information on how to do it. A quick look at research methods textbooks in applied linguistics shows that none of them includes detailed information on how to conduct a literature review. In a study by Zaporozhcz (1987), Ph.D. advisors ( $N = 33$ ) ranked the literature review section the lowest in terms of the amount of help they provided to their students—they reported spending the most time on the supervision of the methods chapter. The lack of interest in guiding students on how to do a literature review is probably because (1) it is considered an easy and transparent process, not a skill that needs to be trained, and (2) there are a myriad of ways of writing a literature review, which makes it challenging to provide a general guidance. However, it can be argued that (1) doing a literature review is not a naturally acquired skill, and (2) despite the variety of styles and approaches, there are some common principles and procedures one could follow in order to write a successful review.

Jesson, Matheson, and Lacey (2011) defined a traditional literature review as “a written appraisal of what is already known ... with no prescribed methodology” (p. 10). This definition suggests that a literature review is not a mere description of previous research; rather, it provides an evaluation of the research. The definition also distinguishes a traditional review from a research synthesis, which is carried out by following a set of well-defined procedures (see Plonsky & Oswald, 2015). Traditional reviews include the introductory sections of the study reports of empirical studies as well as freestanding reviews such as those published in *Language Teaching*. The main purpose of a literature review for an empirical study is to *set the stage* for a new study. The purposes of freestanding reviews, by contrast, are more diverse, such as providing a state-of-the-art review of the research on a certain instructional treatment, clarifying the myths and central issues of a substantive domain, proposing a new research agenda, and summarizing the methods previous researchers have used to measure a certain construct. Because of the diversity of topics and purposes of freestanding reviews, there are no fixed formats to follow. The focus of this section, consequently, will be on the literature reviews for empirical studies, which, of course, overlaps with freestanding reviews in many ways.

Before going into further detail, it is necessary to emphasize that what we are discussing here is how to *do*, not just how to *write*, a literature review; writing a literature review is the final step of the entire process of doing a literature review. The purpose of doing a literature review is trifold: (1) to contextualize the study to be conducted, (2) to inform the study design, and (3) to help the researcher interpret the results in the discussion section. Specifically, when contextualizing the current study, the researcher needs to identify what is known about the topic by discussing the related theories, research, and practices. The researcher also needs to identify what is unknown about the topic, explain how it is informed by, and deviates from, previous studies, and convince the reader of the significance of the current study. An equally important purpose of doing a literature review is to draw on the methodology of existing research to answer the research questions of the current study. Finally, doing a literature review enables the researcher to refer back to the theories and research expounded in the review section when discussing how the findings of this study may contribute to the existing body of knowledge.

The content of a literature review, or the scholarship to be evaluated, is of three types: conceptual, empirical, and practical. Conceptual knowledge concerns theories, including arguments, statements, claims, and terminology. Empirical knowledge refers to the findings of empirical studies as well as the methodological aspects of the studies. Practical knowledge can be divided into two types. One refers to the knowledge contributed by practitioners including (1) the findings of action research, such as those reported in articles published in the *ELT Journal* or in the practitioners’ research section of *Language Teaching*

*Research*; (2) guidelines and principles for effective practice, such as the information from teacher guides; and (3) opinions, debates, and discussions on public forums such as the Internet. The other type of practical knowledge pertains to policies and instructions formulated by government agencies to guide the practice of the domain in which the research is situated. These three types of knowledge correspond to three aspects of a research topic: theory (conceptual), research (empirical), and practice (practical). Although it is not a must to be all-inclusive, a literature review should minimally include the theories and research relating to the research topic. However, given the applied nature of our field, the review of the literature would appear incomplete if the practical dimension is left out.

## Stages

The process of doing a literature review can be divided into six stages, which are elaborated in the following sections.

### Stage 1: Defining the Problem

When doing a literature review, the first step is to define the research problem or formulate research questions for the study. Although research questions usually appear at the end of a literature review, in practice a researcher must have them at the beginning of the process. The research questions constitute a flagship guiding the literature review as well as other parts of an empirical study such as study design. Therefore, if the researcher is uncertain about where to start during a literature search, or even how to organize the literature review, the best way is to consider what questions the study seeks to answer and then find information on what theorists, researchers, and practitioners said about the questions. Although the research questions may be fine-tuned as the review process unfolds, they serve as starting points leading you towards the destination and guiding the literature search as well as later stages of the process.

### Stage 2: Searching for the Literature

After formulating the research questions, the next step is to search for the literature to be included in the review. The most common search strategy is to use electronic databases, including (1) domain-general databases (e.g., *Google Scholar*), (2) domain-specific databases in applied linguistics (e.g., *LLBA*), (3)

databases from neighbouring disciplines (e.g., *PsycINFO* in psychology), and (4) databases for Ph.D./M.A. dissertations and theses (e.g., *ProQuest Dissertations & Theses*). One emerging powerful source of information is public academic forums such as Academia ([www.academic.edu](http://www.academic.edu)), which are not only venues for academic communication between researchers but also large repositories of information. Another commonly used strategy is ancestry chasing, that is, mining the reference sections of primary studies (the term *primary* is used to distinguish studies synthesized in a literature review and the review itself) and review articles to find relevant items.

### Stage 3: Selecting Studies

Although a traditional literature review usually does not report how the studies included in the review are selected, the researcher must make decisions, albeit “behind the scene,” on which retrieved studies actually go into the review. Unlike a research synthesis, which must be inclusive, a traditional review is selective. Although the selection criteria are idiosyncratic, some general principles should be adhered to. The first principle is that the selected studies must be representative. Representativeness has two dimensions: influential and diverse. Influential studies refer to milestone or seminal studies on the topic under investigation, which are frequently cited and/or are published in prestigious venues such as journals with high impact factors. By diverse, it is meant the included studies must represent different perspectives, disparate findings, and varied contexts or populations. Critics of traditional reviews argue that the authors may include only studies that support a certain theory, show certain results, or are carried out with a certain methodology. Therefore, it is important to include studies that represent different theories and trends to reduce the likelihood of bias and arbitrariness. The second principle concerns relevance. Retrieved studies may be relevant to the research questions to varying degrees. Given the usually limited space for a literature review, it is important to include the most relevant research. The third consideration is study quality, that is, the reviewed studies should have high internal and external validity.

### Stage 4: Reading the Literature

Two principles should be kept in mind when reading the literature. The first is to read carefully and understand thoroughly. A common problem in doing a literature review is piecemeal reading (Booth, Colomb, & Williams, 1995),

which may cause incomplete or inaccurate understanding. There is no shortcut to a successful review, and familiarity with the literature is the only key. It is advisable to read key articles several times, read alternative explanations in the case of a difficult or complicated theory, and read all hallmark studies. The second principle is to read actively and critically instead of passively and mechanically. While reading, one should not assume that published research is perfect. Instead, consider whether the study in question was conducted using valid methods, whether the results are due to the idiosyncratic versus principled methods, how the study is similar to, and different from, other studies, whether the interpretations are warranted, and how the study informs one's own study in terms of what one can draw on, what improvements one wants to make, or whether one will reorient the focus of one's own study based on what has been learned from this study.

### **Stage 5: Organizing the Data**

Information derived from the retrieved studies constitutes data for the literature review and should be organized in two ways: discretely and synthetically. In a discrete organization, the details about each individual study are recorded in a table or spreadsheet, and this type of information can be labelled *study notes*. Study notes can be arranged alphabetically according to the authors' family names, and the notes about a primary study should include a brief summary of the study, followed by detailed information about the methods, results, and interpretations of the results. As to the literature review of a retrieved study, it would be useful to observe what theories or models the author refers to, how he/she defines the constructs, and how he/she summarizes and critiques previous research. However, the author's comments and interpretations of other studies might be biased and inaccurate, and therefore it is always advisable to read the original articles in case of any uncertainty. Finally, the study notes table should have a section for any comments the reviewer may have on any aspect of the study that merits further attention.

A synthetic organization involves extracting themes, patterns, or trends that have emerged from individual studies. Synthetic notes can be organized after the study notes about each individual study are in place, but it is easier and more efficient to work on both at the same time. Organizing synthetic notes entails categorizing the information from the primary studies and identifying and reflecting on the commonalities and disparities between them. In which way should the information be categorized depends on what has

emerged from the studies and whether categorizing the studies in a certain way leads to an interesting point or a convincing argument. Categorization can be based on study findings. For example, some studies may have found a certain instructional treatment to be effective, while others may not. It would then be necessary to divide them into two categories and ascertain what characteristics each group of studies share that lead to their respective and conflicting findings. In a similar vein, categorization can be done methodologically based on learner population, study context, measures of independent and dependent variables, and so on. Furthermore, methodological information collated from different studies can be the target of synthesis if one purpose of the review is to summarize the methodology of the primary studies. Finally, in addition to empirical knowledge, the synthetic notes should include sections for theoretical and practical knowledge so that all information needed for the review converges in one venue.

## Stage 6: Writing Up the Review

*Components of a literature review.* A literature review typically consists of three components: an introduction, the body of the review, and research questions. The length of the introduction ranges from one or two paragraphs (for a journal article) to a chapter (for a Ph.D. thesis), but the purpose is the same: to give “the reader a sense of what was done and why” (APA, 2001, p. 16). This section of the literature review should provide a succinct overview of the topic, spell out the key issues surrounding the topic, state the significance of the study, identify the gaps in knowledge, explain the aims of the study, and inform the reader of the structure of the literature review. The body of the literature review should contextualize the current study by summarizing the theory, research, and practice of the research topic and justifying the significance of the study. The body of the review leads towards the research questions, and therefore before the research questions are introduced, it is necessary to summarize previous findings and controversies and show the links between previous research and the current study.

*Structuring a literature review.* There is no fixed format as far as how a literature review should be organized, but given the separation between the three types of knowledge—conceptual, empirical, and practical—the macro structure may consist of three major parts dealing with the theories, research, and practices relating to the focus of the current study. For an empirical study, the bulk of the review should be empirical knowledge, namely, the findings and methods of empirical studies.

There are two ways to present the information contributed by empirical studies: thematic and anthological. Thematic presentation is based on the themes emerging from the primary studies (from the synthetic notes), and the flow of information proceeds through arguments. An argument is “the logical presentation of evidence that leads to and justifies a conclusion” (Machi & McEvoy, 2012, p. 65). An argument has three components: claim, evidence, and warrant (Booth et al., 1995). A claim should be substantive and contestable in order to arouse readers’ interest and contribute to the existing body of knowledge. The evidence cited to support the claim needs to be:

1. accurate, that is, incorrect evidence should not be used
2. precise, namely, be specific, avoid being vague, and hedge or use qualifiers if absolute preciseness is impossible
3. sufficient, meaning there should be enough evidence for the validity of the claim
4. authoritative, that is, evidence should be robust and influential
5. perspicuous, which means evidence should be clear and easy to understand

The warrant of an argument is the reasoning linking the evidence and the claim; it is about the logical connections between the evidence and the claim, not the evidence or claim *per se*. Therefore, if there are no logical links between the claim and the evidence, then the argument is not warranted, even though the evidence is sound. In a nutshell, if one elects to present the information about empirical studies thematically, the literature review would be built around different arguments in which claims or themes are reported with supporting evidence from multiple sources or studies.

Anthological presentation means that the review is organized as a collection of individual studies, reporting details on the methods and results of each study—similar to study notes. Although this practice is prevalent in the field of applied linguistics, it is less effective than thematic presentation because the primary objective of a literature review is to critique, synthesize, and show readers what to make of previous findings rather than create an annotated bibliography. This is not to say that we should root out detailed descriptions of individual studies; rather, study details are necessary when they are important for making an argument or when hallmark or seminal studies are reported due to the special place they occupy in a literature review. However, because of the critical nature of traditional reviews, study details, when reported, should be accompanied with comments and critique.

*The critical nature of a literature review.* As Imel (2011, p. 146) pointed out, a literature review “provides a new perspective on the topic ... [and is] more than



the sum of its parts.” Key to developing a new perspective is a critical assessment of the findings and methods of the relevant body of research. The following strategies can be utilized to make a literature review a critical piece of writing:

- Identify the differences and similarities between primary studies in terms of their findings and methods.
- Include all representative findings, not only those that are “cherry-picked” to support your own position. If there are disparities, it is better to explain rather than ignore. For opposing evidence that is uninterpretable, it is better to state that “certain studies support one conclusion and others support another” (APA, 2001, pp. 16–17), rather than provide an unconvincing interpretation.
- Challenge existing theories and research. If you disagree with an argument or claim on reasonable grounds, do not hide your position; if there is clear evidence showing the limitations of a previous study, point them out; propose new or alternative interpretations for previous findings. However, do not stigmatize previous research even though you have found loopholes and limitations.
- Discuss the significance of previous studies. If you have discovered merits of a view, finding, or method, it is important to make them known to the reader. Thus, being critical entails demonstrating not only the weaknesses of previous research but also their strengths and contributions.
- Evaluate and clarify controversies or opposing theoretical positions and discuss how they have influenced the research and practice of the substantive domain.
- Propose new directions, methods, or theories, which may complement, but not necessarily contradict or supersede, existing research.
- Use linguistic devices to show the relationships between ideas and describe the authors’ stances. For example, cohesive expressions, such as “similarly,” “in contrast,” “however,” “therefore,” and so on, are effective in making evident how information is related. Appropriate reporting verbs should be used to accurately capture the authors’ positions and stances, such as “contend,” “argue,” “state,” “observe,” “assert,” “report,” and so on.

## Research Synthesis

Research synthesis grew out of the dissatisfaction with traditional reviews, which are deemed unscientific and subjective. Cooper (2016) argued that like an empirical study, a research synthesis must meet rigorous methodological

standards in order for the conclusions to be trustworthy. Despite the different labels for research synthesis, experts seem to converge on the point that a research synthesis is comprehensive in coverage and transparent in reporting, and its purpose is to reach conclusions based on study findings, which may be used to guide practice and policy-making.

In applied linguistics, two types of research synthesis have emerged: methodological and substantive. Methodological synthesis provides a survey of one or more methodological aspects of the primary research with a view to evaluating whether current practices meet certain criteria and what improvements can be made. Plonsky and associates (e.g., Plonsky & Gass, 2011) have published a series of methodological syntheses assessing the study quality of the primary research in second language acquisition. For example, Plonsky (2013) synthesized 606 empirical studies published between 1990 and 2010 in two major journals: *Language Learning* and *Studies in Second Language Acquisition*, coding the studies for designs, statistical analyses, and reporting practices. The results showed a number of strengths and flaws, and the author proposed strategies to resolve the identified issues.

Substantive syntheses seek to aggregate the results of primary studies and reach conclusions about whether an instructional treatment is effective or a certain relationship exists or how frequently a certain phenomenon occurs. Depending on the way the data is analysed, substantive syntheses can be further divided into three types: thematic synthesis, vote-counting, and meta-analysis. In a thematic synthesis, study findings are reported as themes and categories. A thematic synthesis may appear similar to a traditional review, but it is not, because it has all the characteristics of a research synthesis. In particular, it seeks to reach conclusions based on the totality of research rather than critique some selected studies. However, freestanding state-of-the-art reviews, which fall into the traditional review paradigm, have the potential of being converted to thematic syntheses if they follow more rigorous procedures, transparently report the review methods, and narrow their foci. An example thematic synthesis is Dixon et al.'s (2012) synthesis of 72 empirical studies on the optimal conditions of second language acquisition. The synthesis sought to answer five research questions from four perspectives based on transparent study selection criteria. The article reported the information about each included study in a 20-page table, and the study findings were described in a narrative style.

The second type of substantive synthesis is called vote-counting, in which study findings are analysed by tallying the number of significant and non-significant  $p$ -values reported in primary studies. An alternative, which is based on similar principles, is to average the  $p$ -values generated by the primary

studies. An example vote-counting synthesis is Ellis (2002), who aggregated the results of 11 studies examining the effects of form-focused instruction on the acquisition of implicit knowledge. Ellis conducted the synthesis by counting the number of studies that reported significant or nonsignificant effects, followed by a discussion of the results. Strictly speaking, thematic synthesis, which is discussed above, is one type of vote-counting because the conclusions are based on whether primary studies reported significant findings, even though a thematic synthesis does not overtly count the number of significant  $p$ -values.

The third type of substantive synthesis is meta-analysis, where effect sizes extracted from each primary study are aggregated to obtain a mean effect size as a proxy of the population effect (e.g., Li, 2010). Meta-analysis is a common type of research synthesis, and in fact, research synthesis has, either implicitly or explicitly, been equated with meta-analysis by many experts in this field (e.g., Cooper, 2016). Since Glass (1976) coined the term “meta-analysis,” it has become the most preferred method of research synthesis in various fields such as psychology and medicine. Unlike the dearth of guidelines on how to conduct a traditional literature review, there has been an abundance of publications including journal articles, book chapters, and books, which provide systematic instructions on how to carry out a meta-analysis (see Li, Shintani, & Ellis, 2012). Given that it is the most favoured and common method of research synthesis, the focus of this section is on meta-analysis.

## Meta-Analysis

Meta-analysis is a statistical procedure aiming to (1) aggregate the quantitative results of a set of primary studies conducted to answer the same research question(s) and (2) identify factors that moderate the effects across studies. Thus, a meta-analysis seeks to obtain a numeric index of the effects of a certain treatment or the strength of a relationship existing in the population; it also investigates whether the variation of the effects or relationship can be explained by systematic substantive and/or methodological features of the included studies. In a meta-analysis, the “participants” are the included studies, and the unit of analysis is the effect size contributed by each primary study or calculated based on available information. The variables for analysis are those investigated by the primary researchers or created by the meta-analyst on the basis of the features of the studies (e.g., participant demographics, research context, treatment length, etc.). The effect size, which takes different forms depending on the nature of the construct or the study design, is the building

block of a meta-analysis. Importantly, the effect size is a standardized index that makes it possible to compare the results of different studies.

Basing the analyses on effect sizes overcomes the limitations of the dichotomizing  $p$ -value in null hypothesis significance testing (NHST), which represents the presence or absence of a significant effect but tells nothing about the size of the effect. NHST is sensitive to sample size. For example, a significant  $p$ -value may result from a large sample even though the effect is small, and a nonsignificant  $p$ -value may be associated with a large effect but a small sample. Sun, Pan, and Wang (2010) reported that in 11% of the articles published in selected journals in educational psychology, there was a discrepancy between effect sizes and the results of NHST: medium to large effect sizes were associated with nonsignificant  $p$ -values, while small effect sizes were accompanied with significant  $p$ -values. Because the effect size contains information about the magnitude of an effect and it is exempt from the influence of sample size, its utility has gone beyond meta-analysis. For example, many applied linguistics journals have made it a requirement to include effect sizes in manuscripts reporting empirical studies.

Meta-analysis is considered to be superior to other methods, such as vote-counting, and to a large extent the superiority lies in the usefulness of effect size (Li et al., 2012). First, by aggregating effect sizes across primary studies, meta-analysis provides information about the size of the overall effect, whereas vote-counting can only tell us whether an effect is present. Second, in meta-analysis, effect sizes can be weighted in proportion to sample sizes, while in vote-counting all data points carry the same weight. Third, meta-analysis can provide a precise estimate of an effect or relationship, vote-counting methods can only demonstrate what the majority of the studies show about the construct. Fourth, meta-analysis follows rigorous statistical procedures. The results are likely more robust and credible and can guide practice and policy-making.

## How to Do Meta-Analysis?

An easy way to understand meta-analysis is to construe it as an empirical study, which involves problem specification, data collection, data analysis, and research report writing. In the following, we briefly outline five major stages involved in conducting a meta-analysis along with some of the issues and choices encountered at each stage. For a detailed tutorial on conducting a meta-analysis in applied linguistics, see Li et al. (2012), Ortega (2015), and Plonsky and Oswald (2015).

## Stage 1: Identifying a Topic

As with a primary study, a successful meta-analysis starts with clear, well-defined research questions. However, unlike a primary study where the research questions are based on hypotheses and theories or gaps in previous research, the research questions for a meta-analysis are primarily those examined by the primary studies. Nevertheless, the meta-analyst still has to delineate the research domain based on a thorough understanding of the theories and the research that has been carried out. For instance, a meta-analysis on language aptitude (Li, 2015) requires an unequivocal theoretical and operational definition of aptitude. In educational psychology, aptitude refers to any person trait that affects learning outcomes, including cognitive (e.g., analytic ability) as well as affective variables (motivation, anxiety, etc.). However, in most aptitude research in applied linguistics, aptitude has been investigated as a cognitive construct. Therefore, defining the research topic is no easy matter, and the decisions at this preliminary stage have a direct impact on the scope of the synthesis and how it is implemented at later stages.

## Stage 2: Collecting Data

Data collection for a meta-analysis involves searching for and selecting studies for inclusion in the synthesis. Different from a traditional review, which usually does not report how the reviewed studies are identified and sieved, a meta-analysis must report the details on the strategies, databases, and keywords used during the search, as well as the criteria applied to include/exclude studies. One judgement call at this stage concerns whether to include unpublished studies. Evidence shows that studies that report statistically significant results are more likely to be published or submitted for publication (Lipsey & Wilson, 1993). Therefore excluding unpublished studies may lead to biased results, and this phenomenon is called publication bias. One solution, of course, is to include unpublished studies. However, there have been objections to this recommendation on the grounds that unpublished studies are difficult to secure, that they lack internal and external validity, and so on. Another solution is to explore the extent to which publication/availability bias is present in the current dataset, such as by plotting the calculated effect sizes to see whether studies with small effects are missing, followed by a trim-and-fill analysis to see how the mean effect size would change if the missing values were added, or by calculating a fail-safe  $N$  statistic to probe whether the obtained results would be easily nullified with the addition of a small number of studies.

### Stage 3: Coding the Data

Data coding involves reading the selected studies, extracting effect sizes from the studies, and coding independent and moderating variables. As a start, we would like to point out that most introductory texts ignore the reading stage and emphasize the technical aspects of meta-analysis. However, it is important to recognize that meta-analysis is fundamentally a statistical tool that helps us solve problems, not an end in itself. Eventually the contribution of a meta-analysis lies in the findings and insights it generates, not the sophisticated nature of the statistical procedure. Therefore, while every effort should be made to ensure statistical rigour, careful reading of the study reports is critical to a thorough understanding of the substantive domain, meaningful coding of the data, and accurate interpretation of the results. While reading, the meta-analyst should keep notes about each individual study recording the main findings and methodological details, which can be checked at any stage of the meta-analysis.

While the meta-analyst should read the whole article for each study, effect size extraction relates mainly to the results section of the study report. There are three main types of effect sizes:  $d$ ,  $r$ , and OR (odds ratio), which represents mean difference between two groups, correlation, and probability of the occurrence of one event in a certain condition compared with another, respectively. In a meta-analysis, the effect sizes from the primary studies serve as the dependent variable, and the independent variables are of two types: study-generated and synthesis-generated (Cooper, 2016). Study-generated independent variables are those investigated in primary studies, and these variables should be recorded intact. Synthesis-generated independent variables are not directly investigated in primary studies, that is, they are not manipulated as variables by primary researchers. Rather, they are created by the meta-analyst a posteriori based on the characteristics of the primary research. Synthesis-generated variables also include what Lipsey and Wilson (2001) call study descriptors, which refer to the methodological aspects of the primary research such as participants' age, research setting, and so on.

### Stage 4: Analysing the Data

In most meta-analyses, data analysis follows a two-step procedure: aggregation of all effect sizes that produces an estimate of the population effect, followed by moderator analysis exploring whether the variation of effect sizes is

due to systematic differences between subgroups of studies formed by treatment type, research setting, and so on. Effect size aggregation must be theoretically meaningful. For example, in Li's (2015) meta-analysis on language aptitude, there were two types of studies based on different theoretical frameworks and conducted via two distinguishable methodologies. The effect sizes were aggregated separately because squashing the two study types was theoretically unsound, albeit statistically feasible.

In a meta-analysis, several issues need to be attended to for each analysis. One is to assign different weights to the included studies in proportion to their sample sizes such that large-sample studies carry more weight because they provide more accurate estimates of the population effect. The second is to make sure that one study contributes only one effect size for each aggregation, to prevent effect size inflation, Type I errors, and the violation of the "independence of data points" assumption of inferential statistics. In the event that a study contributes several effect sizes, it is advisable to either pick one or average them, depending on which choice suits the research question for that analysis. The third is to calculate a confidence interval for each mean effect size, which is the range the population effect falls into. A confidence interval that does not include zero means the effect size is significant, and a narrow interval represents a robust effect. Fourth, for each mean effect size, it is necessary to report a measure of variability—either standard error or standard deviation—that shows the distribution of the aggregated effect sizes. Finally, a homogeneity test, which is called  $Q_w$  ("w" in the subscript is the abbreviation for "within-group") test, should be performed for each group of effect sizes to assess the distribution of the effect sizes.

Moderator analysis, which is alternatively called subgroup analysis, is often conducted through the  $Q_b$  ( $b$  in the subscript stands for "between-group") test, which is similar to one-way ANOVA, with the only difference being that the  $Q_b$  test incorporates study weights in the calculation of coefficients and standard errors. A significant  $Q$  value indicates significant differences between the subgroups of effect sizes, and post hoc pairwise  $Q_b$  tests are in order to locate the source of significance. In applied linguistics, a common practice is to use the confidence interval as a test of significance: lack of an overlap between two confidence intervals indicates that the mean effect sizes are significantly different. However, it must be pointed out that the opposite is not true, namely, an overlap does not mean absence of significant differences (see Cumming, 2012, for further details). Therefore, the confidence interval is at best a conservative, if not unreliable, test of statistical significance.



## Stage 5: Writing Up the Research Report

Similar to the report for an empirical study, a meta-analytic report includes the following sections: introduction, methods, results, discussion, and conclusion. The introduction should contextualize the meta-analysis by:

1. elaborating the relevant theories
2. defining the research problem and scope
3. identifying the central issues and controversies
4. explaining the rationale for examining the variables

The methods section reports information about search strategies, selection criteria, the coding protocol, and statistical procedures. Transparent reporting is a defining feature that distinguishes meta-analysis from traditional literature reviews; transparent reporting of methods and analytic procedures makes it possible for other researchers to replicate a meta-analysis to verify the findings.

The results section should include a summary of the methodological aspects of the synthesized studies to provide an overall picture about how studies in this domain have been conducted, followed by the results of the meta-analysis. Valentine, Pigott, and Rothstein (2010) recommended creating a table providing a description of the characteristics and results of each study including participant information, the treatment, and the calculated effect sizes. For the meta-analytic results, each mean effect size should be accompanied with the number of effect sizes ( $k$ ), the confidence interval, a measure of dispersion (standard error or standard deviation), the results of a homogeneity test, and the  $p$ -value for the effect size. Results of a moderator analysis should include the number of effect sizes for each group/condition, the  $Q_b$  value, and the related  $p$ -value.

In the discussion section, the meta-analyst may interpret the results “internally” with reference to the methods of the primary studies and “externally” to theories and other reviews such as state-of-the-art traditional reviews and meta-analyses on this and similar topics. The results can also be discussed in terms of how they can be used to guide practice and policy-making as well as future research.

## Traditional Review and Research Synthesis: Comparison and Integration

Given that traditional reviews and research syntheses have been discussed as two separate approaches in the literature, there seems to be a need to make a direct comparison between them. The purpose of the comparison is not only



to distinguish them but also to stimulate thoughts on how to accurately understand what the two approaches can achieve and how to overcome their limitations and maximize their strengths. Before comparing the two methods of review, it is helpful to provide a taxonomy of literature review as a genre, which, according to Cooper (2016), can be classified on six dimensions: focus, goal, perspective, coverage, organization, and audience. The *focus* of a review refers to the type of information to be included, which may take the form of theories, empirical findings, research methods, and policies/practices. The *goals* to accomplish may include summarizing existing knowledge, evaluating the validity of certain aspects of the primary research, or identifying issues central to the field. *Perspective* refers to whether a reviewer adopts a neutral position or is predisposed to a certain standpoint. *Coverage* concerns whether a review is based on all available studies or a selected set of studies. In terms of *organization*, a review can be structured based on the historical development of the research, the themes that emerged in the literature, or the methods utilized by subgroups of studies. The *audience* of a review may vary between researchers, practitioners, general public, and so on. It is noteworthy that the options for each dimension can be applied jointly when identifying the characteristics of a review or making decisions on what type of review one plans to carry out. For example, the focus of a review can be on both theory and research findings. Similarly, a review can be structured historically, but the research within a certain period can be synthesized thematically.

While Cooper's scheme does not directly distinguish traditional reviews and research syntheses, it helps us understand the differences between the two approaches, which are distinguishable along the following lines (see Table 6.1). First, traditional reviews are based on selected studies while research syntheses are based on all available studies. Certainly, research syntheses may also be selective, following certain screening criteria, but in general they tend to be more inclusive than traditional reviews. Second, the studies included in a traditional review are vetted via the reviewer's expertise or authority; in a research synthesis, however, the primary studies are often assessed based on criteria for study quality. Third, traditional reviews do not follow protocols or procedures, whereas research synthesis follows rigorous methodology. Fourth, one important feature of research synthesis is its transparency in reporting the review procedure or process, which is absent in traditional reviews. A corollary is that a research synthesis is replicable, but a traditional review is not. Fifth, traditional reviews do not often have research questions and often seek to provide overarching descriptions of previous research; research syntheses, in contrast, aim to answer clearly defined research questions. The above arguments and observations seem to suggest that traditional reviews are subjective

**Table 6.1** A comparison between traditional reviews and research syntheses

	Traditional review	Research synthesis
Focus	Theory, research, and practice	Research
Purpose	To justify further research; identify central issues and research gaps; discuss state of the art; critique previous research	To answer one or more research questions by collating evidence from previous research; resolve controversies; guide practice
Coverage	Selective: most relevant and representative; no selection criteria	Inclusive: all relevant studies; based on systematic search and justified selection criteria
Methodology	No prescribed methodology	With transparent methodology
Structure	Organized by themes and patterns	Analogous to the template of a research report, including an introduction, methods, results, discussion, and conclusion
Style	Narrative, critical, and interpretive	Descriptive, inductive, and quantitative
Pros	Flexible; appropriate when the purpose is to critique rather than aggregate research findings	Objective, transparent, and replicable; results may inform practice and policy-making
Cons	Subjective; based on idiosyncratic methods; not replicable	Subject to the quality of available studies; comparing oranges and apples; time-consuming

and may lead to false claims and that research syntheses are more objective and generate robust results that can guide practice and policy-making.

However, it is arbitrary and inaccurate to discount traditional reviews as valueless, and their utility is justifiable on the following grounds. First, a traditional review is often part of a journal article, a master's thesis, or a Ph.D. dissertation, and the primary purpose is to contextualize a new study and draw on existing research methods to answer the research questions of the current study. In this context, it seems less important to include all available studies and reach conclusions based on the totality of the research. Second, traditional reviews are flexible and can be used to synthesize knowledge other than research findings such as theories, practices, and policies. Third, traditional reviews are often critical. Thus, if the purpose of a review is to critique certain aspects of previous research such as the instruments used to measure a certain construct rather than collate information to show the effectiveness of a certain treatment or the presence of a certain relationship, then the traditional approach is more appropriate. Finally, traditional reviews are appropriate when (1) it is premature to conduct a research synthesis because of the lack of research, and (2) aggregation of evidence is not meaningful because the primary studies are carried out using heterogeneous methods.

So what do we make of the status quo of literature review as a field of research in relation to the two different review methods? First of all, the differences between the two methods are suggestive rather than conclusive, and they stand in a continuum rather than a dichotomy. For example, although traditional reviews are more likely to be critical, those adopting a more systematic approach may also critique certain aspects of the research based on their aggregated results. Therefore, it is better to consider the differences in terms of overall emphasis and orientation rather than treat them as polarized disparities. Second, the differences are based on what has been observed about two broad types of review. They are *a posteriori* and descriptive, not stipulated or prescriptive; in other words, they do not have to differ the way they are assumed to be different. For example, traditional reviews have been criticized for being subjective, but there is no reason why they cannot become more objective by following more rigorous procedures, such as conducting more thorough literature searches, including more studies that represent different perspectives, and so on.

Finally, and importantly, we should find ways to integrate traditional literature reviews and more systematic approaches such as meta-analysis, and such initiatives have already taken place in our field. For example, Carpenter (2008) included a meta-analysis of the predicative validity of the Modern Language Aptitude Test in her Ph.D. dissertation when reviewing the literature on language aptitude. In this case, a small-scale meta-analysis is embedded in a traditional literature to explore an important issue, which constitutes an assimilative approach where a meta-analysis plays a supplementary role. Another way is to adopt a more balanced approach where the two review methods are utilized to synthesize (1) different types of knowledge or (2) studies conducted with different methods. In the case of (1), a good example is Xu, Maeda, Lv, and Jinter (2015), who used more traditional methods to synthesize the theoretical and methodological aspects and meta-analysis to aggregate the findings of the primary research. In the case of (2), an example is Li (2017), who conducted a comprehensive review of the research on teachers' and learners' beliefs on corrective feedback. The author meta-analysed the results of the studies conducted using similar methods (e.g., studies using a five-point Likert scale) but described the themes and patterns demonstrated by studies that deviated from the majority, such as those using a four- or three-point scale and those that used qualitative methods (e.g., observations, interviews, or diaries). Therefore, rather than make an *a priori* decision to carry out a certain type of review, we may customize our methodology based on the available data, using or integrating different approaches so as to provide a comprehensive, impartial view of the research domain.

## Resources for Further Reading

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Sons.

This book constitutes a comprehensive, practical guide for both novices and experts on how to carry out a meta-analysis. The book also discusses different practices and judgement calls one may face at each step, draws the reader's attention to pitfalls, and makes recommendations on how to resolve issues and overcome challenges.

Galvan, J. (2014). *Writing literature reviews: A guide for students of the social and behavioral sciences*. Glendale, CA: Pycszak Publishing.

Galvan's book is an extremely useful source of information for students who have no experience in conducting a literature review and for teachers who want to teach students how to do a literature review. Like this chapter, Galvan dissects literature review into a multi-step process rather than treats it as a product that only concerns the write-up or the discourse features of a review. The book provides straightforward, accessible guidelines as well as concrete, real-world examples illustrating how to apply the guidelines.

## References

- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.
- Booth, W., Colomb, G., & Williams, J. (1995). *The craft of research*. Chicago: The University of Chicago Press.
- Bronson, D., & Davis, T. (2012). *Finding and evaluating evidence: Systematic reviews and evidence-based practice*. New York: Oxford University Press.
- Carpenter, H. S. (2008). *A behavioural and electrophysiological investigation of different aptitudes for L2 grammar in learners equated for proficiency level*. Ph.D. dissertation, Georgetown University.
- Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Thousand Oaks, CA: SAGE.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and metaanalysis*. New York: Routledge.
- Dixon, L., Zhao, J., Shin, J., Wu, S., Su, J., Burgess-Brigham, R., & Snow, C. (2012). What we know about second language acquisition: A synthesis from four perspectives. *Review of Educational Research*, 82(1), 5–60.

- Ellis, R. (2002). Does form-focused instruction affect the acquisition of implicit knowledge? A review of the research. *Studies in Second Language Acquisition*, 24(2), 223–236.
- Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Education Researcher*, 5, 3–8.
- Imel, S. (2011). Writing a literature review. In T. Rocco & T. Hatcher (Eds.), *The handbook of scholarly writing and publishing* (pp. 145–160). San Francisco, CA: Jossey-Bass.
- Jesson, J., Matheson, L., & Lacey, F. (2011). *Doing your literature review: Traditional and systematic approaches*. Los Angeles, CA: SAGE.
- Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, 60(2), 309–365.
- Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, 36(3), 385–408.
- Li, S. (2017). Teacher and learner beliefs about corrective feedback. In H. Nassaji & E. Kartchava (Eds.), *Corrective feedback in second language teaching and learning* (pp. 143–157). New York, NY: Routledge.
- Li, S., Shintani, N., & Ellis, R. (2012). Doing meta-analysis in SLA: Practices, choices, and standards. *Contemporary Foreign Language Studies*, 384(12), 1–17.
- Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and Behavioural treatment: Confirmation from meta-analysis. *American Psychologist*, 48(12), 1181–1209.
- Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Machi, L., & McEvoy, B. (2012). *The literature review: Six steps to success*. Thousand Oaks, CA: Corwin.
- Ortega, L. (2015). Research synthesis. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 225–244). London: Bloomsbury.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35, 655–687.
- Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, 61(2), 325–366.
- Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). New York, NY: Routledge.
- Sun, S., Pan, W., & Wang, L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, 102(4), 989–1004.
- Valentine, J., Pigott, T., & Rothstein, H. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, 3(2), 215–247.

- Xu, Y., Maeda, Y., Lv, J., & Jinther, A. (2015). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, 33(4), 497–528.
- Zaporozhetz, L. (1987). *The dissertation literature review: How faculty advisors prepare their doctoral candidates*. Ph.D. dissertation, The University of Oregon.



# 7

## Research Replication

Rebekha Abbuhl

### Introduction

This chapter traces the history of replication research in the field of applied linguistics, culminating in a discussion of current views of replication research as a means of evaluating the internal and external validity of a study, providing better explanations of phenomena of interest, and ultimately, driving both theory and pedagogy forward. The chapter discusses different types of replication studies (e.g., exact, approximate, conceptual) with recent examples from the field. Controversies surrounding the use of this type of research—including whether replications of qualitative studies are possible or desirable—are discussed. The author concludes with current recommendations to facilitate replication research in applied linguistics.

### History

Replication is so engrained in the physical and natural sciences that it rarely provokes discussion: it is taken as self-evident that repeating experiments is needed to confirm and test the generalizability of results (Hendrik, 1990).

---

R. Abbuhl (✉)

Department of Linguistics, California State University at Long Beach,  
Long Beach, CA, USA

e-mail: [Rebekha.Abbuhl@csulb.edu](mailto:Rebekha.Abbuhl@csulb.edu)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_7](https://doi.org/10.1057/978-1-137-59900-1_7)

However, the treatment of replications in applied linguistics (and in the social sciences in general) has been more contentious. As a relatively young field, applied linguistics has prioritized innovation and originality for most of its short history. This emphasis has spurred tremendous growth and an ever-expanding research agenda—but at the same time, it is at least partially responsible for the widespread view that replications are the poor cousins of original research (Porte, 2012). And despite early attempts to argue to the contrary (e.g., Polio & Gass, 1997; Santos, 1989; Valdman, 1993), replications have remained scarce in applied linguistics. Polio (2012b), for example, examined studies conducted between 1990 and 2009 in six major journals in the field and identified only 24 replication studies explicitly labeled as such. Similarly, Marsden and Mackey (2013) [as reported in Mackey & Marsden, 2016] found that in the database *Language and Linguistic Behavior Abstracts*, there were only 40 such studies dated between 1973 and 2013.

A series of developments has helped to shift that view in applied linguistics. In 2007, *Language Teaching* became the first journal in the field to have a strand specifically devoted to replications; a subsequent symposium organized by the editor of that journal, Graeme Porte, was held in 2009, covering various issues related to using, interpreting, and teaching replications. In 2012, the first book-length treatment of replications in the field was published (Porte, 2012), and since then renewed attention has been given to this topic (e.g., Basturkmen, 2014; Bikowski & Schulze, 2015; Bitchener & Knoch, 2015; Casanave, 2012; Chun, 2012; Gass & Valmori, 2015; King & Mackey, 2016; Markee, 2017; Matsuda, 2012; Mu & Matsuda, 2016; Plonsky, 2012, 2015; Polio, 2012a, 2012b; Porte, 2012, 2013; Porte & Richards, 2012; Sasaki, 2012; Schmitt, Cobb, Horst, & Schmitt, 2017; Smith & Schulze, 2013; Webb, 2015; Willans & Leung, 2016).

## Current Issues

There are many factors behind the recent surge of interest in replications. One, of course, is the increasing maturity of applied linguistics as a field: Discovery, for a young area of inquiry, is paramount, but for more mature fields, confirmation of results is as equally important (Sakaluk, 2016). The surge of interest may be related to the field's growing sophistication with respect to its use and interpretation of statistics. Traditionally, a statistically significant finding was considered tantamount to a generalizable finding.



However, as a number of researchers have pointed out, this is not the case; external replications are needed to provide information about the generalizability of any given result (e.g., Nassaji, 2012).

Another reason for the increasing attention to replication is the current replicability crisis facing the sciences in general and social sciences, such as psychology. Controversial claims have been made that most published research results are false (e.g., Ioannidis, 2005; see also Pashler & Harris, 2012). In addition, there have been highly publicized cases of fraud (see e.g., King & Mackey, 2016; Pashler & Wagenmakers, 2012, for discussions) and questionable research practices (see e.g., John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011, for overviews). These questionable research practices include not mentioning all dependent measures, selectively reporting statistically significant results (and omitting those that are not), misrepresenting  $p$ -values (e.g., reporting that a  $p$ -value greater than a particular criterion, say 0.05, was actually less than that), claiming that an unexpected finding had actually been predicted from the beginning, stopping data collection when the hoped-for result had been obtained, excluding data in order to obtain the hoped-for result, and falsifying data (John et al., 2012). In John et al.'s (2012) survey of over 2000 psychologists, the majority admitted to engaging in one or more of these practices, leading John et al. to conclude that "these practices may constitute the de facto scientific norm" (p. 524).

It has long been noted that the social sciences have a "file drawer" problem (Rosenthal, 1979, p. 638), in that statistically significant results are far more likely to be published than those that are not. This may help explain the prevalence of these questionable research practices: if publication is more likely when the results are statistically significant, and if career considerations (e.g., retention, tenure, promotion) hinge on publications, then it is little wonder why such practices continue.

As these practices inflate Type I errors (i.e., false positives, or erroneously rejecting the null hypothesis), researchers in many fields have called upon the use of replications as a safeguard mechanism (e.g., Burman, Reed, & Alm, 2010, for public finance; Benson & Borrego, 2015, for engineering education; McNeeley & Warner, 2015, for criminology; Mezas & Regnier, 2007, for management, among many others). Efforts across a range of fields have also been made to encourage more replications, as we will discuss below. Nevertheless, controversy has continued, in part because of the misunderstandings surrounding the definition and interpretation of replications. We turn to the issue of defining in the next section.

## Overview of Design

A tripartite distinction is typically made between exact, approximate, and conceptual replications (Abbuhl, 2012a, 2012b; Language Teaching Review Panel, 2008; Porte, 2012; Porte & Richards, 2012; see, however, Polio, 2012b, for a slightly different classification). It should be understood from the outset, however, that these three types exist along a continuum, from the most faithful to the original study (exact) to the least (conceptual). Each of these will be discussed in turn below.

The least common (but most well known) type of replication in applied linguistics is the *exact replication* (also known as direct or literal). This involves identifying a methodologically sound study and repeating the study as exactly as possible (i.e., the same tasks, setting, and type of participant) in order to confirm the original findings. It is recognized that some small changes will be inevitable (e.g., the gap in time) even though every effort is made to be as faithful to the original study as possible (Earp & Trafimow, 2015).

In the field of applied linguistics and in the social sciences in general, *approximate replications* are much more common than exact replications (for recent examples, see Booth, 2013; Johnson & Nicodemus, 2015). In this type of replication, the original study's methodology is adhered to in most respects, but one or two of the non-major variables—such as the context, participant first language (L1), or task—are changed in order to determine the generalizability of the original results. In other words, there is a deliberate change in order to determine whether the results of the original study hold for a new population, setting, language, and so on. Sample Study 7.1 provides an example of an approximate replication.

### Sample Study 7.1

Johnson, M., & Nicodemus, C. (2015). Testing a threshold: An approximate replication of Johnson, Mercado & Acevedo 2012. *Language Teaching*, 49(2), 251–274.

#### Research Background

One question in the second language (L2) writing literature concerns the effect of pre-task planning on written production. The original study (Johnson, Mercado, & Acevedo, 2012) sought to examine the effect of pre-task planning on the fluency and complexity of L2 writing. Johnson et al. (2012) found that although pre-task planning had a small effect on writing fluency, it did not significantly affect grammatical or lexical complexity.

#### Research Problems/Gaps

As their results differed from those found in previous studies, Johnson et al. (2012) suggested that the characteristics of their participants (e.g., their rela-

tively low proficiency in the L2) may have been partially responsible. In particular, they hypothesized that writers who have not reached a certain level of L2 proficiency may not benefit from pre-task planning as their cognitive resources may already be stretched by composing in the L2.

### Research Method

- To test this hypothesis, Johnson and Nicodemus (2015) conducted an approximate replication of the original study, changing the participants to a group of L1 writers.
- The researchers divided the 90 L1 writers into a control group with no pre-task planning and three groups with different types of pre-task planning.
- The writing produced by the students was analyzed in the same manner as the original study, with the fluency, grammatical complexity, and lexical complexity calculated.

### Key Results

Johnson and Nicodemus (2015) found that pre-task planning had no effect on the written fluency, grammatical complexity, or lexical complexity of their L1 writers. This result led the researchers to conclude that Johnson et al.'s (2012) hypothesis (that writers need to have a certain level of proficiency in the language in order to benefit from pre-task planning) could not be not supported.

### Comments

This study provides evidence that approximate replications have a role to play beyond determining the generalizability of an original study. Had the results of Johnson and Nicodemus' (2015) replication differed markedly from those in the original (viz., if there had been a large effect for pre-task planning for the L1 writers), this would have supported the original study's explanation for their results (viz., that their L2 participants had not reached a threshold level of proficiency to benefit from the planning). The consistency of results between the replication and original allowed Johnson and Nicodemus (2015) to refute the original study's explanation of their results.

At the least faithful end of the continuum is the *conceptual replication*. Here, a new research design (which may involve changing the independent and/or dependent variables, operationalizations of phenomena, and participant population) is used to investigate the same general idea or hypothesis as the original study. The purpose of such a replication may be to test the generalizability of relationships to new sets of variables within a larger model, or alternatively, to determine to what extent the findings of the original study were artifacts of its own methodology. Recent examples of this type of replication include Frankenberg-Garcia (2014), Leow (2015), Lim and Godfroid (2015), Rott and Gavin (2015), and Yoon and Polio (2017). An example of a conceptual replication can be found in Sample Study 7.2.

## Sample Study 7.2

Yoon, H.-J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, 51(2), 275–301.

### Research Background

The original study, Lu (2011), used 14 different measures of syntactic complexity to examine a corpus of writing produced by English as a second language (ESL) writers at four different levels of proficiency. The corpus contained both timed and untimed essays, 16 different topics, and different genres (argumentative and narrative). One of the main findings was that there was greater syntactic complexity in the argumentative essays than in the narrative essays.

### Research Problems/Gaps

Yoon and Polio (2017) sought to determine whether differences observed in Lu (2011) were most likely due to developmental reasons or to genre factors.

### Research Method

- In a conceptual replication of Lu (2011), Yoon and Polio (2017) examined a range of writing development measures, including syntactic complexity, syntactic accuracy, lexical complexity, and fluency on two different topics.
- They recruited ESL students ( $N = 37$ , at the top two levels of a university program) and a group of native speakers ( $N = 46$ ) to see which differences could be attributed to genre and which to proficiency.
- A longitudinal component was added to examine changes over time.

### Key Results

Yoon and Polio (2017) found interactions between genre and time in the L2 essays. For example, in the narrative essays, there was a significant increase in two of the syntactic complexity measures; however, no such increase was seen in the argumentative essays. Comparing the genres, the researchers found that several syntactic complexity measures were greater in the argumentative writing than in the narrative writing. Differences between the genres were also seen in the lexical complexity and lexical diversity measures (but not in fluency or in accuracy). Similar patterns were also seen in the native speaker writings.

### Comments

Yoon and Polio's supportive conceptual replication provided evidence that Lu's (2011) results were not simply artifacts of the writing development measures used or the proficiency level of the participants. Given the scarcity of research on genre effects in the field of L2 writing, Yoon and Polio's results highlight the importance of addressing this variable when investigating writing development.

## Challenges

This section presents several challenges in replication research, such as interpretations of research findings, qualitative research replications, bias issues, and reporting and data sharing issues.

### Interpretation of Results

One of the most common concerns regarding replication research centers on the interpretation of results. Replications are typically classified as “successes” or “failures” depending on whether or not they report the same findings as the original study. In the case of a “successful” exact replication, the findings of the original study are confirmed and researchers can have more confidence in the internal and external validity of the original study (e.g., Language Teaching Review Panel, 2008). However, it has also been maintained that such findings lack novelty and that the researcher is merely (and unimaginatively) verifying the results of the original study (Hendrik, 1990). As Bronstein (1990) puts it, in such a case “you’ve just demonstrated something that we already knew” (p. 73). If the exact replication “fails,” on the other hand, researchers may be tempted to conclude that the findings of the original study are false and need to be overturned. Recent researchers have pointed out that this is not the case (replication “failures” may be due to many reasons, including undetected moderators [i.e., subtle and currently unknown factors that affect the results], human error, low statistical power in the replication, and/or other factors, e.g., Makel & Plucker, 2015; Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2012), but this uncertainty contributes to the view that unsuccessful replications are “uninterpretable” (Bronstein, 1990, p. 74).

However, it needs to be kept in mind that just as no single original study is proof for a particular phenomenon (just evidence for), no single replication will either prove or falsify the original study. Replications should be seen as informative rather than as conclusive (Earp & Trafimow, 2015). For example, if an exact replication obtains the same findings as the original study, it should be seen as additional evidence that the original study is valid. If it does not reach the same findings, it should be seen as evidence that more research is necessary (Crandall & Sherman, 2016), not that it is a failure. In this light, the terms success and failure should be replaced with supportive and non-supportive.

Moving on to the next type of replication, a supportive approximate replication should be seen as providing evidence that the results of the original study can be generalized, for example, to a new language, population, or context. A non-supportive approximate replication may suggest that the original results cannot be generalized, but more importantly, it should serve as a call for further research on the topic. Similarly, a supportive conceptual replication can provide evidence that a particular finding is not an artifact of the original study's methodology; a non-supportive conceptual replication should again serve as a springboard for more research. Were the differences in methodologies responsible for the non-supportive results? Were there undetected moderators, and if so, what are they? Addressing these questions, as Crandall and Sherman (2016) note, can provide opportunities for theoretical advances.

## Qualitative Research Replications

Another controversy surrounding replications is whether they are possible or even desirable for qualitative studies (e.g., Casanave, 2012; Matsuda, 2012; Sasaki, 2012). Traditionally, the argument against qualitative replications was that the verification of results in this type of inquiry was a reductionist and fruitless endeavor: the particularities of the researcher, the participant(s), and the setting would make it highly unlikely that another researcher could study the same participant (or even the same type of participant) in the same setting using the same methodological tools and come up with the same results. In addition, as Schofield (2002) notes, “[t]he goal is *not* to produce a standardized set of results that any other careful researcher in the same situation or studying the same issue would have produced. Rather it is to produce a coherent and illuminating description of and perspective on a situation that is based on and consistent with detailed study of that situation” (p. 174, italics in original).

However, much of this concern over the verification of qualitative findings (and hence the antipathy to replications among some qualitative researchers) seems to stem from conflating replication with exact replication (Porte & Richards, 2012). Due to the interpretive and situated nature of qualitative inquiry, researchers' concerns over exact replications are not without merit. This does not mean, however, that all qualitative replications should be abandoned or even that qualitative exact replications are impossible.

For example, let's say that an instrumental case study was conducted on a small group of English as a second language (ESL) writers to determine their perceptions of written corrective feedback as they progressed through univer-

sity. In contrast to an intrinsic case study (where the primary focus is to explore and understand the case at hand, without an attempt to generalize across cases), in an instrumental case study, understanding the phenomena being studied is more important than the cases themselves (Stake, 2005). Let's further say that the researcher made full use of sound qualitative research practices (e.g., rich description, triangulation, audit trails, coding checks, member checking, long-term observation). With such a study, readers can make determinations as to whether any of the information provided applies to their own local context.

However, the study may also raise some interesting questions: for example, if another researcher used the same sound methodology that the original researcher employed, what perceptions would she or he uncover in a slightly different population, say English as a foreign language writers? Would the context impact the students' perceptions? A supportive approximate replication here would suggest that context plays a minimal role in student attitudes toward written corrective feedback; a non-supportive approximate replication, rather than being a failure, would uncover both theoretically and pedagogically interesting information—in particular, that context does play a role. In a similar vein, a conceptual replication of our ESL writer study could involve changing aspects of the methodology—for example, replacing researcher-guided interviews with peer-guided ones, or supplementing the interviews with diary data. A supportive conceptual replication here would suggest that the methodology does not exert undue influence over the students' responses. A non-supportive conceptual replication, rather than undermining the original study or proving it false, would suggest that methodological changes can influence student responses.

But what about exact or nearly exact replications in qualitative research? If a qualitative researcher were to take this approach, she or he could investigate the same type of participant, in the same setting, using the same methodology (all the while recognizing that the particularities of the researcher and participant could affect the results). A supportive exact replication here would provide evidence that these particularities are not major factors when studying the phenomenon at hand; a non-supportive exact replication would provide the equally valuable evidence that those factors do likely play an important role. Future studies could then work to determine how much of a role they play and under what conditions they exert an influence.

When we abandon using the terms success and failure to describe replications, and when we recognize that replicating some types of qualitative studies (e.g., instrumental case studies) could offer valuable information to the field, some of the current debate over qualitative replications may be dispelled. As

Porte and Richards (2012) note, this kind of reconceptualization could help reconfigure “transferability issues in [qualitative research], moving away from questions about whether findings might be transferable and toward the issues of transferability itself: In what respects cases might be transferable, for what reasons they might be transferable, etc.” (p. 289). It is an approach that a number of qualitative researchers have advocated (e.g., Golden, 1995; Markee, 2017; Schofield, 2002), and one that may open up additional areas of inquiry.

## Bias Issues

Another challenge concerning replication studies is the bias issue. As Makel and Plucker (2015) note, “there is ... a catch-22: Replication is good for the field but not necessarily for the individual researcher” (p. 158). Researchers have long noted that replications are a low-prestige endeavor, commonly associated with a lack of creativity and a confrontational personality (e.g., Bronstein, 1990; Earp & Trafimow, 2015; Easley, Madden, & Dunn, 2000; Hubbard & Armstrong, 1994; Makel & Plucker, 2015; Makel, Plucker, & Hegarty, 2012; Mezas & Regnier, 2007; Neuliep & Crandall, 1990). Real or perceived editorial biases, as well as concerns over external funding and career advancement, may further dissuade researchers from conducting replications (Bronstein, 1990; Earp & Trafimow, 2015; Neuliep & Crandall, 1990). Recent events in psychology (where the replication author was attacked via social media and threatened by the original author) also lend support to the view that replications continue to be a risky endeavor (see LeBel, 2015, for a discussion). There are signs that the situation may be improving (Mu and Matsuda’s [2016] survey of second language writing researchers did not uncover biases toward replications), but it would be premature at this point to declare that original and replication studies are on equal footing.

## Reporting and Data Sharing Issues

Another issue that impacts the field’s ability to conduct and interpret replications concerns reporting practices. Over the years, numerous calls have been made for authors to report full methodological details in their studies so as to facilitate replication (e.g., Polio & Gass, 1997). Researchers interested in conducting exact replications, of course, need these details to make the replication as faithful to the original as possible, but even researchers doing approximate or conceptual replications require this information in order to interpret any differences in findings between their own studies and



the original. In the past, full methodological information was often excluded due to page restrictions imposed by journals, but with the advent of online supplementary materials and databases, such restrictions are gradually proving to be less of a barrier.

Recent researchers have also advised authors to be more transparent with respect to data analysis. Quantitative researchers have been urged to be more consistent in their reporting of descriptive statistics (means, standard deviations), confidence intervals, effect size, and statistical power (e.g., Larson-Hall & Plonsky, 2015; Liu & Brown, 2015; Plonsky, 2013); qualitative researchers have been encouraged to provide more information on their coding processes and analytical decisions (e.g., Moravcsik, 2014; Richards, 2009). Such practices will not only help in the interpretation of results but also in the comparison of studies, which is central to both replications and meta-analyses.

Related to these reporting concerns is the issue of data sharing. Previous studies have provided evidence that authors may be reluctant to share their raw data (or, in cases of storage failure, unable to) (e.g., Plonsky, Egbert, & LaFlair, 2015). Sharing data (when not prohibited by privacy concerns) has been encouraged by a number of bodies, including the American Psychological Association and the Linguistics Society of America.

## Future Directions

As noted by Reis and Lee (2016), “pointing to the importance of replication is a little bit like arguing that we should help the elderly and babies: everyone agrees in principle but how to make it happen can often become vexing” (p. 2). Thankfully, in recent years great strides have been made in applied linguistics and in other fields to facilitate replication research.

With respect to making full methodological information available, a number of recent developments have occurred, from journals explicitly stating that “Methods sections must be detailed enough to allow the replication of research” (from the journal *Language Learning*) to the use of supplementary online materials. The latter can take the form of additional files posted to the publishing journal’s website (as in the case of *Language Learning*, for example) to large-scale repositories containing materials used in published research (e.g., the *Instruments for Research into Second Languages [IRIS]*, Marsden, Mackey, & Plonsky, 2016).

A number of journals (e.g., *Language Teaching* and *Studies in Second Language Acquisition*) have also actively called for replication studies. *Language Teaching* has taken the additional step of publishing articles that

provide recommendations as to which articles in the field should be replicated, along with recommendations as to how to carry out those replications (e.g., Bitchener & Knoch, 2015; Gass & Valmori, 2015; Schmitt et al., 2017).

There have been interesting developments outside the field of applied linguistics as well. For example, in psychology, researchers have suggested that a “replication norm” be established, encouraging researchers “to independently replicate important findings in their own research areas in proportion to the number of original studies they themselves publish per year” (LeBel, 2015, p. 1). Psychologists have also created websites that allow researchers to upload and view unpublished replications in their field (e.g., [psychfiledrawer.org](http://psychfiledrawer.org)). Actively encouraging replications—whether through a replication norm or otherwise—as well as developing methods of making such studies publically available, will help remove some of the real and perceived barriers to replication research.

Another suggestion from the field of psychology is to “explore small and confirm big” (Sakaluk, 2016, p. 47). According to Sakaluk, in the initial stages of exploration, small-scale studies would help develop hypotheses about particular phenomena. In the confirmatory stages, researchers would then seek to replicate the findings from the first stage using large samples (thus, if a non-significant finding was found, researchers could rule out low statistical power as being responsible). If individual researchers do not have access to sufficient participants for the confirming stage, a many labs approach (e.g., Klein et al., 2014; Open Science Collaboration, 2012; Schweinsberg et al., 2016) could be taken. Here, multiple independent labs pre-register their hypotheses, materials, and data analysis plans (so as to minimize the questionable research practices discussed above). The study would then be carried out, and the results of the separate replications aggregated in a meta-analysis. The beginnings of such a project in applied linguistics are discussed in Marsden et al. (2016). Pre-registered protocols could also be used to minimize the file drawer problem. For example, LeBel (2015) noted that in several journals in psychology, the results of pre-registered studies are published regardless of the results.

Replications can help researchers verify study results, examine questions of generalizability, and in the process, better understand phenomena of interest. Perhaps most importantly, though, accepting replications—and in particular, multiple replications—as a normal part of the research process can help bring home the point that no single study has the ability to conclusively separate the spurious from the real.

## Resources for Further Reading

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(S1), 127–159.

The authors provide recommendations concerning data reporting that will facilitate study interpretation, meta-analyses, and replications.

Mackey, A., & Marsden, E. (2016). *Advancing methodology and practice: The IRIS repository of instruments for research into second languages*. New York: Routledge.

Recognizing the importance of data sharing and transparency in the field, Mackey and Marsden present a collection of articles based on the Instruments for Research into Second Languages (IRIS), a publically available repository for materials used in second language research.

Markee, N. (2017). Are replication studies possible in qualitative second/foreign language classroom research? A call for comparative re-production research. *Language Teaching*, 50(3), 367–383.

In this article, Markee presents a strategy for replicating qualitative studies in second language classroom research (the comparative re-production method, a common practice in ethnomethodological conversation analysis).

Porte, G. (Ed.). (2012). *Replication research in applied linguistics*. Cambridge: Cambridge University Press.

This edited collection of articles is the first (and to date, only) book-length treatment of replications in the field of applied linguistics. The collection covers a variety of topics, including the history of replications in applied linguistics, the relationship between statistical interpretations and replications, teaching replications at the graduate level, and writing them up. Two examples of replication studies are also provided.

## References

- Abbuhl, R. (2012a). Practical methods for teaching replication to applied linguistics studies. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 135–150). Cambridge: Cambridge University Press.
- Abbuhl, R. (2012b). When, when, and how to replicate research. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition* (pp. 296–312). Oxford: Wiley-Blackwell.
- Basturkmen, H. (2014). Replication research in comparative genre analysis in English for academic purposes. *Language Teaching*, 47(3), 377–386.
- Benson, L., & Borrego, M. (2015). The role of replication in engineering education research. *Journal of Engineering Education*, 104(4), 388–392.
- Bikowski, D., & Schulz, M. (2015). Replication and evaluation in CALL. *CALICO Journal*, 32(2), i–v.
- Bitchener, J., & Knoch, U. (2015). Written corrective feedback studies: Approximate replication of Bitchener & Knoch (2010a) and Van Beuningen, De Jong & Kuiken (2012). *Language Teaching*, 48(3), 405–414.
- Booth, P. (2013). Vocabulary knowledge in relation to memory and analysis: An approximate replication of Milton's (2007) study on lexical profiles and learning style. *Language Teaching*, 46(3), 335–354.
- Bronstein, R. (1990). Publication politics, experimenter bias and the replication process in social science research. *Journal of Social Behavior and Personality*, 5(4), 71–81.
- Burman, L., Reed, W., & Alm, J. (2010). A call for replication studies. *Public Finance Review*, 38(6), 787–793.
- Casanave, C. (2012). Heading in the wrong direction? A response to Porte and Richards. *Journal of Second Language Writing*, 21(3), 296–297.
- Chun, D. (2012). Replication studies in CALL research. *CALICO Journal*, 29(4), 591–600.
- Crandall, C., & Sherman, J. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66, 93–99.
- Earp, B., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6(621), 1–11.
- Easley, R., Madden, C., & Dunn, M. (2000). Conducting market science: The role of replication in the research process. *Journal of Business Research*, 48(1), 83–92.
- Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL*, 26(2), 128–146.
- Gass, S., & Valmori, L. (2015). Replication in interaction and working memory research: Révész (2012) and Goo (2012). *Language Teaching*, 48(4), 545–555.
- Golden, M. (1995). Replication and non-quantitative research. *PS: Political Science and Politics*, 28(3), 481–483.

- Hendrik, C. (1990). Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior and Personality*, 5(4), 41–49.
- Hubbard, R., & Armstrong, J. (1994). Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, 11(3), 233–248.
- Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532.
- Johnson, M., Mercado, L., & Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing*, 21(3), 264–282.
- Johnson, M., & Nicodemus, C. (2015). Testing a threshold: An approximate replication of Johnson, Mercado & Acevedo 2012. *Language Teaching*, 49(2), 251–274.
- King, K., & Mackey, A. (2016). Research methodology in second language studies: Trends, concerns, and new directions. *Modern Language Journal*, 100(s1), 209–227.
- Klein, R. A., Ratliff, R. A., Vianello, M., Adams, R. A., Jr., Bahník, S., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A “ManyLabs” replication project. *Social Psychology*, 45(3), 142–152.
- Language Teaching Review Panel. (2008). Replication studies in language learning and teaching: Questions and answers. *Language Teaching*, 41(1), 1–14.
- Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(S1), 127–159.
- LeBel, E. (2015). A new replication norm for psychology. *Collabra*, 1(1), 1–13.
- Leow, R. (2015). The roles of attention and (un)awareness in SLA: Conceptual replication of N.C. Ellis & Sagarra (2010a) and Leung & Williams (2012). *Language Teaching*, 48(1), 117–129.
- Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen’s 2009 study. *Applied Psycholinguistics*, 36(5), 1247–1282.
- Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing*, 30, 66–81.
- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers’ language development. *TESOL Quarterly*, 45(1), 36–62.
- Mackey, A., & Marsden, E. (2016). *Advancing methodology and practice: The IRIS repository of instruments for research into second languages*. New York: Routledge.

- Makel, M., & Plucker, J. (2015). An introduction to replication research in gifted education: Shiny and new is not the same as useful. *Gifted Child Quarterly*, 59(3), 157–164.
- Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, 7(6), 532–542.
- Markee, N. (2017). Are replication studies possible in qualitative second/foreign language classroom research? A call for comparative re-production research. *Language Teaching*, 50(3), 367–383.
- Marsden, E., & Mackey, A. (2013, June). *IRIS and replication*. Paper presented at the 9th International Symposium on Bilingualism, Singapore.
- Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository for research into second languages* (pp. 1–21). New York: Routledge.
- Matsuda, P. (2012). On the nature of second language writing: Replication in a post-modern field. *Journal of Second Language Writing*, 21(3), 300–302.
- Maxwell, S., Lau, M., & Howard, G. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70(6), 487–498.
- McNeeley, S., & Warner, J. (2015). Replication in criminology: A necessary practice. *European Journal of Criminology*, 12(5), 581–597.
- Mezias, S., & Regnier, M. (2007). Walking the walk as well as talking the talk: Replication and the normal science paradigm in strategic management research. *Strategic Organization*, 5(3), 283–296.
- Moravcsik, A. (2014). Transparency: The revolution in qualitative research. *PS: Political Science & Politics*, 47(1), 48–53.
- Mu, C., & Matsuda, P. (2016). Replication in L2 writing research: Journal of second language writing authors’ perceptions. *TESOL Quarterly*, 50(1), 201–219.
- Nassaji, H. (2012). Significance tests and generalizability of research results: A case for replication. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 92–115). Cambridge: Cambridge University Press.
- Neuliep, J., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, 5(4), 85–90.
- Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660.
- Pashler, H., & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531–536.
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors’ introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530.

- Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 116–132). Cambridge: Cambridge University Press.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35(4), 655–687.
- Plonsky, L. (2015). Quantitative considerations for improving replicability in CALL and applied linguistics. *CALICO Journal*, 32(2), 232–244.
- Plonsky, L., Egbert, J., & Laflair, G. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, 36(5), 591–610.
- Polio, C. (2012a). No paradigm wars please! *Journal of Second Language Writing*, 21(3), 294–295.
- Polio, C. (2012b). Replication in published applied linguistics research: An historical perspective. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 47–91). Cambridge: Cambridge University Press.
- Polio, C., & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition*, 19(4), 499–508.
- Porte, G. (2012). *Replication research in applied linguistics*. Cambridge: Cambridge University Press.
- Porte, G. (2013). Who needs replication research? *CALICO Journal*, 30(1), 10–15.
- Porte, G., & Richards, K. (2012). Focus article: Replication in second language writing research. *Journal of Second Language Writing*, 21(3), 284–293.
- Reis, H., & Lee, K. (2016). Promise, peril, and perspective: Addressing concerns about reproducibility in social-personality psychology. *Journal of Experimental Social Psychology*, 66, 148–152.
- Richards, K. (2009). Trends in qualitative research in language teaching since 2000. *Language Teaching*, 42(2), 147–180.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.
- Rott, S., & Gavin, B. (2015). Comprehending and learning from Internet sources: A conceptual replication study of Goldman, Braasch, Wiley, Greasser and Brodowska (2012). *CALICO Journal*, 32(2), 323–354.
- Sakaluk, J. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology*, 66, 47–54.
- Santos, T. (1989). Replication in applied linguistics research. *TESOL Quarterly*, 23(4), 699–702.
- Sasaki, M. (2012). An alternative approach to replication studies in second language writing: An ecological perspective. *Journal of Second Language Writing*, 21(3), 303–305.
- Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, 50(2), 212–226.



- Schofield, J. (2002). Increasing the generalizability of qualitative research. In A. Huberman & M. Miles (Eds.), *The qualitative researcher's companion* (pp. 171–203). Thousand Oaks, CA: SAGE.
- Schweinsberg, M., Madan, N., Vianello, M., Amy Sommer, S., Jordan, J., Tierney, W., ... Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55–67.
- Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.
- Smith, B., & Schulze, M. (2013). Thirty years of the CALICO Journal—Replicate, replicate, replicate. *CALICO Journal, 30*(1), i–iv.
- Stake, R. (2005). Qualitative case studies. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (pp. 443–466). Thousand Oaks, CA: SAGE.
- Valdman, A. (1993). Replication study. *Studies in Second Language Acquisition, 15*(4), 505.
- Webb, S. (2015). Learning vocabulary through meaning-focused input: Replication of Elley (1989) and Liu & Nation (1985). *Language Teaching, 49*(1), 129–140.
- Willans, F., & Leung, C. (2016). Empirical foundations for medium of instruction policies: Approximate replications of Afolayan (1976) and Siegel (1997b). *Language Teaching, 49*(4), 549–563.
- Yoon, H.-J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly, 51*(2), 275–301.





# 8

## Ethical Applied Linguistics Research

Scott Sterling and Peter De Costa

### Introduction

Conducting research ethically is paramount for the continuing success of any research field. And increasingly, many journals often stipulate that ethical requirements are met before submitting a paper, with some journals asking contributing authors directly to produce evidence of Institutional Review Board (IRB) approval<sup>1</sup> from their respective institutions. At the same time, however, it is important to note that research ethics takes on a different role when the data being collected and analyzed comes from human beings. Research ethics has been defined in a number of ways including stated definitions, such as that offered by the British Economic and Social Research Council (ESRC, 2015, p. 43) as “the moral principles guiding research, from its inception through to completion and publication of results and beyond” (p. 4). Admittedly, there are many definitions and viewpoints for potential ethical frameworks, too many to be listed here. In light of a space limitation, we offer two such viewpoints below as put forward by Pimple (2002) and Emanuel, Wendler, and Grady (2013).

---

S. Sterling (✉)

Department of Languages, Literatures, and Linguistics,  
Indiana State University, IN, USA  
e-mail: [scott.sterling@indstate.edu](mailto:scott.sterling@indstate.edu)

P. De Costa

Department of Linguistics and Germanic, Slavic, Asian and African Languages,  
Michigan State University, East Lansing, MI, USA  
e-mail: [pdecosta@msu.edu](mailto:pdecosta@msu.edu)

Pimple (2002) breaks down ethical research into three components:

1. Truth in reporting and representing data
2. Fairness in citing and using the work of others
3. Wisdom to only conduct meaningful and useful research

In this view, research is said to be ethical not if it follows a set guideline or checklist but instead looks at the nature of the research and researcher. Is the researcher being truthful in the reporting of his or her information, or is there potential fabrication of data or hiding of non-significant results? In terms of qualitative data, are participants being fairly represented or has the author cherry-picked data to make a point? Fairness revolves around properly citing work and not plagiarizing. Additionally, it could align with fairness in conducting group-based research projects and ensuring that all authors are fairly attributed. Finally, wisdom to conduct meaningful research discusses the need that research is useful to society or science and that researchers are not bringing about undue and unnecessary harm.

An additional conceptualization of ethical research can be seen in Emanuel et al. (2013), who draw on a more positivist view of research as compared with Pimple. These authors suggest that ethical research needs to have:

1. value to science or society
2. scientific validity
3. reliable independent review
4. respect for persons
5. balanced risk-to-benefit ratio
6. fair participant selection
7. truly informed consent

The seven items listed by Emanuel et al. (2013) can be broken into two large themes: (1) value of research (Items 1, 2, and 3) and (2) ethically conducting research (Items 4, 5, 6, and 7). Similar to Pimple (2002), Emanuel et al. discuss the need for research to be useful or have value. For applied linguistics research that might mean pedagogical implications for research but it could also entail furthering our knowledge about the nature of language. A potential project that wouldn't have value would be one that investigates if it were possible to learn a second language as an adult. There is a plethora of research that already shows that adult second language acquisition (SLA) is possible and so this project would not add any real value or knowledge to the field. In fact, a study like this would take away resources that could be spent on more valuable projects such as looking at how adult SLA is accomplished or at a particular interesting feature. Scientific validity is an issue that has been discussed heavily in the last 15 years (Norris & Ortega,

2000; Plonsky & Gass, 2011). The overall goal of this research trend is to point out various methodological and statistical issues the field faces. Ensuring that the studies conducted are done in methodologically rigorous ways might not seem like a question of ethics but how the field uses its resources and ensuring that the results produced are accurate are crucial. Results from linguistics studies can be used by governments or other policy makers to make decisions that affect the lives of millions; basing these policies on a poorly conducted study or misinterpreted data could spell disaster. Finally, an area that has seen little review in the field is just how research is reviewed during the peer review stage.

Another large area of research ethics discussed in Emanuel et al. (2013) that impacts the field of applied linguistics relates to how the field conducts research. A fair selection of participants can be difficult to achieve in applied linguistics. Truly blinded and randomized sampling procedures are usually not available as participants are often selected due to their linguistic background. However, those who conduct research in the classroom are often confronted with issues such as balancing the benefits from being in an experimental group as opposed to a control group. A balanced risk-to-benefit ratio is generally taken to mean that any risks taken on by a participant are balanced by the benefit. Asking a student to contribute to a research project that takes away from their instructional time is a burden, one that might be too large if all the benefits go to other stakeholders such as the researcher, institutions, or future students. Ensuring actual informed consent can be difficult (see Sterling, 2017). Consent documents are difficult to read, often not even read by participants, and often information is purposefully withheld in order to not bias the data.

These definitions provided by Pimple (2002) and Emanuel et al. (2013) highlight the fact that research ethics are consistently negotiated and do not stop after research publications. Both Pimple (2002) and Emanuel et al. (2013) note that research should be conducted only if it is deemed to be useful or worthwhile, a sentiment made clear in Ortega (2005a), who emphasized the social utility of applied linguistics research. Additionally, all of the definitions of research ethics imply that ethical research goes beyond simply doing what is regulated by governmental or institutional bodies and that a large burden of conducting ethical research falls on the researcher.

Building on developments in ethical practices observed in applied linguistics research, this chapter focuses on three main points revolving around the agency that researchers are required to exercise when conducting ethical research. First, issues related to the general lack of ethics training received by applied linguists are highlighted. Second, we focus on IRBs in terms of how these organizations potentially assist and restrict research, as well as the reliance that many researchers place on the IRB. Finally, we will offer advice in

the form of questions that applied linguistics researchers should ask themselves when conducting research.

## Historical Development

Applied linguistics is an international and methodologically diverse field. As such, opinions on what constitute best practices will always be situated within the context of the research, the home institution or funding body, and within an individual researcher's moral and ethical framework. It can be difficult to know what or if major ethical issues have occurred in applied linguistics due to a general lack of publications on the topic. Little has been published explicitly on research ethics when compared to the overall number of methodologically oriented publications produced every year, though we are seeing that change as research methodology and meta-research are becoming more prevalent in the field. That said, however, the issue of ethical practices in applied linguistics has been circulating within the field's broader discourse for more than four decades. For one, it was foregrounded in TESOL's "Guidelines for Ethical Research in ESL," which appeared in a 1980 issue of the *TESOL Quarterly*. The primary focus of this document and that of many other texts that came out in the 1980s (e.g., Brown, 1988) and the early 1990s (e.g., Hatch & Lazaraton, 1991) was the logistical aspects of conducting research. Simply put, the earlier literature sought to highlight the formal procedures associated with setting up a research project. This phenomenon was underscored by Dufon (1993), who pointed out that much of the methodology literature at the time emphasized "theoretical and methodological approaches, statistics, validity, replication, and so forth" (p. 158).

By the mid-2000s, however, research methods books in applied linguistics started to address ethics more explicitly, due in part to the increase in IRB involvement in the research process (Duff, 2008). This deepening interest in carrying out ethical research was exemplified in the coverage of ethics in Mackey and Gass (2005), McKay (2006), Dörnyei (2007), and Phakiti (2014) and has become more prominently addressed in book chapters (e.g., De Costa, 2015; Sterling, Winke, & Gass, 2016), journal articles (e.g., De Costa, 2014; Mahboob et al., 2016), recent methodology volumes (e.g., De Costa, 2016a; Mackey & Gass, 2016; Paltridge & Phakiti, 2015), and special issues of several journals such as *The Modern Language Journal* (Ortega, 2005b), *TESL Canada Journal* (Kouritzin, 2011), and *Diaspora, Indigenous and Migrant Education* (Ngo, Bigelow, & Lee, 2014).

## Current Core Issues

As applied linguists who are interested in research ethics, the authors are constantly having discussions with our colleagues and students about research ethics and are often asked to provide advice for various situations. While anecdotal evidence is often best avoided, it is difficult to find much actual empirical evidence on the general beliefs of researchers in the field. One major issue that we often encounter in our conversations is an overall feeling of apathy toward the topic of research ethics. This apathy can be seen in comments such as “who cares” or “if the IRB says it is fine, then it is fine.”

Apathy can also be seen in the lack of education in research ethics in PhD training programs (see Sterling et al., 2016, for a discussion of ethics training in applied linguistics graduate programs). This lack of training along with a general *creep* of federal regulations into research ethics (Haggerty, 2004) has led to consequences, namely, an overreliance on IRBs as the gold standard for research ethics and research ethics training. However, IRBs are designed to ensure that federally mandated ethical practices are followed when human beings are involved in a research project. While helpful, IRBs are not supposed to be used as a model of ethical correctness. Thus, field-specific ethical training and the development of linguistic ethical research are needed for research within applied linguistics. In one of the few studies investigating applied linguistics researchers' views on research ethics, Sterling et al. (2016) and Sterling and Gass (2017) found that applied linguists place a high degree of faith into the IRB to make ethical decisions for them. Replacing one's agency in making decisions on complex ethical issues and relying instead on a nebulous ethical agency could result in situations in which a university's interests are put above those of our individual participants.

In Sterling et al. (2016) and Sterling and Gass (2017), participants were asked to read ethically challenging scenarios that were designed to fall into an ethical “gray zone.” All scenarios contained issues that while not always purely unethical should have been perceived as problematic. The participants in the study, who were all applied linguistics researchers, largely rated each scenario as having low to no ethical conflict. In other words, even though the fictional situations contained multiple areas of questionable practices, the majority of participants did not perceive them as ethically challenging. In open comments, participants indicated that the situations were ethical because an IRB had approved of them, highlighting the fact that applied linguistics researchers have started to place more responsibility

on the IRB than on their own ethical judgments. This loss of agency might result in scholars who are not properly trained to handle the complex realities of research, which will almost always contain gray areas where no right/wrong decision can be made. Finding no ethical issues present in the various situations, and calling on the fact that the IRB signed off on their imaginary study, further indicates that researchers are not always clear about their ethical duties or the IRB's.

Another topic that comes up in conversation is an overreliance on “general rules of thumb” toward research ethics that are not substantiated with empirical evidence. Issues such as paying student-participants compared to using extra credit, participant understanding of research, length of consent form, participant motivation to be in research, and many others are passed from researcher to researcher, probably in informal contexts as suggested by Sterling et al. (2016). These rules of thumb might be provided during informal conversations in the hallway, through meetings with a mentor, or even passed on through less trained colleagues such as other graduate students or language instructors not specifically trained in research ethics. These rules of thumb might also be passed along during more formal classroom settings as well. While these rules of thumb might contain useful knowledge, there is currently limited data available for how well the rules actually work. A clear example can be found in the length of consent forms. Often the general dictum is that shorter is better. However, Sterling (2017) shows that consent forms written for ESL studies tend to be around two pages long and fluctuate widely on their vocabulary and reading comprehension levels. In fact, writing short consent forms might require researchers to use more complex language and jargon. Thus, the general advice of keeping consent forms short might actually lead more difficult forms and just less ethical practices.

There is a general lack of strong empirical evidence for many of the ethical practices that are field specific. Often, papers on research ethics are taken from the experiences of the author or are written as position papers. Empirical investigations into whether or not policies or guidelines are actually “best practice” are rarely, if ever, investigated. This lack of direct evidence for best practices means that mentors are often left using untested rules of thumb when conducting research or passing on critical ethical information to future researchers. Devising future training materials might be challenging due to the dearth of information we currently know about best practices within the field, making this area of investigation not only ripe for future researchers but of paramount importance for the advancement of the field.

One of the main issues with research ethics in applied linguistics research is the minimal amount of education and training dedicated to the topic. When surveyed, many in the field indicated that research ethics was isolated to a single day/unit of a research methods class and was mainly covered as part of IRB certification or during the IRB application itself (Sterling et al., 2016). Two recent studies (Sterling et al., 2016; Sterling & Gass, 2017) have found similar trends. The authors found a divide between the topics covered that were considered to be *procedural ethics* (Guillemin & Gillam, 2004), items covered during IRB training, and those that might be considered academic integrity items. Procedural ethics were covered more extensively in formal graduate school training, either during IRB training or in research methods courses. By contrast, the academic integrity items, mentorship, authorship, collaboration, and peer review were only discussed informally, if at all.

Taken together, the issues from this section highlight the fact that research ethics in applied linguistics has seen limited empirical research into best practices. We have also discussed the trend of applied linguistics researchers shifting their ethical responsibilities from themselves and onto mandatory ethical review boards. A fear going forward for the applied linguistics research community is that future scholars will not receive quality training on research ethics.

## Challenges and Controversial Issues

In this section, we discuss five different macro-topics that are likely to be played out in research conducted in applied linguistics. Our questions do not have any definite solutions, but seek to raise awareness about the challenges and controversies surrounding the conduct of ethical applied linguistics research. We use the following questions to highlight several facts. First, research is full of pitfalls that are not easily avoidable. At some level, almost any decision will have a negative outcome for someone, even if that negative outcome is negligible. The second reason we selected these questions is to underscore that applied linguistics research ethics need to be considered from the inception of a project to publication and beyond. Finally, these questions highlight the fact that IRBs should not be completely entrusted with ethical decisions. There are many issues that go beyond basic federal and institutional mandates that scholars should consider. In the best-case scenario, a university IRB is staffed with dedicated people trained in law, research methodology, and ethics. Though highly qualified, such individuals cannot be expected to understand the intricacies, complexities, and ethical considerations involved



in all domains of research involving human subjects. As a consequence, it is up to the researcher to constantly consider these issues.

The five questions that are discussed in this section are:

1. Who defines what a language community is and who speaks for them?
2. How do we balance a participant's right to confidentiality with his/her desire to be known for participating in research?
3. How do we ensure informed consent has been collected, given that cultures have differences in opinions on research ethics and the terminology used in the consent process?
4. What right do participants have to their data once a project has been completed? Can researchers continue to use the data they collected for various analyses, or does every new analysis require a renewal of consent?
5. What ethical role does a researcher have after the research paper is published? Are researchers responsible for how their data is appropriated by governments or teachers, and would they need to withdraw a paper if it contains data that has been found to be inaccurate at a later date?

### **Question 1: Who Defines What a Language Community Is and Who Speaks for Them?**

The first question involves a topic that is often discussed in linguistic field method textbooks but one that is alarmingly difficult to answer. On the surface, this is an easy question. I study language X, so my research community constitutes of people Y. To gain access to language X, I would need to ask someone from that language group. However, this type of rationalization may unravel very quickly. Not everyone who speaks language X is part of community Y, and not all people who live in or around community Y are members of that community.

More often than not, language communities in applied linguistics research are actually imaginary constructs created by researchers. For example, even though a researcher might consider "Spanish 101 students" a community, it is quite unlikely that students taking that course feel a strong sense of connection to their classmates. In fact, many language students probably feel closer affiliations to other people from their home country or region regardless of language than they do to the language learning community as a whole.

The question of group identity becomes even more complicated when we view the outcomes of research. Taking the example of English language learners (ELL), who benefits from ELL research? It is unlikely that the ELL partici-



pants in the study will gain anything from the research project, as they will graduate or move out of ELL courses before the paper sees publication. One line of rationalization often used when considering the benefits to research is that a study could impact future cohorts of students, which might include siblings or children of current participants but that is difficult to argue for in terms of community membership, especially with potentially imaginary groups. In the end, it is likely that participants are being asked to absorb all the risks and likely see none of the benefits while others will gain all the benefits at no risk to themselves.

Another common complaint by applied linguists relates to blanket IRB policies that do not appear to work well in applied linguistics research. Typically this complaint surfaces in comments related to linguistic research not being inherently risky, consent being difficult to obtain in some cultures, or confusion over how to treat non-native English speakers. However, this type of complaint is unfair as IRBs are typically not staffed by linguists. If you find yourself in a situation where a general policy will not work with your intended data collection, the best suggestion is having an in-person meeting with your IRB. As with many faceless organizations, it is easy to vilify the IRB as an entity that only wants to make your life more difficult and is guarding the interests of your institution. However, one also needs to remember that many IRB members are volunteer faculty who truly care about protecting people. The worst-case scenario from having a discussion with your IRB is that they will not change their minds and you will find yourself in the same situation where you started. In the best-case scenario, both sides learn from each other, your project is deemed in compliance with the required guidelines, and future applied linguists from your institution working in a similar context will profit from your experience by having an easier time getting the IRB (assuming that its members do not change) to understand their difficult situations. Sample Study 8.1 provides an illustration of this very issue.

### Sample Study 8.1

Duff, P.A., & Abdi, K. (2016). Negotiating ethical research engagements in multilingual ethnographic studies in education: A narrative from the field. In P.I. De Costa (ed.), *Ethics in applied linguistics research: Language researcher narratives* (pp. 121–141). New York: Routledge.

#### Research Background

This chapter examines the ethical dilemmas encountered by the second author (Abdi) in as she prepared to embark on a two-year ethnographic multiple-case study of transnational children moving between Canada and China.

**Research Problems**

Taking at its starting point the unexpected challenges that emerge while conducting ethnographic research, this chapter provides helpful insights into behind-the-scenes interaction that often goes on in order to gain approval from the ethics board of a university. Specifically, the chapter details the complexities surrounding interactions between researchers, the departmental authorities who need to sign off on applications to carry out research, and the staff from the office of research services that oversees ethical review applications.

**Research Method**

Even though the study described in this chapter is based on ethnography, the chapter takes on the form of a narrative in that it describes the unexpected tensions that occur when carrying out field research that involves internationally mobile child participants.

**Key Results**

The central tension encountered by Abdi was having to consider the demands of her ethical review board which enforced protocols that favored a predetermined and static study design. Consequently, she had to grapple with a protracted and time-consuming amendment process and having to educate the board about the nuances involved in conducting research in China.

**Comments**

The chapter underscores how given the evolving nature of ethnography in particular and research in general, some level of procedural flexibility is needed in order for researchers to ensure that ethical research practices are implemented.

## **Question 2: How Do We Balance a Participant's Right to Confidentiality with His/Her Desire to Be Known for Participating in Research?**

In the United States, confidentiality is promised to most participants when they agree to take part in research, and it is often a stipulated IRB requirement. Requiring confidentiality is often not a problem for many large-scale quantitative research projects, as large volumes of data are collected and all traces of participant individuality are removed. However, with many qualitative projects, this is not often the case. Generally, the fewer participants in a research project, the more difficult it is to conceal their identity. If a researcher is collecting data on one of the last four speakers of a language, then there is a 25 percent chance of guessing who the participant is.

Keeping data confidential is important for IRBs, but in some projects, this requirement may contradict what the participant wants. Some participants want it to be known that they participated in the research, or that they contributed to the preservation of a language. If a participant wants to be unanon-

ymous, should the researcher comply with this demand? If an adult participant wants to exercise her agency and claim the fact that she participated in the research, should the researcher object?

An example of this problem occurred in research that Sterling was conducting. He initially wanted to offer his participants the chance to select their own pseudonym for the project. However, in his earlier work, several participants had opted to use their real names for a variety of reasons, the main being that they had nothing to hide and that they trusted Scott. However, participants in this particular study did not have any real indication of what would be asked of them at the onset of the study, nor did they know how they would be represented in the final write-up. Additionally, Scott later discovered that participants, who had all signed a written consent form and verbally agreed that they fully understood the research study, had no idea that he was not a teacher in their ELL program or that he was a linguist. He also learned that they were not even sure what he would do with the data. In the end, Scott opted to use a numbering system for identification instead as he concluded that the participants were not adequately informed enough to make the decision to revoke their confidentiality.

A counterpoint to this argument might be the use of *member checks*, which occurs when a researcher asks participants to read a manuscript prior to publication to ensure that the data accurately represents the participant. Even if a participant agreed to be part of a research project at the beginning and wanted to expose his/her identity, there would still be a chance before publication for the participant to reconsider this decision. While a member-checks strategy is highly recommended, this counterpoint is an excellent place to illustrate the complexities surrounding applied linguistics research and shows that not all research is the same and that providing a “simple solution” will not work in all cases. In Sterling’s study discussed earlier, the data were collected through focus groups. By the time the data were analyzed and written up, the vast majority of the participants had left not only the language program in which the data was collected, but also had graduated from the university and presumably left the country as well. The feasibility of tracking down 40–50 former students in various countries around the world, who will likely have forgotten that they even took part in a one-hour focus group interview years earlier, is a daunting task. Additionally, member checks assume that participants are not only literate but also literate enough to comprehend a scientific document. The current trend in academic publishing would also require that the participant be able to read English at a high level, or that the researcher translate the document into the L1 of the participant, an additional burden on a likely already time-stressed situation.

Ultimately, in many ways, this question of confidentiality will be left to the researcher to decide. Again, offering untested rules of thumb will likely only lead to practices that will probably not work that well in many cases. Your opinions on the agency of participants, the voice that participants want to adopt, and other issues must be entered in to any calculation. One suggestion, however, is to judge the level of seriousness that participants put into answering the question of using their names. Do your participants strongly press for exposing their true identity, or are they indifferent? Returning to Scott's example, his participants appeared to be extremely indifferent to which name was used with many saying something along the lines of "just use my name or whatever." It would be prudent to explain all the possible negative consequences that could surface if their identity becomes known, even if they seem extremely unlikely. In the end, you are responsible for the safety of all your participants both during and after a research project has ended.

### **Question 3: How Do We Ensure Informed Consent Has Been Collected, Given That Cultures Have Differences in Opinions on Research Ethics and the Terminology Used in the Consent Process?**

It is tempting to think that simply responding affirmatively to the question, "Do you agree to be part of my research?", should be sufficient evidence that someone has agreed to take part in a research project. Yet what consent is, and what it looks like, changes between studies and cultures. The basic idea behind informed consent, that a participant understands and voluntarily agrees to be part of a research project, is an idea that many researchers feel comfortable with. Informed consent is not a universal idea that transcends all cultures. In some cultures people are suspicious when asked to sign legal-like documents, and in others a respected person such as a teacher can grant consent for an entire group to be part of a research project.

While cultural differences can be problematic for obtaining informed consent, it should also be noted that consent documents likely make little sense to anyone not familiar with research, even if the participant and researcher share a culture. In experimental research, the true nature of the project is often hidden at the onset of data collection to avoid biasing the data pool, meaning that people have to agree to take part in research but are not actually told what kinds of things are expected of them. In many instances, the full objective is disclosed after a study, but again there are many instances when

disclosing the focus of a study might affect the results if participants talk with each other. Qualitative research encounters similar issues, in that it is not always possible to draw a clear line between when research is being conducted and when a casual conversation is taking place. As mentioned, notions such as the benefits to community or need for confidentiality might be concepts that make sense to the research community; however, these concerns probably have very little real value to or its impact may not be fully realized by someone outside the academy. For an illustration, see the sample study presented in Sample Study 8.2.

### Sample Study 8.2

Lee, E. (2011). Ethical issues in addressing inequity in/through ESL research. *TESL Canada Journal*, 28(31), 31–52.

#### Research Background

Research ethics obligations stretch far beyond the data collection and retention process and at times these responsibilities will be split between various stakeholders in a research project. An example of unforeseen problems with implications for different groups can be found in a report by Ena Lee (2011) who reports a predicament she found herself in during the course of an ethnographic case study of an ESL program. Her stated goal for the project was to observe how the ESL program mixed culture and language instruction, and how that mixture affected the identities of those in the program.

#### Research Problems

Unlike traditional views of protecting participants from harm or reducing risks, Lee's article dealt with the long-term needs of protecting identity, confidentiality, and the dangers associated with the loss of those features. While applied linguistics tends to be a low-risk field, that doesn't mean participants face no dangers. Instead, problems can arise at any time before, during, and after a research study.

#### Research Method

This study takes a narrative, anecdotal perspective and provides the reader with a detail account of how a single author struggled to navigate a difficult ethical issue.

#### Key Results

During Lee's original study, she noticed that one of her instructor-participants was being racialized by the other staff members in the program. Lee wanted to publish data about this incident but recognized that her instructor-participant would be easy to identify in any report and thus at risk of losing her job. Of course, Lee had received consent to conduct this study, but the participants' treatment at the language center would not have been discussed then. Lee was left feeling of guilt at knowing that an unfair practice was taking place and not

being in a position to do anything about it. As a researcher ethnographer, Lee is expected to be objective and not place value judgments on a situation. While one might feel that treating someone differently because of their race is wrong, a researcher supposed to observe and report back to a larger community. In the end, Lee decided to not publish this portion of the data until the participant moved to a new job and when her career was no longer in jeopardy if she was exposed.

#### Comments

Lee's story is interesting as it shows the unexpected consequences that can take place during a research study. It also highlights the ongoing need to think ethically and responsibly that go beyond the end of data collection. Finally, it should be mentioned that no part of this story is likely to surface in IRB training and that Lee was forced to take control over her own ethical understanding in order to make a decision.

Hence, if participants are not able to know what is expected of them in a research project, can we ever really claim that participants are informed? How do we address assumptions that are not explicitly stated in the documents? As applied linguists we understand that data will be presented at conferences, that graduate student assistants might be asked to code data, or that a different researcher might be invited to look at the data for a reliability check. But do our participants share this belief? The answer to this question remains unclear, but it seems unlikely.

The advent of the internet has created new public spaces for people to post their thoughts in either textual or audio formats. These public forums afford applied linguistics researchers the opportunity to investigate daily language usage by people around the world, but does this mean that we can analyze this information without consent from the author? (For further discussion of conducting ethical online research, see Gao & Tao, 2016.) Most people would say "yes," because this communication is in the form of texts that exist in the public sphere; however, they are also generally in the form of social media posts intended for friends and family members. For instance, is it ethical to analyze how one's aunt discusses political issues of gender and identity on Facebook? What about using a website that hacks people's phones and uploads personal text messages that were never intended to be read by others? For many, the last example might seem unethical, but is using someone's hacked personal communication different from analyzing the letter correspondence or diary of someone from history? Again, we find ourselves running into a gamut of issues that are difficult to answer. The simple question of "do you agree to take part in this research" is alarmingly oversimplified. In short, consent in research is a much larger topic than we can discuss in this chapter but nevertheless requires further consideration.

#### **Question 4: What Right Do Participants Have to Their Data Once a Project Has Been Completed? Can Researchers Continue to Use the Data They Collected for Various Analyses, or Does Every New Analysis Require a Renewal of Consent?**

Through the use of digital corpus linguistic tools, we have the ability to sift through thousands of hours of recorded data and look for similarities between when a person uses the past tense and when they use the past perfect in English. Yet, unless data were specifically collected as part of a corpus, should researchers be able to repurpose the information? Do participants have any say as to what happens to their data after they have been collected?

Reusing data is a fairly common research practice. Both authors have been in positions when we were studying a particular piece of data and noticed that it contained interesting elements that we never really considered at the onset of the project. Scott collected data in an ELL class where his students were asked to have a discussion with four strangers. What he never expected was to see a high frequency of humor used to navigate difficult topics such as terrorism, homosexuality, and harmful stereotypes. After reviewing the transcripts, it was clear that this data would be interesting to investigate and Scott collected retroactive consent from the participants. This was easy enough to do since the course was still in progress. But what would his options have been had he waited a few more weeks and his ability to track down the participants lessened? In Peter's case, it wasn't his original intention to visit the dormitory in which the female participants in his school-based ethnography stayed. Initially, his plan was to conduct observations within the premises of the school, but after learning about the tight living conditions under which they lived, and in order to better understand the sociolinguistic realities of the scholarship students with whom he worked, he decided to visit their dormitories to learn about the out-of-school dynamics that also influenced their language learning outcomes. Fortunately, he was successful in applying for permission to conduct his closely monitored visit, which in turn shed new light on their learning experience (De Costa, 2016b).

As noted, after participants have finished their part of a research project, they typically scatter and can be extremely difficult to track down. Thus, the ability to gain retroactive consent to use data for future projects is problematic. Many readers might be wondering if it is possible to simply file an IRB application as already existing data. The answer is usually "yes." However, applying for data as already existing is akin to telling the IRB that risks will be

minimal to the participants. While reusing data is not itself an ethical violation, the problem here is that we would be equating gaining IRB approval with absolute ethicality. We might also be violating participants' trust by exposing their data in ways they had never agreed to.

Put simply, gaining permission to collect data is only the first necessary step in data collection. As highlighted in this section, there will never be a perfect answer for when it is acceptable to reuse data and when new consent should be collected. It is probably fairly common practice to reuse data when the data collected and the new analyses are closely aligned. Using interview data that were originally intended to focus on learner written identity to investigate how learners use various tenses when writing a story is probably an acceptable practice to most. Using pieces of audio from the interview data just described as part of another research project, one where you are attempting to elicit opinions from strangers on accents from native speakers of the language, decidedly goes beyond the consent form signed.

### **Question 5: What Ethical Role Does a Researcher Have After the Research Paper Is Published? Are Researchers Responsible for How Their Data Is Appropriated by Governments or Teachers, and Would They Need to Withdraw a Paper If It Contains Data That Has Been Found to Be Inaccurate at a Later Date?**

For many people, publication is the final stage of a research project. Once a paper has been submitted, the data needs to be safeguarded, but some applied linguists would argue that this is the end of the ethical contract. Yet, we also need to consider what role do researchers play in the continuing use of their study. Shohamy (2004) describes how she and her colleagues published data that were later used by the Israeli government to limit the number of Ethiopian students allowed into Israeli schools. This was never the intention of the researchers, but their actions had negative ramifications for Ethiopian students as the findings were used to “limit the number of students from Ethiopia in Israeli schools and to transfer others to ‘more successful’ schools, a discriminatory, unethical policy that our research was seen to legitimize” (p. 730). Are the researchers responsible? Do they need to make public statements or defend their wrongly cited work? As it can be seen, there are no clear answers to this type of question.

Another issue that has surfaced in other fields is the question of whether papers that contain inaccurate information should be retracted from published



sources. Imagine researchers finding correlations between specific genes believed to control some aspect of language, but later that the gene in question is found to have nothing to do with language at all and that the original data were anomalous. Should the authors retract their paper or do we allow the publication to stand since it was published in good faith and with the best understanding of the phenomena at the time? Or does a question like this open the door for papers to be pulled due to a lack of support of a particular theory or methodological assumption?

Another issue to consider in publication is the confidentiality that we can actually guarantee to our participants. Ironically, participants in a study have the least amount of confidentiality from those with the most influence on their lives. For example, a study might be focused on investigating the amount of grammar knowledge that Spanish students gain over the course of a single year. If the students do particularly poorly on the task, the results could be interpreted as being a fault of the instruction. Most people who read this article will have no idea who the instructor is. However, anyone involved in the project is likely to remember the researcher's name and could probably piece together enough information to figure out the identity of the instructor. While unlikely, the information found in the study could have ramifications on the career of the teacher.

In sum, the various issues brought up in this section highlight the fact that researchers have a moral obligation that stretches beyond just safeguarding data after publication. What exactly this obligation is must be decided upon by the individual researcher who has to wrestle with the dilemma of conducting her research in an ethical manner.

## Resources for Further Reading

De Costa, P. I. (Ed.). (2016). *Ethics in applied linguistics research: Language researcher narratives*. New York: Routledge.

This edited volume contains chapters from various scholars in the field with each chapter focusing on observation and advice of scholars in applied linguistics. Chapter topics range from education in research ethics to the ethical use of technology in data collection. The works in this volume are useful to researchers at any phase of their career and will generate discussion in graduate-level classes.

Mackey, A., & Gass, S. (2016). *Second language research: Methodology and design* (2nd ed.). New York: Routledge.

Chapter 2 in this book provides helpful information about codes of ethics and the procedures that need to be carried out to ensure that informed consent is obtained and participation protection is observed. This updated edition also details ethical concerns surrounding online data collection.

Mahboob, A., Paltridge, B., Phakiti, A., Wagner, E., Starfield, S., Burns, A., ... De Costa, P. I. (2016). TESOL quarterly research guidelines. *TESOL Quarterly*, 50(1), 42–65.

This article gives an overview to the *TESOL Quarterly* research guidelines, one of the few explicitly stated guidelines for applied linguistics research. Areas of focus include the review process, brief introduction to common methodologies, and a section centered on research ethics. The research ethics section discusses the macro-/microethical divide as well as other areas covered during this current chapter.

Ortega, L. (2005a). For what and for whom is our research? The ethical as transformative lens in instructed SLA. *Modern Language Journal*, 89(3), 427–443.

This article can be found in a special issue of *The Modern Language Journal* focusing on ethics, methodology, and epistemology (also edited by Ortega). Ortega argues for applied linguistics research to take a more social perspective including: (1) social utility should be used to judge research, (2) research is never conducted in a vacuum and thus always has a value, and (3) epistemological diversity is benefit to a field.

## Note

1. We will use the term IRB or Institutional Review Board throughout this chapter. An IRB is the American version of ethical review boards that are common in many countries. The general goal of an IRB is ensure that governmental requirements of ethical behavior are being followed. Additionally, IRBs provide support for technical ethical issues such as ensuring that data is securely kept and that participants are being treated fairly.

## References

- Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.
- De Costa, P. I. (2014). Making ethical decisions in an ethnographic study. *TESOL Quarterly*, 48(3), 413–422.
- De Costa, P. I. (2015). Ethics in applied linguistics research. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 245–257). London: Bloomsbury.
- De Costa, P. I. (Ed.). (2016a). *Ethics in applied linguistics research: Language researcher narratives*. New York: Routledge.
- De Costa, P. I. (2016b). *The power of identity and ideology in language learning: Designer immigrants learning English in Singapore*. Dordrecht: Springer.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.
- Duff, P. A. (2008). *Case study research in applied linguistics*. New York: Lawrence Erlbaum Associates.
- Dufon, M. A. (1993). Ethics in TESOL research. *TESOL Quarterly*, 27(1), 157–160.
- Emanuel, E. J., Wendler, D., & Grady, C. (2013). What makes clinical research ethical? *The Journal of the American Medical Association*, 283(20), 2701–2711.
- Gao, X., & Tao, J. (2016). Ethical challenges in conducting text-based online applied linguistics research. In P. I. De Costa (Ed.), *Ethics in applied linguistics research: Language researcher narratives* (pp. 181–194). New York: Routledge.
- Guillemin, M., & Gillam, L. (2004). Ethics, reflexivity, and “ethically important moments” in research. *Qualitative Inquiry*, 10(2), 261–280.
- Haggerty, K. D. (2004). Ethics creep: Governing social science research in the name of ethics. *Qualitative Sociology*, 27(4), 391–414.
- Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle and Heinle.
- Kouritzin, S. (Ed.). (2011). Ethics in cross-cultural, cross-linguistic research [Special issue]. *TESL Canada Journal*, 28(2).
- Lee, E. (2011). Ethical issues in addressing inequity in/through ESL research. *TESL Canada Journal*, 28(31), 31–52.
- Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mahboob, A., Paltridge, B., Phakiti, A., Wagner, E., Starfield, S., Burns, A., ... De Costa, P. I. (2016). TESOL quarterly research guidelines. *TESOL Quarterly*, 50(1), 42–65.
- McKay, S. L. (2006). *Researching second language classrooms*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ngo, B., Bigelow, M., & Lee, S. (Eds.). (2014). Introduction: What does it mean to do ethical and engaged research with immigrant communities? Special issue:

- Research with immigrant communities. *Diaspora, Indigenous and Migrant Education*, 8(1), 1–6.
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528.
- Ortega, L. (2005a). For what and for whom is our research? The ethical as transformative lens in instructed SLA. *Modern Language Journal*, 89(3), 427–443.
- Ortega, L. (Ed.). (2005b). Methodology, epistemology, and ethics in instructed SLA research [Special issue]. *Modern Language Journal*, 89(3), 315–488.
- Paltridge, B., & Phakiti, A. (Eds.). (2015). *Research methods in applied linguistics: A practical resource*. London: Bloomsbury.
- Phakiti, A. (2014). *Experimental research methods in language learning*. London: Bloomsbury Publishing.
- Pimple, K. D. (2002). Six domains of research ethics: A heuristic framework for the responsible conduct of research. *Science and Engineering Ethics*, 8(2), 191–205.
- Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, 61(2), 325–366.
- Shohamy, E. (2004). Reflections on research guidelines, categories, and responsibility. *TESOL Quarterly*, 38(4), 728–731.
- Social Research Council. (2015). ESRC Framework for Research Ethics. Retrieved from <http://www.esrc.ac.uk/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015/>
- Sterling, S. (2017). Investigating the complexity of consent forms in ESL research. [Manuscript in preparation].
- Sterling, S., & Gass, S. (2017). Perspectives from the field on ethical behavior in applied linguistics and second language research. [Manuscript in preparation].
- Sterling, S., Winke, P., & Gass, S. (2016). Training in research ethics among applied linguistics and SLA researchers. In P. I. De Costa (Ed.), *Ethics in applied linguistics research: Language researcher narratives* (pp. 15–37). New York: Routledge.



# 9

## Writing a Research Proposal

Sue Starfield

### Introduction

One of the earliest documents that doctoral scholars have to write is a research proposal in which they provide a rationale and motivation for the research study they plan to undertake. It is a cognitively challenging activity that “demands thinking logically through the entire project from beginning to end” (Rogers, Zawacki, & Baker, 2016, p. 58). As Paltridge (1997) tells us, even though your research focus may change and develop over time, the proposal is “a key first step in the research process” (p. 62). Throughout your future career as a researcher, there will be opportunities to write different kinds of research proposals such as, for example, grant funding proposals. In some contexts, doctoral students write dissertation grant proposals (Cheng, 2014) seeking funding for their doctoral study. While there are many similarities between doctoral research proposals and grant funding proposals, they are written for different purposes, the main one being the funding component of the grant proposal. In this chapter, however, the focus will be on the research proposal doctoral students in most contexts are expected to write and submit for feedback and approval prior to being allowed to commence their study.<sup>1</sup>

---

S. Starfield (✉)

School of Education, UNSW Sydney, Sydney, NSW, Australia

e-mail: [s.starfield@unsw.edu.au](mailto:s.starfield@unsw.edu.au)

In a post on her academic writing blog, Pat Thomson argues that the purpose of a research proposal is a *bid* by the author to gain acceptance into a community: “to demonstrate that the researcher has the capacity to produce disciplinary knowledge.” She writes that “In order to do so, the proposal writer must show familiarity with the “right” language, knowledge production practices, existing debates and taken for granted “truths” of the relevant scholarly community” (<https://patthomson.net/2014/06/23/the-research-proposal-as-writing-work/>). Clearly, becoming an “insider” to disciplinary ways of thinking and writing is part of the process of becoming a doctoral scholar and developing a research proposal is a stage in the process. At the same time, however, there are a number of components of a research proposal that are common across all disciplines. In fact, there is already much advice available to students on how to put together a research proposal both online and in manuals and handbooks (e.g., Creswell, 2014; Paltridge & Phakiti, 2015; Paltridge & Starfield, 2007; Punch, 2012; Tracy, 2013).

While completed doctoral theses and dissertations are widely available online on sites such as *ProQuest Dissertations & Theses Global* (<http://www.proquest.com/products-services/pqdtglobal.html>), proposals, on the other hand, typically have only a small number of readers, and it is quite difficult to gain access to examples of successful research proposals. Punch (2012) contains examples of research proposals from several disciplines that use qualitative, quantitative and mixed methods approaches. He also provides a useful list of other manuals that contain examples of research proposals (pp. 138–140). There is also now at least one online open access database containing sample dissertation research proposals from a range of disciplinary fields and adopting different methodologies ([http://www.ut-ie.com/s/sample\\_diss.html](http://www.ut-ie.com/s/sample_diss.html)). Although none of these proposals is from an applied linguistics field, and proposal formats will differ due to disciplinary and methodological differences, it is certainly worthwhile to look at a number of research proposals as you set out to conceptualise your own study and draft your proposal.

The core constituents of a research proposal typically listed in much of the standard advice given to commencing doctoral students are (see e.g., Paltridge, 1997):

- The aims of the research project
- The scope of the research project
- The significance and originality of the research project in relation to previous work in the field
- The research approach or method to be adopted
- A provisional chapter plan
- A provisional timetable for future progress

While these fairly abstract notions provide a starting point for conceptualising both the research and the proposal, they fail to signal what Pat Thomson (cited above) is putting forward in her blogpost: that the work of the proposal is to gain acceptance of the arguments we are making in our proposal for why our study is needed. We are trying to persuade our readers—probably not more than two to three people in the case of a student’s research proposal—that we have a topic worthy of study that is relevant and of interest to members of our field. In the remainder of this chapter, I want to provide you with some writing and thinking tools that, based on my experience, can help develop your ability to put the case for your applied linguistics research project.

## Tools for Developing a Research Proposal

In a course I taught for a number of years for PhD students in the Faculty of Arts and Social Sciences at my university on developing a research proposal, we began with an activity that directly addressed the *work* of the proposal using the *four questions framework* (see also Starfield & Paltridge, 2014), which asks students to briefly respond (in writing) to the following prompts as illustrated in Fig. 9.1.

These four questions lie at the heart of the research proposal and through attempting to answer them, students can begin to generate both their

1. What is the question I am trying to answer?

2. Why is it worth answering?

3. How have other people tried to answer it?

4. How am I going to go about answering it?

Fig. 9.1 The four questions framework

thinking and their writing. The students who took my course found that returning to these four questions as they drafted and redrafted their research proposal was very helpful in conceptualising the focus of their study. Of course, over the course of six months their initial question might evolve or change substantially, but using the framework allowed them to keep focussed on the core purposes of a research proposal.

“What is the question I am trying to answer” refers to the aims of the project and the potential research questions which need to be clearly articulated. Using the first person, “I” encourages ownership of the project which is important as students transition from an undergraduate identity to one in which they have a much greater degree of agency in defining and designing the parameters of their study.

“Why is it worth asking” refers to the “significance and the originality” of the proposed research or what is also called the “so what” or “why bother” question. In other words, what will your study contribute to the field you are researching? Over the period of developing your research proposal, you should become able to more clearly articulate how your study will contribute. There are of course many ways in which we can articulate such a contribution (see Fig. 9.3 below), and it may take quite a bit of drafting, discussion and redrafting before the nature of this contribution takes shape.

The question that follows: “how have other people tried to answer it” reminds us, as indicated in the list above of the core constituents of a proposal, that claims to significance and originality have their origin in previous work in the field and have to be argued for. The fourth and final question gets us thinking about the methodology we want to adopt—your investigative approach and why it is the most appropriate approach to investigating your specific question.

So, rather than just providing a set of headings such as aims, literature review or methodology, the four questions framework asks you to consider the *functions* of the different components of the proposal or as I said earlier, the *work* it is trying to do. In the following sections, I address each of the four questions in more detail.

## What Is the Question I Am Trying to Answer?

Research questions do not emerge fully formed from anyone’s head: they are the outcome of much thinking, writing and going back and forth to the literature. Looking at successful research proposals or completed dissertations can thus be a bit misleading as they do not convey the many iterations of the



questions that would have led to the set of neatly posed questions in the final proposal. Continuing to use the four questions framework as you progress in your thinking and reading can help you get closer to the final version. You might also want to try the writing prompts suggested by Rowena Murray (2002, p. 104) to kick-start writing about the context/background to your study:

- My research question is .... (50 words)
- Researchers who have looked at this subject are .... (50 words)
- They argue that .... (25 words)
- Researcher A argues that .... (25 words)
- Researcher B argues that .... (25 words)
- Debate centres on the issue of .... (25 words)
- There is still work to be done on .... (25 words)
- My research is closest to that of Researcher A in that .... (50 words)
- My contribution will be .... (50 words)

Writing prompts can be very helpful ways of generating thinking and writing, especially when clear word limits are provided. Working through the prompts leads you to your contribution, as they encourage you to think about how understanding debates in the field and what still needs to be done can help frame your study's contribution and how all the four questions are seamlessly related. As Murray points out, working through the prompts helps to establish the focus and direction of your thesis.

Paltridge and Starfield (2007, p. 59) suggest a number of ways to refine a research question. These include:

- Read broadly and widely to find a subject about which you are passionate. Immerse yourself in the literature, use your library, read the abstracts of other recent theses and dissertations, check dissertations on the web.
- Narrow your focus to a single question: be disciplined and not overambitious.
- Be prepared to change or modify your question if necessary.
- Be able to answer the question "Why am I doing this project?" (and not a different one).
- Read up-to-date materials—ensure that your idea is achievable and no one else has done or is doing it.
- Consult other students who are further down the track, especially those who have the same advisor as you.
- Discuss your ideas with your advisor and lots of other people.

- Work through the implications of your research question: consider existing materials and ideas on which it is based, check the logic, spell out methods to be used.
- Condense your research question(s) into two sentences, write them down—with pride—and post them next to your computer, where you can see them daily. Change the question(s) if needed.
- Ask yourself: What will we know at the end that we did not already know?

## How Have Others Tried to Answer It?

Another activity I used encouraged students to conceptualise their “research space” (Feak & Swales, 2011) using a visualisation activity. Students are given a handout with a simple Venn diagram of three intersecting circles (see Fig. 9.2) and asked to group the key literatures or authors in one of the three circles—the intersection of the three circles is where the topic of the thesis is located—what Swales (2004) calls their research “niche.” This activity is best done on hard copy in my experience. Physically writing in the circles promotes a different kind of thinking than the more linear writing or typing on a page and often encourages students to see connections not apparent earlier (see also Starfield & Paltridge, 2014, pp. 112–115).

It is vitally important that you come to understand that the literature review section is not simply a listing or summary of everything you have read or an attempt to convince the reader that the writer is knowledgeable about the

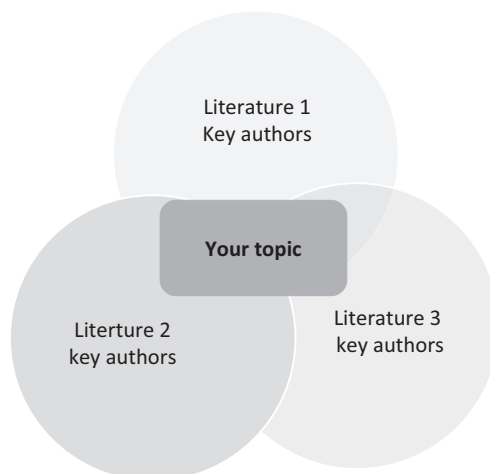


Fig. 9.2 Visual prompt for a literature review

work of others (see also Li & Wang, Chap. 6). As Rudestam and Newton (2007) point out, you are using the literature to develop both a coherent argument and a conceptual framework for your study so that your reader should be able to conclude: “Of course, this is the exact study that needs to be done at this time to move knowledge in this field a little further along” (pp. 63–64).

In the early stages of doing your literature review, it may be that what you are writing does look like a list or summary—a collection of annotated bibliographies put together as you read widely in your topic area. This is entirely to be expected and I encourage my students to develop a note-making system, either in paper format or probably more usefully as a soft copy file, to make brief notes on whichever article, book chapter or book they are reading. Noting down the author’s theoretical perspectives, main arguments and evidence for these as well as the article’s relationship to your own work will all be helpful at this initial stage and later on. This preliminary writing is the beginnings of our writing and actively engages our thinking too as the activity causes us to not just passively consume the texts we’re reading but to engage with the thinking of the authors of these texts and locate their writing within broader theoretical frames. Many researchers now use *EndNote* and other similar software for these same tasks. From early on in your research career, it is a good idea to develop a method of annotated reading and notetaking that works for you as you will be doing a lot of reading.

As your literature review develops over time, it must move beyond the inventory/summary stage into the argument that supports the *bid* your proposal is making. Rudestam and Newton (2007) suggest several quite simple but not immediately obvious ways in which the literature review can begin to sound more like your “take” on the topic so that your own “voice” begins to emerge:

- Try to avoid beginning your sentences with “Jones said ...”; “Smith found ...”—this shifts the focus of your reviews from your own argument to the work of others.
- Try to “develop a theme and then cite work of relevant authors” (p. 65) to support your arguments or to provide examples or counterexamples of your point.
- Try to limit excessive quoting. This can also lessen your authority and control.
- Try to avoid reporting everything. Be selective—“build an argument not a library” (p. 66).

## Why Is It Worth Answering?

The second question in the four questions framework asked you to think about why your research is worth doing. Of course it is, but what we do not always succeed in doing is persuading others that it is. Sometimes this is because the idea of doing research that is significant and original seems overwhelming and impossible. How am I, a PhD student, going to produce this kind of research is a not uncommon thought. The now quite well-known comment by an experienced thesis examiner, “it’s a PhD not a Nobel Prize” (Mullins & Kiley, 2002, p. 386) acknowledges the anxiety experienced by many doctoral students that their contribution may not be judged to have a sufficient degree of originality, that is, be worthy of a Nobel prize!

Rowena Murray helpfully provides a list of ways in which we can think about originality. It is not, she argues, simply saying something no one has said before. What her list does is provide us with ways of thinking about where the originality of our work might lie and how we can best articulate this. What is important, and what doctoral scholars in the early stages of their studies sometimes find difficult, is to begin to think about the ways in which their work might demonstrate an original or significant contribution. My students find the list below, from Murray (2002, p. 59), gets them thinking about this dimension of the proposal, and, in addition provides some tools for having a conversation with their supervisor.

- You say something no one has said before.
- You do empirical work that has not been done before.
- You synthesise things that have not been put together before.
- You make a new interpretation of someone else’s material/ideas.
- You do something in this country that has only been done elsewhere.
- You take an existing technique and apply it to a new area.
- You work across disciplines, using different methodologies.
- You look at topics that people in your discipline have not looked at.
- You test existing knowledge in an original way.
- You add to knowledge in a way that has not been done before.
- You write down a new piece of information for the first time.
- You give a good exposition of someone else’s idea.
- You continue an original piece of work.

The activity in Fig. 9.3 is set up to help you think about the contribution of your study. One of the criteria thesis examiners are asked to consider is the extent to which the study makes a significant contribution to the field. It’s important to

<p>Why is it worth answering? Why bother?</p> <p>My topic/question is significant because ....</p> <p>My topic/question is contributing ..... [select one or more of the boxes below and explain briefly in what ways your study will contribute to your field]</p>	
<p>Theoretically</p>	<p>Empirically</p>
<p>Socially/politically</p>	<p>To practice/policy</p>

**Fig. 9.3** How is my study contributing?

think about the specific ways in which your study may be contributing. Certainly, your proposal should identify these and, as stated above, locate your contribution in relation to previous work in the field. The first prompt you are asked to write to focuses on the significance of your study, while in the quadrants beneath you are asked to consider the nature of your contribution and jot down some reasons that support your claims. Your study does not have to contribute across all four domains: it may only be contributing in one of the four or its significant contribution may lie elsewhere. For example, if your work is in education or Teaching English to Speakers of Other Languages (TESOL), you may see your work as contributing to pedagogy rather than to any of the four quadrants. Alternatively, as an applied linguist you may see your contribution as being

more within the realm of methodology or research design. Use the worksheet in Fig. 9.3 to brainstorm ideas about your study's significance and contribution as you work through versions of your proposal draft.

## How Am I Going to Go About Answering It?

This question relates to the investigative approach you will use to answer your research question. As with the question about how others have answered it, the response to this question develops into an argument for why the approach you have chosen to adopt is the best approach for answering your questions. The methodology and methods section of your proposal needs to put forward a logical justification for your choice of research paradigm and research methods that will assure your supervisor and panel members that your proposed study is both viable and feasible within the time available.

Chapters 3, 4 and 5 of this volume discuss the three main research paradigms used by applied linguists: quantitative, qualitative and mixed methods. As Richard Young argues in Chap. 2, a student's choice of paradigm or methodology is often shaped by the "habits of mind" that draw them to a particular graduate school or advisor or that they develop in graduate school. Your choice of investigative approach—how to answer your research question(s)—will be shaped to some extent by these habits of mind but also by the nature of your topic, by the conversations you have with your peers as well as by the prevailing zeitgeist. Christine Casanave (2014, p. 59) advises doctoral students not to "embark on a methodological approach that you are likely to hate" and also recommends seeking out an "adviser who is philosophically compatible with you" and "who can guide your development." It is important to think through these issues of methodological choice as you are committing to a number of years, if not a lifetime, of working within the particular paradigm you choose to adopt for your dissertation.

An important distinction that we point out in our book (Paltridge & Starfield, 2007) is between methodology and methods; although the two terms are often used as if they were interchangeable, in my view they refer to quite different aspects of the research process. Methodology refers to the research paradigms alluded to above—the epistemologies or ways of knowing that shape the kinds of knowledge our investigative approach can help us uncover (see also Phakiti & Paltridge, 2015 for further discussion of research paradigms in applied linguistics research). For example, within quantitative research, psycholinguistically oriented studies, such as those described in Chaps. 14 and 15 of this volume into the processes underlying language learning are likely to lead to experimental studies of how students learn that seek to measure that learning.

More qualitative methodologies may lead to studies that involve trying to understand the perceptions and identities of those being studied through methods such as interviews and/or observation (e.g., Chaps. 11 and 12 by Matthew Prior and Fiona Copland, respectively). Using a mixed methods approach could involve bringing together the strengths of quantitative and qualitative approaches to adopt methods such as a large-scale questionnaire/survey followed by interviews with participants who have completed the questionnaire to gather more in-depth information.

## Your “Two-Pager”

When you have worked through the activities and writing prompts in the chapter, I recommend you try to write a “two-pager” as suggested by Punch (2012, p. 80). The instructions are quite simple:

- Write no more than two pages using single spacing.
- Describe as clearly and directly as possible what your proposed research is trying to find out and how it will do it.

Focus more on:

- **What** am I trying to find out?
- **How** am I going to do it?

and **less** on context, background and literature.

Your two-pager is a work-in-progress document, but it is an important first step in drafting the proposal. I always asked my students to work on a two-pager about half way through the semester, and they would then swap with one another in class and provide peer feedback on each other’s drafts. If you can find a peer to do this activity with, I think you will notice the benefit.

## So What Will My Research Proposal Look Like?

Up until now, I have been advising you on how to think and write about the different constituents of your proposal—how to conceptualise it. In this section, I want you to now give it a more formally recognisable shape and form as you get ready to submit it to your advisor for review. Table 9.1 lays out the most commonly used section headings for thesis proposals and summarises the purpose of each section.

**Table 9.1** Thesis proposals: structure and purpose (based on Paltridge & Starfield, 2007, p. 61)

Section	Purpose
Title	To summarise, in a few words, what the research will be about
Summary	To provide an overview of the study which you will expand on in more detail in the text which follows
Overall purpose	To present a clear and concise statement of the overall purpose of the research
Relevant background literature	To demonstrate the relationship between the proposed study and what has already been done in the particular area; that is, to indicate the “gap” that the study will fill
Research question/s	To locate the study within a research paradigm. To provide an explicit statement of what the study will investigate
Definitions of terms	To provide the meaning of the key terms that have been used in the research question/s
Research methodology	To give an illustration of the steps the project will go through in order to carry out the research
Anticipated problems and limitations	To show awareness of the limitations of the study, what problems may be met in carrying it out and how they will be dealt with
Significance of the research	To say why the study is worth carrying out
Resources required/budget	To say what resources the research will require—and what other costs may be anticipated in carrying out the study
Ethics	To provide a statement as to how participants will be advised of the overall nature of the study and how informed consent will be obtained from them
Proposed table of contents	To give an overview of the scale and anticipated organisation of the thesis or dissertation
Timetable	To give a working plan for carrying out, and completing, the study
References	To provide detailed references and bibliographic support for the proposal
Appendix	To provide examples of materials that might be used, or adapted, in the study

## Criteria for Assessing Research Proposals

Once your proposal is ready to be submitted to your advisors or panel, you may be wondering how they will evaluate it. Cadman (2002) surveyed and interviewed thesis supervisors (advisors), asking them to prioritise the particular features they expected to see in a research proposal. If you have worked through this chapter and the suggested activities, her findings may not surprise you. The supervisors indicated that they gave most value to:



- the logic of the student's argument
- a well-focussed research question, set of research objectives or hypothesis
- the width and depth of the student's reading
- the feasibility of the student's project
- a critical approach to the literature
- justification of the project through the literature
- understanding of current issues on the student's topic
- matching of methodology and methods to the research questions

Now read through your proposal, imagine you are the supervisor of your thesis or a member of your panel. Review your proposal using the criteria listed above and make changes you think are needed.

## Conclusion

Embarking on a doctoral thesis or dissertation is a major life event that is, I believe, in the majority of cases, a life-transforming one. Having said that, it is a huge investment on many levels and should not be engaged in without serious consideration. While this chapter has focussed on writing a research proposal, I would like to recommend an extremely thoughtful book that every prospective doctoral student should read long before getting to the proposal drafting stage. Aptly titled *Before the Dissertation* (Casanave, 2014), it explores (and explodes) many of the myths and realities of the doctoral journey honestly and clearly.

## Resources for Further Reading

Casanave, C. (2014). *Before the dissertation: A textual mentor for doctoral students at early stages of a research project*. Ann Arbor: University of Michigan Press.

This wise and useful book is essential reading for those either contemplating or in the early stages of doctoral study.

Creswell, J. (2014), *Research design* (4th ed.). Thousand Oaks, CA: SAGE.

A key text for students designing a research project, Creswell covers all the angles.

Punch, K. (2012). *Developing effective research proposals* (2nd ed.). London: SAGE.

Shorter and sharper than Creswell, this book is a thorough beginner's guide to developing a research proposal, written in a highly accessible style.

## Note

1. In the North American context, PhD submissions are known as dissertations, while in countries with British higher education traditions, they are referred to as theses. In this chapter, I use them interchangeably to refer to the written submission of a doctoral candidate for examination.

## References

- Cadman, K. (2002). English for academic possibilities: The research proposal as a contested site. *Journal of English for Academic Purposes*, 1, 85–104.
- Casanave, C. (2014). *Before the dissertation: A textual mentor for doctoral students at early stages of a research project*. Ann Arbor: University of Michigan Press.
- Cheng, Y.-H. (2014). Dissertation grant proposals as “writing games”: An exploratory study of two L2 graduate students' experiences. *English for Specific Purposes*, 36, 74–84.
- Creswell, J. (2014). *Research design* (4th ed.). Thousand Oaks, CA: SAGE.
- Feak, C. B., & Swales, J. M. (2011). *Creating contexts: Writing introductions across genres*. Ann Arbor, MI: University of Michigan Press.
- Mullins, G., & Kiley, M. (2002). It's a PhD, not a Nobel prize. *Studies in Higher Education*, 27, 369–386.
- Murray, R. (2002). *How to write a thesis*. Maidenhead: Open University Press.
- Paltridge, B. (1997). Thesis and dissertation writing: Preparing ESL students for research. *English for Specific Purposes*, 16, 61–70.
- Paltridge, B., & Starfield, S. (2007). *Thesis and dissertation writing in a second language: A handbook for supervisors*. London: Routledge.
- Paltridge, B., & Phakiti, A. (2015). Developing a research project. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 260–278). London: Bloomsbury.
- Phakiti, A., & Paltridge, B. (2015). Approaches and methods in applied linguistics research. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 5–25). London: Bloomsbury.
- Punch, K. (2012). *Developing effective research proposals* (2nd ed.). London: SAGE.

- Rogers, P. M., Zawacki, T. M., & Baker, S. E. (2016). Uncovering challenges and pedagogical complications in dissertation writing and supervisory practices: A multimethod study of doctoral students and advisors. In S. Simpson, N. A. Caplan, & M. Cox (Eds.), *Supporting graduate student writers* (pp. 52–77). Ann Arbor, MI: University of Michigan Press.
- Rudestam, K., & Newton, R. (2007). *Surviving your dissertation*. Thousand Oaks, CA: SAGE.
- Starfield, S., & Paltridge, B. (2014). Generic support for developing a research proposal. In S. Carter & D. Laurs (Eds.), *Developing generic support for doctoral students* (pp. 112–115). London: Routledge.
- Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.
- Tracy, S. J. (2013). *Qualitative research methods: Collecting evidence, crafting analysis, & communicating impact*. Chichester: Wiley-Blackwell.



# 10

## Writing a Research Article

Betty Samraj

### Introduction

As the “principal site for knowledge-making” (Hyland, 2009a, p. 67), the research article has been the focus of numerous discourse studies with two major goals: one, to understand the epistemologies of different disciplines and, the other, to inform the teaching of writing. The prominence granted to the discourse organization of the research article in Swales’ (1990) monograph *Genre Analysis: English in Academic and Research Settings* inspired a plethora of studies on the research article, the structure of its main sections in terms of functional moves and constituent steps, and rhetorical and linguistic features characterizing this genre. Comparisons of this genre across disciplines and languages have contributed to our understanding of disciplinary and cultural values in academic discourse. In addition, researchers have also compared parts of the genre to one another (e.g., results, discussions, and conclusions by Yang and Allison (2003), and abstracts and introductions by Samraj (2005) and the research article to other genres such as textbooks (e.g., Kuhi & Behnam, 2011) in an effort to increase our understanding of this prestigious genre.

At the same time, researchers have also noted “the growing dominance of English as the global medium of academic publications” (Lillis & Curry, 2010,

---

B. Samraj (✉)

Linguistics and Asian/Middle Eastern Languages, San Diego State University,  
San Diego, CA, USA

e-mail: [bsamraj@sdsu.edu](mailto:bsamraj@sdsu.edu)

p. 1) and have pointed out that “the reward systems within which scholars work increasingly ... foreground English-medium publications” (Lillis & Curry, 2010, p. 48). The need for scholars to publish in English-medium journals, especially those who might be considered “off-network” (Belcher, 2007, p. 2), makes English for research publication an urgent issue and has been a stated motivation for studies focusing on this prestige genre.

Since another recent handbook chapter has provided an overview of research on research articles in general (see Samraj, 2016, in the *Routledge Handbook of English for Academic Purposes*), the current chapter will limit its attention to studies on research articles from applied linguistics. In this chapter, I will discuss the findings from genre analyses of applied linguistics research articles, first, attending to analyses of the overall organization of the research article. Following a short review of studies on the macro-structure of applied linguistics research articles, I will consider the organization of the conventional main sections, abstracts, introductions, methods, results, discussions, and conclusions. Second, I will discuss studies reporting on the use of rhetorical features, such as metadiscourse and academic criticism that have been analyzed in applied linguistics research articles. As I discuss these findings, I will point to ways students can apply these findings to their own writing of research articles. The final section of this chapter will provide further ways for novice writers to draw on research findings to shape their own writing in applied linguistics as well as implications for EAP instruction.

## Macro-structure of Research Articles

Most studies on the organization of research articles have focused on the type and order of functional moves and steps in particular sections, such as the introduction and discussion section. A move is defined as a “discoursal or rhetorical unit that performs a coherent communicative function” in written or spoken discourse (Swales, 2004, pp. 228–229), and moves can be realized by one or more steps. Generally, research articles have been assumed to have an Introduction-Method-Results-Discussion (IMRD) structure, although some studies have questioned this assumption (Lin & Evans, 2012). The most comprehensive study of macro-organization and section headings in applied linguistics articles is one by Ruiying and Allison (2004), where they distinguish between primary and secondary applied linguistics research articles. Their analysis of research articles reporting primary research points to variability in the macro-organization of applied linguistics articles, although all contain the three sections, Introduction, Method, and Results. This variability arises from

the presence of nonconventional headings such as *experimental design* in place of the conventional *method* and the use of content headings such as “L2 reading strategies” when reporting results. Ruiying and Allison (2004) also note the presence of additional sections, such as *theoretical basis* and *literature review* between the conventional introduction and methods sections and the section *pedagogic implications* close to the end of the research article. As shown by a later study discussed below (Lin, 2014), the presence of macro-organizations other than the conventional IMRD structure can have an impact on the structure of a conventional section, such as the introduction. Given this, writers should not immediately assume an IMRD structure when writing a research article but consider possible variations that are used by published writers in applied linguistics articles.

Ruiying and Allison (2004) postulate a macro-structure of *introduction*, *argumentation*, and *conclusion* for secondary research articles, which report critical reviews and syntheses of research. They further distinguish three kinds of organization for the argumentation section of such secondary research articles according to their overall purpose: “theory-oriented, pedagogy-oriented, and (pedagogic) application-oriented” argumentations (Ruiying & Allison, 2004, p. 275). For example, the pedagogy-oriented argumentation has a problem-solution or demand-supply pattern in contrast to a point-by-point pattern in the argumentation of a theory-oriented article. Writers preparing review articles then should consider the purpose of their syntheses and determine the structure of the argumentation based on this purpose. See also Li and Wang (Chap. 6) on traditional literature reviews and research syntheses.

## Abstracts

Research article abstracts, long known to be more than an objective summary of the research article, are said to have become more of a stand-alone genre (Gillaerts & Van de Velde, 2010). Santos’s (1996) early study of abstracts from applied linguistics articles postulated five moves to account for this part-genre: (1) situating the research, (2) presenting the research, (3) describing the methodology, (4) summarizing the results, and (5) discussing the research, with moves two and three being obligatory. Santos found that moves one and five, where persuasive work is conducted in situating and justifying the research in terms of previous research and in connecting the results to research or the real world, were the least frequent in the abstracts in comparison to other moves. However, Hyland’s (2000) cross-disciplinary study using a similar framework found the frequencies of these moves in social

science abstracts (including those from applied linguistics) to be much higher than their frequencies in abstracts from the hard sciences, indicating that these moves that connect the study reported to previous research and generalize the findings are important in the discipline of applied linguistics, even if they are less frequent than the other moves in abstracts.

Pho (2008), in a comparison of abstracts from applied linguistics and educational technology, employing Santos' (1996) framework, also revealed that the first move, situating the research, and last move, discussing the research, although less frequent than the other three abstract moves, were more frequent in applied linguistics than in the educational technology abstracts, providing further support for the importance of these moves in applied linguistics academic writing. Melander, Swales, and Fredrickson (1997) found that linguistics abstracts in English produced by Swedes were more likely to exclude introductions and conclusions (somewhat similar in function to moves one and five in Santos' (1996) framework) than those produced by American English writers.

The analyses of abstracts in applied linguistics research articles discussed here indicate that although contextualizing the study and discussing research results may not be obligatory rhetorical functions of abstracts, they play a more important role in applied linguistics abstracts than abstracts from some other disciplines. Authors of applied linguistics research articles should therefore consider including these functional moves which perform more rhetorical work than some of the other more common moves in abstracts.

## Introductions

Introductions in linguistics research articles have been well-studied, also with a focus on cross-disciplinary and cross-linguistic variation. Swales' (1990, 2004) Create A Research Space (CARS) framework is frequently used in these studies. This framework includes three rhetorical moves of establishing a territory, establishing a niche, and presenting the present study. A simple version of the CARS model given in Feak and Swales' (2011, p. 55) volume, *Creating Contexts: Writing Introductions Across Genres*, is given in Table 10.1 below.

An early study of introductions in language studies research articles in Swedish (Fredrickson & Swales, 1994, p. 15) noted the infrequent use of move two, "establish a niche" in the CARS framework. A more recent study (Sheldon, 2011) comparing the structure of introductions in applied linguistics research articles produced in Spanish by native speakers, and in English by native speakers and non-native speakers with Spanish as a first language (L1)

**Table 10.1** Moves in empirical research article introductions**Move 1: Establishing a research territory**

- (a) showing that the general research area is important, central, interesting, problematic, or relevant in some way (optional)
- (b) introducing and reviewing items of previous research in the area (obligatory)

**Move 2: Establishing a niche (citations to previous literature possible)**

- (a) indicating a gap in the previous research
- (b) extending previous knowledge in some way

**Move 3: Presenting the present work (citations to previous literature possible)**

- (a) outlining purposes or stating the nature of the present research (obligatory)
- (b) listing research questions or hypotheses (probable in some fields but rare in others (PISF))
- (c) announcing principal findings (PISF)
- (d) stating the value of the present research (PISF)
- (e) indicating the structure of the research paper (PISF)

produced some complex results. Spanish speakers writing in their L1 were more likely to include this second move than those writing in English as L2 (second language). This finding seems to contradict earlier studies capturing the absence of this move in research article introductions in languages other than English such as Malay research articles in agriculture (Ahmad, 1997), while at the same time indicating that *establish a niche* might be a difficult rhetorical function to perform in EAL, and that student writers might need more practice and help with producing this important move.

Belcher's (2009) study of the use of explicit gap statements, for example, by stating that research in a particular area has been limited, (in contrast to implicit gap statements where academic criticism is hedged) by writers categorized as speakers of English as an international language (EIL) and those considered native English (EL) speakers resulted in the unexpected finding of EIL authors overwhelmingly preferring an explicit gap statement both at the beginning and end of a ten-year period (1996 to 2006) from which the data were gathered. EIL female writers also showed a greater preference for explicit gap statement over their female EL counterparts. Both male and female EL writers grew in their preference for explicit gap statements over this ten-year period. The increasing pressure to publish in competitive journals with low acceptance rates is presented as a possible reason for EIL writers, especially females, adopting what can be construed as a survival strategy (Belcher, 2009, p. 231). The above studies focusing on the *establish a niche* move in applied linguistics research articles should caution us against simple generalizations about the difficulty that any rhetorical move might pose to EAL (or EIL) writers of research articles. However, the increasing frequency in use of explicit gap statements in applied linguistics research article introductions underscores



the need for novice writers to acquire this rhetorical move for publishing success.

Some studies on the structure of introductions have focused on sub-disciplines within applied linguistics. Two sub-disciplines, second language acquisition and second language writing, are the foci of a study by Ozturk (2007), which revealed the variability inherent in an interdisciplinary field such as linguistics. Research article introductions from the journal *Studies in Second Language Acquisition* display the three-move structure of the CARS model much more frequently than the introductions from the *Journal of Second Language Writing*, where introductions manifested a variety of rhetorical organizations. Another study focusing on these same sub-disciplines (Rodriguez, 2009) noted the more frequent use of centrality claims, that is, statements that assert the importance of the topic being explored, in terms of real-world relevance in introductions from second language writing and a preference for centrality claims foregrounding research vigor and interest in second language acquisition research. Novice writers, then, would benefit from exploring the functional organization of research article introductions and the centrality claims employed in the journals to which they wish to submit their manuscripts because of intra-disciplinary variation.

Variations or innovations in introduction structure in applied linguistics have also been identified in research articles where literature reviews are found as sections between introductions and methods, an increasingly common structure in a variety of disciplines (Lin, 2014). The study by Lin (2014) showed that deviations from the conventional IMRD research article structure can have an impact on the rhetorical organization of other part-genres, such as introductions and methods. Two groups of introductions were identified in articles with a subsequent literature review section. One set included introductions with the regular CARS structure while the other nontraditional or *orientation* introduction included a two-move structure, where the first move identified key issues and the second move presented the study, similar to the third move in the CARS model. Lin (2014) reported that these orientation introductions did not contain substantial niche establishment although they contained a sub-move where the value of the research issue was explicated. Since the structure of the literature review was not analyzed in such research articles, it is not clear if niche establishment was more prevalent in that part-genre. What a writer may consider though is that the presence of a separate literature review might alter the shape of the introduction and its persuasive strength.

## Methods

The methods section in research articles remained relatively unexplored after Swales' (1990) first monograph focusing on this genre. However, the last decade has seen an increase in interest in this section following Swales' (2004) discussion of methods as being on a cline with clipped (fast) texts on one end and elaborated (slow) ones on the other, although few studies have been conducted on methods sections in applied linguistics research articles.

In one study, Lim (2011a) reports on analyses of particular steps in the move *delineating sample procedures* found in methods sections of experimental reports from applied linguistics, motivated both by the need to explicate cross-disciplinary variation in organizational structure and his experience with the challenges posed by particular features of methods construction to his L2 writers in Malaysia. Using both qualitative and quantitative analyses, he identified the structure of the steps *describing the sample/participants* and *justifying the sampling procedures* and the grammatical features that experienced writers frequently use with these steps. Because of the limited research on methods sections in applied linguistics research articles, novice writers might compare the characteristics given by Swales (2004) for clipped and elaborated texts against published research articles from their sub-discipline of applied linguistics.

## Results, Discussions, and Conclusions

The most comprehensive analysis of sections that follow methods in research articles is Yang and Allison's (2003) qualitative study that identified the linear and hierarchical structure of these sections while explicating the complex ways in which results, discussions, and conclusions interrelate. The two-level analysis in terms of moves and steps captured differences across these sections that are not just due to the presence of unique moves but also differences in frequencies and development of the same moves in different sections. The primary communicative function of *reporting results* of the results section is seen in the multiple iterations of the move *reporting results* and the relative infrequency of the move *commenting on results*, which is not only obligatory but also more extensively developed in discussion sections. The conclusions section contains three moves, *summarizing the study*, *evaluating the study*, and *deductions from the research*, all of which can also appear in a discussion section. However, the same moves vary in their constituent steps across the two sections. In addition, the value of each move in a section is

impacted by the other moves also present. The focus of the discussion sections is the commentary on specific results while the focus with the conclusion is a more general one on the overall results and an evaluation of the study as a whole (Yang & Allison, 2003).

A number of other studies have focused on just one of these sections that follow the methods section (e.g., Peacock, 2002) or even a move in a particular section such as the *comments on results* move in discussions (Basturkmen, 2009). Following early analyses of the discussion section (e.g., Holmes, 1997; Hopkins & Dudley-Evans, 1988), Peacock (2002) engaged in a multidisciplinary analysis, contrasting native and non-native authors, of what has been identified as a challenging part-genre for novice writers. Language and linguistics was one of the seven disciplines included in this study that identified *finding*, *claim*, and *reference to previous research* as obligatory moves in discussions across disciplines, using a single-level framework of nine moves. Discussion sections from language and linguistics are characterized by more frequent use of the move *reference to previous research*, greater cycling of moves and less frequent use of *recommendations for further research*. In a later study, Dujsik (2013) used Peacock's (2002) framework to analyze discussions from five applied linguistics journals and revealed that most moves in the texts in his corpus occurred with similar frequencies to Peacock's results.

Lim (2010), asserting the need for research on research articles from specific disciplines to prevent overemphasis on some rhetorical features in English for Academic Purposes (EAP) writing instruction, analyzed the move of *commenting on results* in results sections (not discussion sections as in Basturkmen, 2009), where such comment moves are also found. Comparing the structure of the move in education and applied linguistics research articles, Lim (2010) revealed that the steps, *explaining a finding*, *evaluating a finding*, and *comparing a finding with the literature*, were all much more prevalent in the applied linguistics articles in contrast to the education articles, leading him to conclude that education results sections were "comment-stripped" (p. 291).

In another study on the reporting of results, Lim (2011b) again contrasts education and applied linguistics to explore the steps used in the move *paving the way for research findings* (labeled in Yang and Allison's (2003) study as *preparatory information*) as well as the linguistic structures that characterize these steps. Lim's (2011b) study identified four specific steps that pave the way to a report of results: (1) indicating the structure of the result section to be presented, (2) providing background information to the results to be reported, (3) reiterating research questions/purposes, and (4) stating location of data. Interestingly, the first three steps were more common in the applied linguistics articles with mean frequencies at least twice as high as those for the

education articles. The last step of indicating the location of data in tables and graphs had similar frequencies in both sets of data. This study, like others discussed earlier in this chapter, sheds light on the discursive preferences exhibited in applied linguistics research articles, which Lim (2011b, p. 743) refers to as “unpredictable complexity” found in the real world of discourse that is quite different from the idealized discourses often held out as the standard in language teaching.

In a study comparing discussion sections in student-produced theses and published research articles in applied linguistics (specifically language teaching), Basturkmen (2009) focused on the construction of argument in one move in discussions, *commenting on results*. Her fine-grained analysis provides a helpful picture of the construction of elaborate arguments in this move that is part of a common *results-comment* cycle in discussions. Alternative explanations for results, references to literature in support of explanations, and evaluation of explanations contribute to the complex comments move. Student-produced discussions contained the same steps as experienced writers, but the student writers tended to include many more results and compared their results to those in the literature and, importantly, provided far fewer alternative explanations and were less likely to extend their findings to general theory as were the expert writers.

Conclusions in research articles, although not foregrounded in the traditional IMRD structure, have received some attention, especially after Yang and Allison (2003) identified the main rhetorical moves in conclusions (mentioned earlier) and specified their relationship to those that constitute the discussion section. One such study (Amnuai & Wannaruk, 2013) compared the structure of conclusions in applied linguistics research articles produced in international journals and those produced by Thai writers in English journal publications by high-ranking government universities in Thailand using Yang and Allison’s (2003) move framework. This study revealed that 35% of the conclusions in the Thai journals contained just one move. More importantly, only around 20% of the Thai corpus contained the move *evaluating the study* and 45% contained the move *deductions from the research*, significantly lower than the frequencies found in the international corpus. As the authors of this study conclude, non-native and inexperienced writers need to realize the importance of evaluating their studies, contextualizing their findings, and generalizing their research findings in the conclusion.

The results discussed so far in this section point to a number of discursive preferences seen in applied linguistics research articles that novice writers could adopt to produce successful research articles instead of idealized discourse that might be held up in standard language teaching. They could

provide background information and reiterate research questions before reporting their results. Other useful strategies to adopt would be commenting on results, moving from results to generalizations about results, and providing alternative explanations for the results being discussed. More than one study has also revealed the importance of intertextual links to previous research in different moves in both the results and discussion sections, indicating that novice writers should embed their own work in previous research in these sections. Providing evaluations of their studies and pointing to implications from their research in their conclusions might also help novice writers produce successful research articles.

### **Conclusion on Macro-structure of Research Articles**

The current discussion of the rhetorical organization of applied linguistics research articles has indicated that even articles reporting empirical findings may exhibit structures other than the conventional IMRD structure. Studies of the rhetorical structure of sections have revealed the importance of certain rhetorical moves in applied linguistics research articles. Connecting the study being reported to previous literature in the field has been shown to be important in the abstracts (introduction or situating-the-study move) and introductions (establish a niche move) in addition to the discussion section. Furthermore, generalizing from results was also shown to be important in applied linguistics abstracts (Santos, 1996). The analysis of move structure has also highlighted the role of commentary in both the results and discussion sections of applied linguistics research articles (Basturkmen, 2009; Lim, 2010). In fact, complex argumentation can be built in discussion sections through the use of alternative explanations for results and their evaluation in the commentary of the results reported (Basturkmen, 2009). Novice writers need to be mindful of these general features of research articles from applied linguistics while also considering intra-disciplinary variation in writing norms in applied linguistics.

### **Rhetorical Features Characterizing Applied Linguistics Research Articles**

In seeking to write a successful research article, the author has to demonstrate membership in the target disciplinary community not only by producing a text that follows the generic structure in terms of moves and steps valued by

expert members of the community but also by manifesting the sort of author persona valued in this genre in that target disciplinary community. The sort of author persona constructed in academic writing has been explored in a range of studies that can be broadly construed as studies of metadiscourse (e.g., Hyland, 2005; Lorés Sanz, 2008). As Kuhl and Behnam (2011, p. 98) state, “metadiscourse is a principled way to collect under one heading the diverse range of linguistic devices writers use to explicitly organize texts, engage readers, signal their own presence and signal their attitudes to their material.” The findings from the studies discussed below then show novice writers how to manage their authorial presence and their relationships with their readers when writing a research article.

A comparative study of metadiscourse in a number of academic genres in applied linguistics has revealed the nature of metadiscourse that characterizes research articles (Kuhl & Behnam, 2011). Through a detailed and substantive analysis of a range of metadiscourse features, such as evidentials (explicit references to other sources), hedges (e.g., the modal *may*), directives (e.g., *imperatives*), and reader pronouns (e.g., *you*), Kuhl and Behnam (2011, p. 116) show that “language choices reflect the different purposes of writers, the different assumptions they make about their audiences, and the different kinds of interactions they create with their readers.” Their findings showed that the use of evidentials and hedges to indicate deference to the academic community was high in research articles while the use of directives and reader pronouns, which convey an imposition on the reader, was rare. In contrast, the latter set was common in introductory textbooks, a low prestige academic genre. Interpersonal resources such as self-mention and explicit references to other texts were also more valued in research articles than the other academic genres analyzed. Hence, acquiring metadiscoursal norms that would allow an author to engage in “a dialogism that is a manifestation of positive politeness and communality” (Kuhl & Behnam, 2011, p. 121) would be essential for novices to be successful in producing research articles.

A subset of the textual realizations of interpersonal meanings has also been examined in sections of linguistics research articles, such as abstracts and discussions. The use of metadiscourse resources in research article abstracts has been the focus of a number of studies. Lorés Sanz (2008) compared author visibility, which conveys authority and originality, in abstracts and other sections of research articles from three areas in linguistics (English for specific purposes, pragmatics, and general linguistics) through an analysis of the use of the first person pronoun. The author’s voice is shown not to be very strong in abstracts but strongest in the results sections of research articles where his/her contributions to the field are foregrounded. Interestingly, author presence

was found to be more muted in the discussions and conclusions sections where the use of other linguistic features, such as impersonal active constructions and agentless passive constructions, resulted in the construction of a more objective stance.

Interpersonality in applied linguistics research article abstracts was considered from a diachronic perspective in Gillaerts and Van de Velde's (2010) study that analyzed interactional metadiscourse (Hyland, 2005), specifically hedges, boosters, and attitude markers, in texts from 1982 to 2007. The study revealed a drop in use of interactional metadiscourse, particularly due to a drop in boosters (e.g., *clearly*) and attitude markers (e.g., *is misleading*), prompting the authors to speculate whether this showed the move of applied linguistics as a discipline toward the norms held by the hard sciences. In contrast, the use of hedges remained strong in this period, and the authors, in fact, showed a rise in the use of a combination of hedges, boosters, and attitude markers. The combination of features mitigates author stance in abstracts, which Gillaerts and Van de Velde (2010, p. 137) postulate could be due to the increase in size of the applied linguistics discourse community. The continued relative frequency of hedges in applied linguistics abstracts is in line with Kuhi and Behman's (2011) finding about the importance of hedges in applied linguistics research articles.

Applied linguistics research articles in English form a part of the corpus of a multifaceted cross-disciplinary, cross-linguistic study of academic writing called the KIAP project (Fløttum, 2010). Dahl and Fløttum (2011) examined the construction of criticism as part of the KIAP project and focused on research article introductions from linguistics and economics and considered how these criticisms were constructed as authors made new claims within established disciplinary knowledge. They further analyzed each criticism along three dimensions: whether the author was explicitly visible (writer mediated), whether the criticism was specifically directed at an author (personal/impersonal), and whether the criticism was hedged. The total number of instances of criticism was higher in the linguistics introductions than the economics introductions although criticisms were found in a greater number of economic texts. Most criticisms in both disciplines were unhedged and not writer mediated. The linguistics introductions included a larger proportion of criticisms that were author directed, hence, more pointed, than those in the economics texts. Dahl and Fløttum (2011, pp. 273, 278) argue that the main function of the criticism in both disciplines is "showing the uniqueness and originality of the writer's findings" and that authors' "new claims often take the form of a posited difference between established and new knowledge."



Another study that is also part of the KIAP project explored the influence of language and discipline on the construction of author identity and polyphony (or other voices) in research articles (Fløttum, 2010). This study explored the construction of three common author roles (author as researcher, writer, and arguer) constructed in research articles from three disciplines, linguistics, economics, and medicine, in three languages, English, French, and Norwegian. Based mainly on an analysis of the use of the first person pronoun and accompanying verbs, linguistics authors are said to assume all three author roles and to be most “clearly present of the three discipline profiles, as well as the most explicitly argumentative and polemical authors” (Fløttum, 2010, p. 273). In contrast, economics authors are researchers and writers (text guides) and less explicitly argumentative.

In another study using the corpus from the KIAP project, Dahl (2004) analyzed the use of specific kinds of metadiscourse in research articles from the same three disciplines, economics, linguistics, and medicine, and languages, Norwegian, French, and English. Rhetorical metatext, which marks the rhetorical acts in the argumentation analyzed through use of verbs that refer to discourse acts, such as *discuss* and *argue*, was found to the same extent in the linguistics and economics articles in English. The results point to argumentation being much more a part of the knowledge-construction process in these two disciplines than in medicine, where, according to Dahl (2004, p. 1820), the results are said to “reside outside the texts.” Locational metatext, where the author points to the text itself or its component parts, was used more by economics authors than linguistics authors in both English and Norwegian. However, the linguistics texts contained more locational metatext than the medical research articles. Dahl (2004) posits two reasons for the greater use of rhetorical and locational metatext in economics and linguistics: the relative youth of these disciplines and the need for results in these two disciplines to be more subjectively interpreted. A relatively lower use of metadiscourse was found in the French texts for all disciplines and seemed to indicate less author presence and responsibility for argumentation structure and sign posting in research articles no matter the discipline.

These studies on author presence and metadiscourse in research articles in applied linguistics have yielded several key findings. Research articles in applied linguistics on the whole are characterized by the presence of hedges, evidentials, self-mention, and references to other sources, which enable authors to be deferential toward the disciplinary community, and acknowledge the value of previous research while asserting the author’s own place in the community (Kuhi & Behnam, 2011). While the use of some interactional metadiscourse might have decreased over time, the use of hedging has



remained relatively strong in applied linguistics research articles (Gillaerts & Van de Velde, 2010). Although an author's presence in a research article varies across different sections, authorial presence and explicit argumentation as knowledge-construction are valued in applied linguistics writing (Dahl, 2004; Fløttum, 2010; Lorés Sanz, 2008). In addition, academic criticism is not uncommon (Dahl & Fløttum, 2011).

These findings from analyses of metadiscourse and author presence in applied linguistics research articles reveal specific ways in which metadiscourse is important in applied linguistics research articles in English and clearly point to the need for novices and non-native speakers to focus their attention on the linguistic features that construct an appropriate author persona and writer-reader relationship. A simple activity that novice writers in an academic writing class could engage in would be a comparison of applied linguistics research articles from which these features have been removed and unmodified research articles containing these elements of metadiscourse, author presence, and argumentation. Such an activity can focus the novice writer's attention on the functions performed by these discursual features. Junior scholars in applied linguistics could themselves compare sections of the research article such as methods and discussions with those from another discipline for explicit use of criticism, argumentation, hedges, self-mention, and reference to previous research in order to raise their awareness of practices in applied linguistics research articles.

## Implications for Writing a Research Article and Conclusions

Several studies have discussed the growing pressure on EAL writers to publish in English-medium research journals and the challenges they may face (Flowerdew, 2014; Hyland, 2009b). Flowerdew's (2014) chapter on "English for research publication purposes" provides a helpful overview of these issues, including the need for EAL writers to appropriately interpret manuscript reviewer comments, the power relationships between writer and supervisor and writer and editor in academic publication and the roles played by literacy brokers, those other than named authors such as editors and translators, in the publication process (Lillis & Curry, 2010).

One of the early stages in the process of being published in English is learning to write a research article in English. Courses in EAP (or English for specific academic purposes (Flowerdew, 2016)) and English for research

publication processes have benefitted from the findings from discourse studies of the research article both of the macro-structure of various sections of the research article and rhetorical features that characterize this genre. The results of cross-disciplinary studies or those that focus on the unique features of the research article from a particular discipline have underscored the need for EAP courses that acknowledge disciplinary variation in genres and conventions of academic communities. Many of these findings have been transformed into excellent teaching materials in volumes, such as those by Swales and Feak (2000, 2004) and Feak and Swales (2009, 2011), used to familiarize students with the discourse and linguistic tools needed to attain success for writing in various essential social contexts, thereby acculturating them into a variety of target disciplinary communities.

The results of analyses of the research article from the sub-discipline of applied linguistics have not merely identified linguistic features but the communicative functions expressed by organizational structures and linguistic choices, such as the use of the first person pronoun. These results can be employed as data in EAP courses for rhetorical consciousness-raising tasks, an important insight from Swales (1990) and a “fundamental feature of his pedagogic approach,” where students are made aware of the linguistic features of a genre and their connection to communicative functions (Flowerdew, 2015, p. 104; Flowerdew, 2016). The growing use of corpora and computational techniques in EAP research has also had an impact on the teaching of EAP, especially the teaching of writing of the research article. Lee and Swales (2006) and Charles (2014), among others, report on the use of student-built corpora of research articles in advanced EAP courses with students from multiple disciplines.

While the results of the studies on applied linguistics research articles can be used in the design of EAP materials and tasks, with or without electronic corpora and computational tools, in what has been labeled as a pragmatist approach, it might serve us well to remember that Swales (1997, p. 381) refers to his approach to teaching academic writing to advanced students as liberation theology because in his EAP course he seeks to free his students from “consuming attention to the ritualistic surfaces of their texts, ... from dependence on imitation, on formulas, and on cut-and-paste anthologies of other writers’ fragments” among other things. Bearing this in mind, practitioners should develop EAP tasks and materials that promote discovery-based analysis of relevant data that raise their students’ rhetorical consciousness and lead to the writing of successful research articles while ensuring the maintenance of some rhetorical diversity (Mauranen, 1993).

Applied linguistics students not attending an EAP course can use research articles from a preferred applied linguistics journal to explore the disciplinary features discussed in this chapter. Keeping in mind the advice of scholars such as Swales (1997) and Mauranen (1993), I present here some suggestions for novice writers based on aspects of applied linguistics research articles reviewed in this chapter. Those seeking to write research articles in applied linguistics might benefit from considering certain dimensions of this genre in this discipline. They could analyze research articles from journals they are targeting as a venue for their own work in order to answer the questions given in Table 10.2, which focus on text structures. As discussed earlier, a number of rhetorical features have also been analyzed in applied linguistics research articles. The same set of research articles from the journal selected as a publication venue can be used by novice writers to explore the questions provided in Table 10.3, which can focus the writer's attention on a few select rhetorical functions. Seeking to answer the questions given in these two tables would focus writers' attention on some key dimensions to consider when writing a research article without limiting the students' options to merely adopting language choices from published texts. Instead these questions could help a writer maintain

**Table 10.2** Dimensions to consider when constructing a research article

---

**Overall organization:**

1. What should be the main sections of the research article?
2. Should I have a separate literature review between the introduction and methods section?

**Abstract:**

1. Should I have a preliminary move where I connect my research to previous work?
2. Should I end with a move that states the implications of my study?

**Introduction:**

1. Should I include all the moves in the CARS model?
2. Should I provide a gap? How explicit should the gap be?

**Methods:**

1. What features of the clipped and elaborated methods should I include?

**Results:**

1. Should I focus solely on reporting on results?
2. Should I also comment on results by connecting to previous research and by evaluating or explaining a finding?

**Discussion:**

1. How should I comment on results?
2. Should I provide alternate explanations for results and evaluate them?
3. Should I draw on previous literature?

**Conclusion:**

1. After providing a summary of the study, should I evaluate it?
  2. Should I provide deductions from the study?
-

**Table 10.3** Discovering norms for use of metadiscoursal features**Intertextual links:**

1. Where in the research article and how should I refer to previous literature?
2. Should references to other authors be explicit?

**Author presence and strength of claims:**

1. Should my authorial role be explicit in various sections?
2. What discourse functions warrant use of the first person?
3. How should I criticize author claims explicitly?
4. How much hedges, boosters, and attitude markers should I use in making claims to establish new knowledge?

some rhetorical diversity while adhering to genre convention in his/her sub-discipline.

Applied linguists have performed a number of studies on the research article from their own field, which can inform pedagogy as discussed above. These studies seem timely given the growing number of EAL graduate students in the field. Given the diversity in foci within applied linguistics (Kaplan, 2010), a greater number of intra-disciplinary studies on the research article can enhance our understanding of academic conventions in this field. Further, research articles employing quantitative and qualitative methodologies can also be compared to add to this understanding of discourse norms in applied linguistics.

## Resources for Further Reading

Feak, C., & Swales, J. M. (2009). *Telling a research story: Writing a literature review*. Ann Arbor, MI: The University of Michigan Press.

This volume focuses on the literature review and includes valuable information on the use of metadiscourse, taking a stance, and choices in citations in literature reviews.

Feak, C., & Swales, J. M. (2011). *Creating contexts: Writing introductions across genres*. Ann Arbor, MI: The University of Michigan Press.

Introductions from a few academic genres (such as proposals and book reviews) are attended to in this volume, with the most attention paid to research article introductions. The various moves in research article introductions and essential language features are the focus.

Samraj, B. (2016). Research articles. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 403–415). New York, NY: Routledge.

This chapter provides a review of studies on research articles from a variety of disciplines, not just applied linguistics. As such, it captures some of the variation in disciplinary norms manifested in the research article.

Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.

This volume discusses the nature of a number of research genres including the Ph.D. defense and research talks. Especially relevant is the chapter on the research article, which provides a comprehensive discussion of the standard research article, the review article, and short communications.

Swales, J. M., & Feak, C. (2009). *Abstracts and the writing of abstracts*. Ann Arbor, MI: The University of Michigan Press.

This volume, like the others listed below, can be used by instructors or independent researcher-users to teach or learn about the structure of a particular academic genre and the linguistic choices that characterize that genre. This volume focuses on different kinds of abstracts (such as conference abstracts) but pays particular attention to the research article abstract. The carefully constructed tasks will develop the user's rhetorical awareness of the genre and provide practice in producing the genre.

Swales, J. M., & Feak, C. (2011). *Navigating academia: Writing supporting genres*. Ann Arbor, MI: The University of Michigan Press.

This fourth volume in this series perhaps is the least focused on the writing of a research article. However, it does contain useful information regarding communication surrounding the publication process such as responding to reviewer comments. It also includes some information on the author biostatement that accompanies the research article.

## References

Ahmad, U. K. (1997). Research article introductions in Malay: Rhetoric in an emerging research community. In A. Duszak (Ed.), *Culture and styles of academic discourse* (pp. 273–304). Berlin: Mouton de Gruyter.

- Amnuai, W., & Wannaruk, A. (2013). A move-based analysis of the conclusion sections of research articles published in international and Thai journals. 3L; language, linguistics and literature. *The Southeast Asian Journal of English Language Studies*, 19(2), 53–63.
- Basturkmen, H. (2009). Commenting on results in published research articles and masters dissertations in language teaching. *Journal of English for Academic Purposes*, 8(4), 241–251.
- Belcher, D. D. (2007). Seeking acceptance in an English-only research world. *Journal of Second Language Writing*, 16(1), 1–22.
- Belcher, D. D. (2009). How research space is created in a diverse research world. *Journal of Second Language Writing*, 18(4), 221–234.
- Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes*, 35, 30–40.
- Dahl, T. (2004). Textual metadiscourse in research articles: A marker of national culture or of academic discipline? *Journal of Pragmatics*, 36(10), 1807–1825.
- Dahl, T., & Fløttum, K. (2011). Wrong or just different? How existing knowledge is staged to promote new claims in English economics and linguistics articles. In F. Salager-Meyer & B. Lewin (Eds.), *Crossed words: Criticism in scholarly writing* (pp. 259–282). Bern and Berlin: Peter Lang.
- Dujtsik, D. (2013). A genre analysis of research article discussions in applied linguistics. *Language Research*, 49(2), 453–477.
- Feak, C., & Swales, J. M. (2009). *Telling a research story: Writing a literature review*. Ann Arbor, MI: University of Michigan Press.
- Feak, C. B., & Swales, J. M. (2011). *Creating contexts: Writing introductions across genres*. Ann Arbor, MI: University of Michigan Press.
- Fløttum, K. (2010). Linguistically marked cultural identity in research articles. In G. Garzone & J. Archibald (Eds.), *Discourse, identities and roles in specialized communication* (pp. 267–280). Bern: Peter Lang.
- Flowerdew, J. (2014). English for research publication purposes. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 301–321). Chichester: Wiley-Blackwell.
- Flowerdew, J. (2015). John Swales's approach to pedagogy in genre analysis: A perspective from 25 years on. *Journal of English for Academic Purposes*, 19, 102–112.
- Flowerdew, J. (2016). English for specific academic purposes writing: Making the case. *Writing & Pedagogy*, 8(1), 5–32.
- Fredrickson, K., & Swales, J. (1994). Competition and discourse community: Introductions from Nysvenska studier. In B. Gunnarsson, P. Linell, & B. Nordberg (Eds.), *Text and talk in professional contexts*. ASLA: Sweden.
- Gillaerts, P., & Van de Velde, F. (2010). Interactional metadiscourse in research article abstracts. *Journal of English for Academic Purposes*, 9(2), 128–139.
- Holmes, R. (1997). Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purposes*, 16(4), 321–337.

- Hopkins, A., & Dudley-Evans, T. (1988). A genre-based investigation of the discussion sections in articles and dissertations. *English for Specific Purposes*, 7(2), 113–121.
- Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow: Pearson.
- Hyland, K. (2005). *Metadiscourse*. New York, NY: Continuum.
- Hyland, K. (2009a). *Academic discourse: English in a global context*. London: Bloomsbury Publishing.
- Hyland, K. (2009b). English for professional academic purposes: Writing for scholarly publication. In D. Belcher (Ed.), *English for specific purposes in theory and practice* (pp. 83–105). Ann Arbor, MI: University of Michigan Press.
- Kaplan, R. (Ed.). (2010). *The Oxford handbook of applied linguistics* (2nd ed.). New York, NY: Oxford University Press.
- Kuhi, D., & Behnam, B. (2011). Generic variations and metadiscourse use in the writing of applied linguists: A comparative study and preliminary framework. *Written Communication*, 28(1), 97–141.
- Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25(1), 56–75.
- Lillis, T., & Curry, M. J. (2010). *Academic writing in a global context: The politics and practices of publishing in English*. New York: Routledge.
- Lim, J. M. H. (2010). Commenting on research results in applied linguistics and education: A comparative genre-based investigation. *Journal of English for Academic Purposes*, 9(4), 280–294.
- Lim, J. M. H. (2011a). Delineating sampling procedures: Significance of analyzing sampling descriptions and their justifications in TESL experimental research reports. *Ibérica*, 21, 71–92.
- Lim, J. M. H. (2011b). ‘Paving the way for research findings’: Writers’ rhetorical choices in education and applied linguistics. *Discourse Studies*, 13(6), 725–749.
- Lin, L. (2014). Innovations in structuring article introductions: The case of applied linguistics. *Ibérica*, 28, 129–154.
- Lin, L., & Evans, S. (2012). Structural patterns in empirical research articles: A cross-disciplinary study. *English for Specific Purposes*, 31, 150–160.
- Lorés Sanz, R. (2008). Genres in contrast: The exploration of writers’ visibility in research articles and research article abstracts. In S. Burgess & P. Martín-Mártn (Eds.), *English as an additional language and research publication and communication* (pp. 105–123). Bern: Peter Lang.
- Mauranen, A. (1993). Cultural differences in academic discourse—problems of a linguistic and cultural minority. *Afinla-Import*, 23(51), 157–174.
- Melander, B., Swales, J. M., & Fredrickson, K. (1997). Journal abstracts from three academic fields in the United States and Sweden: National or disciplinary proclivities? In A. Duszak (Ed.), *Culture and styles of academic discourse* (pp. 251–272). Berlin: Mouton de Gruyter.



- Ozturk, I. (2007). The textual organisation of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes*, 26(1), 25–38.
- Peacock, M. (2002). Communicative moves in the discussion section of research articles. *System*, 30(4), 479–497.
- Pho, P. D. (2008). Research article abstracts in applied linguistics and educational technology: A study of linguistic realizations of rhetorical structure and authorial stance. *Discourse studies*, 10(2), 231–250.
- Rodriguez, M. (2009). *Applied linguistics research articles: A genre study of sub-disciplinary variation*. Unpublished MA thesis, San Diego State University.
- Ruiying, Y., & Allison, D. (2004). Research articles in applied linguistics: Structures from a functional perspective. *English for Specific Purposes*, 23(3), 264–279.
- Samraj, B. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes*, 24(2), 141–156.
- Samraj, B. (2016). Research articles. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 403–415). New York, NY: Routledge.
- Santos, M. B. D. (1996). The textual organization of research paper abstracts in applied linguistics. *Text*, 16(4), 481–499.
- Sheldon, E. (2011). Rhetorical differences in RA introductions written by English L1 and L2 and Castilian Spanish L1 writers. *Journal of English for Academic Purposes*, 10(4), 238–251.
- Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Swales, J. M. (1997). English as tyrannosaurus rex. *World Englishes*, 16(3), 373–382.
- Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.
- Swales, J. M., & Feak, C. (2000). *English in today's research world: A writing guide*. Ann Arbor, MI: University of Michigan Press.
- Swales, J. M., & Feak, C. (2004). *Academic writing for graduate students: Essential tasks and skills*. Ann Arbor, MI: University of Michigan Press.
- Swales, J. M., & Feak, C. (2009). *Abstracts and the writing of abstracts*. Ann Arbor, MI: University of Michigan Press.
- Yang, R., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes*, 22(4), 365–385.



# Part II

## Research Instruments, Techniques, and Data Sources

There are eight chapters in this part of the Handbook. In each chapter, the authors discuss current and core issues and key research strategies, instruments, and techniques for data gathering. They address challenges and controversial issues, as well as limitations and future directions. The authors also provide examples from sample studies and annotated resources for further reading.

- In Chap. 11 (Interviews and Focus Groups), Matthew Prior discusses the philosophical and methodological background to conducting and evaluating interview and focus group research in applied linguistics. Prior discusses the strengths and limitations of various approaches and deconstructs “commonsense” assumptions associated with interviewing by examining the influence of two prevalent perspectives: “interview as research instrument” and “interviews as social practice.” Prior also gives attention to issues related to rapport, language choice, interculturality, and the “naturalness” of interview data.
- In Chap. 12 (Observation and Fieldnotes), Fiona Copland provides an overview of current approaches to observation with a particular focus on fieldnotes as data. Drawing on a number of recent observational studies in educational and workplace settings, Copland shows how approaches to observation respond to the contextual realities of the research site such as access, relationships, and level of intrusion. The author also discusses the status of fieldnotes as data and how they can be analyzed to develop empirical findings.

- In Chap. 13 (Online Questionnaires), Jean-Marc Dewaele weighs the advantages and disadvantages of data collection through online questionnaires, as opposed to the traditional pen-and-paper method. Online questionnaires allow researchers to reach out to a larger, more diverse, and more motivated pool of potential participants (i.e., not just their own students). The chapter considers these issues and offers practical advice on how to develop questionnaires, how to run them, and how interpret to interpret the findings.
- In Chap. 14 (Psycholinguistic Methods), Sarah Grey and Kaitlyn M. Tagarelli provide insights into the mental representation of language. The authors discuss common psycholinguistic tasks and measures together with their strengths and limitations. They review traditional behavioral measures such as response time as well as newer measures, including eye-tracking and event-related potentials. They also discuss the major experimental paradigms that are employed with psycholinguistic tasks and measures and consider the research conducted with these paradigms.
- In Chap. 15 (“SLA Elicitation Tasks”), Susan Gass deals with two common elicitation procedures in second language acquisition (judgment tasks and elicited imitation). The chapter begins with a discussion of the uses and misuses of judgment data. The major focus is on the practicalities involved in collecting acceptability/grammaticality judgment data, although mention is also made of other types of judgment data (e.g., truth-value judgments) and an alternative to judgment data, namely, magnitude estimation.
- In Chap. 16 (Introspective Verbal Reports: Think-Alouds and Stimulated Recall), Melissa Bowles provides an overview of the use of verbal reports and synthesizes research that has examined their validity, finding them to be valid, if implemented appropriately. The chapter includes a concise guide to the proper use of verbal reports in language research, from data collection to analysis, concluding with a discussion of the method’s limitations and possible triangulation with other data sources.
- In Chap. 17 (Corpus Research methods for Language Teaching and Learning), Magali Paquot offers a critical overview of how corpus-based methods and the subsequent descriptions of language use have been used to inform the publication of reference works (most particularly grammars and learner dictionaries), language teaching syllabus design, and the content of teaching materials (e.g., textbooks). The chapter focuses on how corpus-derived frequency information and (genre-based) lexicogrammatical descriptions can be used to help with decisions about what to teach and how to teach it.

- Finally, in Chap. 18 (Digital Discourses Research and Methods), Christoph A. Hafner provides an overview of the unique affordances and constraints of digital tools and how such tools affect, among other things, the kinds of meanings we can make and the kinds of relationships we can have. Hafner examines important kinds of digital texts and interactions, having regard to (1) how the study of such digital discourses relates to key questions in applied linguistics and (2) how particular digital discourses can be studied, considering issues related to the collection and analysis of data.



# 11

## Interviews and Focus Groups

Matthew T. Prior

### Introduction

Given the person-centered, experiential focus of much applied linguistics research, there is perhaps no investigative activity more widespread than interviewing. Its appeal can be traced in large part to its immediacy and grounding in common sense. After all, when you want to find out what people know, believe, or feel about language-related matters; when you want to document their personal histories and explore their present circumstances; when you want them to comment on hypothetical scenarios, explain their motivations, and describe their imagined futures—you *ask* them. Interviewing, so the logic goes, offers a means to obtain information unavailable to direct observation and to understand individuals in their own words. The interviewer's function, therefore, is to *elicit* and *report* these uncovered facts and first-person perspectives. Or is it?

In this chapter, I reflect upon this “commonsense logic” by examining some of the philosophical and methodological principles and practices of interview<sup>1</sup> and focus group research in applied linguistics. I begin with a brief background to the history of research interviewing, outlining some of the key issues and discussing various approaches along with their associated strengths and challenges. I provide a sample study and an annotated list of suggested resources for further reading.

---

M. T. Prior (✉)

Department of English, Arizona State University, Tempe, AZ, USA

e-mail: [Matthew.Prior@asu.edu](mailto:Matthew.Prior@asu.edu)

## Definitional Matters

Despite its ubiquitous presence in the research domain (and beyond), interviewing is neither self-explanatory nor a neutral transaction (i.e., it is not atheoretical, apolitical, or a-ethical): it is a dynamic interaction requiring a responsible and active interviewer and a responsive and willing interviewee. Sometimes the interview is described as *methodology*, *approach*, *method*, *technique*, or *tool*—terminology that may evoke very distinct (and discipline-specific) conceptual and procedural lenses. Interviews can be designed as stand-alone studies as well as preliminary, follow-up, or complementary components within qualitative, quantitative, or mixed methods projects. In its diversity, interviewing embodies a “complex, multidimensional collection of assumptions and practices” (Gubrium, Holstein, Marvasti, & McKinney, 2012, p. x). As a professional and creative activity, it is a “craft,” a “skill,” and an “art” (Brinkmann & Kvale, 2015; Rubin & Rubin, 2012; Weiss, 1995) that can only be honed through practical, “real-world” experience.

The *institutional frame* of the interview, with its alternating *question-answer format*, distinguishes it from ordinary or spontaneous conversation. Thus, a working definition of the research interview would be: *a prearranged interaction for the specific purposes of gathering and/or generating information, usually recorded, involving participants whose roles or “situated identities” (Zimmerman, 1998; i.e., “interviewer” and “interviewee”) are predetermined.* We must also acknowledge the common *asymmetric participation format* (e.g., the interviewer usually sets the topic and questions; the interviewee is often restricted in the scope of possible or normative responses).

## Historical Development

In recent decades, much has been made of our “interview society” (Atkinson & Silverman, 1997) and its preoccupation with personal inquiry and “generating empirical data about the social world by asking people to talk about their lives” (Holstein & Gubrium, 2004, p. 140). Due to the embeddedness of interviewing and interview-like practices in daily life, most people in the industrialized world are socialized into numerous interview genres (e.g., journalistic interviews, TV talk shows, employment interviews, police questioning, doctor-patient interactions, healthcare surveys, political polls) by the time they reach adulthood (Edley & Litosseliti, 2010).

The quintessential component of the social scientist's toolkit, the research interview, has been described as an "'indigenous' method distinctive to contemporary Western culture" (Atkinson, 2015, p. 95). Although interviewing practices and their concomitant expectations may vary across cultural contexts (Briggs, 1986), the influence of broadcast, print, and digital media in the latter part of the previous century, combined with the modern hyper-connected era, has brought the interview society global (Gubrium & Holstein, 2002). In many countries, foreign language proficiency interviews are a part of educational systems as well as the military, government, and even business (Fulcher & Davidson, 2013). Personal interviews also function as one of the primary gatekeeping procedures for asylum seekers and other migrant groups worldwide (e.g., Maryns, 2012).

I will turn now to the "interviewing society" of academic research by considering two influential traditions: *survey interviewing* and *(auto)biographical interviewing*.

## Two Research Interviewing Traditions

### Survey Interviews

The *face-to-face* or *in-person* interview is perhaps the oldest and most common form of survey data collection (Kropf, 2004). In the survey interview, the interviewee is selected as a representative cross-sectional sample of a target population (e.g., recent migrants, pre-service teachers, motivated students, anxious learners). As with other survey methods, effort is made to control for error by standardizing procedures so respondents receive the same question items or prompts in the same order and manner. This format is advantageous in that it allows the collection of data from a large pool of respondents. Though commonly viewed as a quantitative method, it can incorporate qualitative components, such as extended or open-ended responses (Lavrakas, 2008). See also Currivan (2008) on a modified style of survey interviewing labeled "flexible" or "conversational" interviewing.

As many researchers have pointed out, standardized interviews may "pressure respondents into categories that fail to capture their experiences" (Schaeffer, 1991, p. 369). This problem is referred to as "ecological validity" (i.e., the extent to which the opinions, attitudes, and other responses reflect interviewees in their everyday lives). One way to maximize ecological validity and improve instrument design is to make a regular practice of piloting materials to assess the quality of the questions, protocols, and potential responses (Richards, 2003; Roulston, 2010; Seidman, 2013).

## (Auto)Biographical Interviews

A second tradition in interview research that has had particularly strong uptake in applied linguistics is the elicitation of personal histories and first-person *narratives* (e.g., Barkhuizen, Benson, & Chik, 2014; De Fina & Georgakopoulou, 2012). These are often labeled *autobiographical*, *oral history*, or *life-story* interviews. Storytelling is a basic mode of human interaction; however, interviews to elicit personal stories are relatively recent phenomena. This line of research was led mainly by feminist scholars and historians (Goodwin, 2012) following the “narrative turn” (Chase, 2011) in the mid-1980s.<sup>2</sup> These interviews tend to be in-depth and may be conducted once or over multiple sessions. In coding and analysis, researchers focus on participants’ life trajectories, storied experiences, and key events, paying special attention to the connections and meanings speakers attach to them (Atkinson, 1998; Wengraf, 2001).

*Narrative interview* is a term sometimes used synonymously with autobiographical interviews (cf. Wengraf, 2001, on “biographic-narrative” interviewing), but it generally refers to “small-scale” narrative interviews rather than extended life-story interviews (see Flick, 2014, on “episodic interviews”; also Pavlenko, 2007; Prior, 2016; cf. Riemann & Schütze, 1991, for a distinctive “narrative interview” approach). Miller (2014), for example, illustrates narrative interviewing in action in a case study of adult immigrants in the US. Through a close analysis of speakers’ narrative accounts of learning English and navigating their multilingual repertoires, she attends to their discursive construction of agency and sense-making practices within the interview context.

Although the term *sociolinguistic interview* is occasionally used by researchers in relation to autobiographical interviewing, it is most commonly associated with variationist and other sociolinguistic research seeking to elicit spoken data for analysis. Labov (1972) is well known for using sociolinguistic interviews to collect personal narratives from AAVE speakers for samples of spontaneous, casual (i.e., unmonitored) speech. For consistency with the literature, the term “sociolinguistic interview” is best used in studies where the purpose is *linguistic* analysis.

The *ethnographic interview* label is perhaps the most contentious. In cultural anthropology, ethnographic interviews have long been an integral component of field studies (Skinner, 2012; Spradley, 1979), where the aim is to understand and represent an *emic* or “insider’s” perspective through observation and carefully developed relationships with members of the host community. Though many researchers in applied linguistics and neighboring disciplines often use “ethnographic” as a shorthand description for “qualitative,” others (e.g., Atkinson, 2015; Watson-Gegeo, 1988) contend that a qualitative interview is *not* ethnographic if it does not involve essential components

of ethnographic fieldwork (e.g., extended duration, observation, direct participation). However, see Heyl (2001) for an expanded perspective on ethnographic interviewing informed by feminist, postmodern, and related critical and reflexive perspectives.

## Two Core Issues

In an influential paper on the treatment of qualitative interviews in applied linguistics, Talmy (2010) distinguishes between two prevalent perspectives: “interview as research instrument” and “interview as social practice.” Because these get at the conceptual core of interviewing, they deserve careful consideration here and by all interview researchers.

### Interview as Research Instrument

The *interview as research instrument*, or “data collection” perspective, corresponds to what a number of researchers have variously described as “knowledge collection” or “data mining” (Brinkmann & Kvale, 2015, p. 57), “excavation” (Holstein & Gubrium, 2004, p. 141), and “harvesting psychologically and linguistically interesting responses” (Potter, 2004, p. 206). This aligns with the transmission or “conduit metaphor” (Reddy, 1979) model of communication. An implicit assumption here is that interviewees—provided they are willing, linguistically and developmentally competent (though perhaps with some careful scaffolding and other support by the interviewer)—can provide access to their “internal” or psychological worlds and lived experiences when prompted to do so in the interview situation.

Because the interview as research instrument perspective supports the researcher’s goal of collecting data, it has tended to be the “default” in interview studies throughout applied linguistics. The standardized survey interview cited earlier offers a classic example. This perspective can also be found in investigations of specific constructs (e.g., attitudes, beliefs, autonomy, motivation, identity) as well as narrative and ethnographic research, where researchers are interested in assembling the in-depth and diverse responses of groups and individuals. A shared practice among the various interview studies conducted in this vein is that researchers frequently present their analyses and findings in the form of (often decontextualized) quotes, narratives, shared themes, and other material that “emerged” from the data (see Talmy, 2010, 2011, for an illustrative critique).



## Interview as Social Practice

A contrasting perspective emphasizes the ways in which interviews are “active” (Holstein & Gubrium, 2004) and collaborative accomplishments. This constructionist approach corresponds with what Talmy (2010) refers to as “interview as a social practice” or “data generation.” The key difference with the preceding perspective is that here there is no objective or emergent “truth” or experience waiting to be elicited and collected. To quote Holstein and Gubrium (2004), “respondents are not so much repositories of knowledge—treasuries of information awaiting excavation as they are constructors of knowledge in association with interviewers” (p. 141).

This stance should not be understood as an extreme form of “epistemic relativism” that denies interviewees have actual experiences, perceptions, and so on (cf. Potter & Hepburn, 2008). The fact that interviewees *can* and *do* indeed talk about such things when asked or prompted is precisely the point—but what they produce are *descriptions* (i.e., representations), not objective reports. As discourse and interaction researchers (e.g., Baker, 2002; Potter & Hepburn, 2005, 2012; Prior, 2014, 2016; Roulston, 2010; Talmy & Richards, 2011) have shown, a close inspection of interviews reveals how interactants co-organize their turn-taking patterns, identity categories, power relations, and even the goals and structure of the interview activity itself. When we recognize that interviews are a social practice, we must also acknowledge that the interview *products* cannot be separated from the *processes* by which they are generated. This corresponds with Holstein and Gubrium’s (2004) recommendation to attend to both the “whats” and the “hows” of interviews. Richards (2011) offers an instructional model that demonstrates how sensitivity to both aspects can greatly inform interviewer awareness and training in applied linguistics research.

Regardless of their specific approach to interviewing, researchers have a responsibility to make explicit and reflect on their stances and practices so that they are “purposefully connecting the information [whether “collected” or “generated”] to a conceptual or theoretical base” (Arksey & Knight, 1999, p. 44; see also Talmy, 2010).

## Overview of Interview Formats

There is no “typical” or “ideal” interview format. As with all research, the interviewer must select that which best aligns with the research questions and aims. It is equally important that interviewers consider their own personal communicative style (Roulston, 2010). For example, an introverted

interviewer may prefer a more structured interview. A naturally voluble interviewer may find it difficult *not* to talk too much within a more conversational setting. Frequently it is a process of self-discovery. As Weiss (1995, p. viii) notes, interviewing often involves “trial and error—rather a lot of error—and trial again.”

A single interview may use one or hybrid interview formats, and even the most “generic” interview or well-planned *interview schedule* (i.e., a “guide” or “protocol” with a list of instructions and questions for the interviewee) may need to be modified on the spot to accommodate unexpected circumstances. It is useful, then, to conceptualize interviews as a continuum. At the one end are those with a more standardized or rigid schedule. These tend to employ more close-ended questions (e.g., “Where do you use English?”; “On a scale of 0 to 5, how would you rate your motivation to study Spanish?”) that elicit shorter, more restricted interviewee responses (e.g., “I use it at work”; “Three”). At the other end of the continuum are those that take a less structured, more conversational and exploratory approach. These have questions that are more open-ended (e.g., “Can you tell me about the first time you taught English?”) to elicit extended, in-depth responses.

Although the research literature distinguishes various categories and sub-categories of interviews, in the following section I will focus on four: *structured*, *open-ended*, *semi-structured*, and *focus groups*.

## Structured Interviews

The *structured* or *standardized* interview is most clearly exemplified by the survey interview described earlier. The aim of the fixed-delivery and fixed-response format is to increase reliability, reduce bias, and standardize responses for optimal coding and analysis. This rigidity leaves little room for interviewer or interviewee spontaneity (Dörnyei, 2007). Because it inhibits follow-up probes and expansions (unless they are built into the interview schedule), it risks overlooking topics and concerns that could potentially be relevant and informative to the study. It also discourages in-depth responses by placing interviewees into a more passive respondent rather than active participant role (Foley, 2012). However, the structured interview is particularly advantageous when the researcher is faced with time constraints and when seeking a large response sample. It is also a helpful format for novice interviewers (and interviewees) who require more guidance and structure during the interview process. Because the questions and responses are standardized, the data lend themselves to both quantitative and qualitative analysis.

Nevertheless, no matter how rigorously standardized, these interviews are rarely trouble-free; interviewers must still use careful judgment when stimuli are not effective (e.g., misunderstandings occur, repetition or clarification may be necessary, participants may resist or prove unresponsive). When understanding issues arise in interviews with L2 (second or additional language) speakers, Sapsford (2007, p. 127) points out that the interviewer has three options: (1) reject the respondent (due to insufficient language competence), (2) translate the problematic material (thereby altering the interview questions), and (3) collect the data (now likely to be useless because the questions were not fully understood). Another option may be to incorporate an interpreter as a mediator, but this raises other considerations: including privacy and consent, personal comfort, transformation of the message, and cultural appropriateness.

## Open Interviews

*Open* or *open-ended* interviews fall at the more minimally structured or conversational side of the interview continuum. In this format, the researcher takes up the role of active listener to elicit the interviewee's in-depth perspective. The researcher often has in mind a general topic or themes and has prepared a few broad or "grand tour" (Spradley, 1979) questions, "but many of the specific questions are formulated as the interview proceeds, in response to what the interviewee says" (Rubin & Rubin, 2012, p. 31). Whereas in the structured interview, interviewees are confined to a narrow set of responses to questions and topics predetermined by the interviewer, here the interviewee helps shape the direction and content of the interview.

Many researchers refer to this format as "unstructured," but this is really a misnomer (cf. Richards, 2009), in that *all* interviews (and all interactions, for that matter) have an underlying organizational structure. Moreover, the label "unstructured interview" suggests that less work is required of interviewers than if they were conducting a more "structured" or standardized interview. Regardless of how *conversation-like* an interview may appear on the surface—and no matter how much we, as interviewers, may wish to convince interviewees (and ourselves) that we are just "talking" together—because the interview is both *research process* and *product*, it is never just casual conversation. As any experienced interviewer can attest, all interviews require planning and ongoing management. To make an interview "come off" as conversational requires a great deal of communicative, mental, and emotional effort.

Although the open interview is a popular method, it is highly labor intensive and therefore not an ideal choice when the researcher is under time constraints (Arksey & Knight, 1999). These interviews also produce a large amount of data to transcribe, code, and analyze. With so much data to sort, determining where to begin analysis and what to focus on can pose a significant challenge. Open interviews are often used in case studies and for carrying out initial exploratory work before proceeding with a more focused study (Dörnyei, 2007). Some researchers find that because the open interview allows participants to produce responses in their own words, it is especially useful for investigating sensitive topics. The fact that open interviews are largely interviewee-led may also go some way toward addressing the asymmetric power relations that can be found between interviewer and interviewee (Arksey & Knight, 1999).

## Semi-structured Interviews

*Semi-structured* interviews are considered a subtype of the structured interview genre. Dörnyei (2007) points out that most interviews in applied linguistics are of this type, which offers a compromise between the two extremes of *structured* and *open*. Nevertheless, the precise meaning of “semi-structured” can vary widely from study to study. Many interviewers choose to incorporate structured questions at the beginning of the interview and then follow up with open-ended questions later to expand upon earlier points. Richards (2009) cautions against this approach, warning that it can establish an undesirable pattern of short interviewee responses. Instead, he advises beginning interviews with an open question.

A disadvantage with semi-structured interviews is in comparing participants and generalizing findings due to the lack of standardization of procedures (e.g., study participants are all asked different follow-up questions). As with open interviews, novice researchers may mistakenly assume that they can get by with less preparation. Wengraf (2001) warns, “They are *semi*-structured, but they must be *fully* planned and prepared. Improvisation requires more training and more mental preparation before each interview than simply delivering lines prepared and rote-learned in advance” (p. 5; emphasis in original). Nevertheless, because they are adaptable to almost any research setting, semi-structured interviews remain a widely used interview format.

Regardless of the specific format, all interviews are labor-intensive activities. A genuine interest in people is essential—as is adequate preparation. Interviewing requires flexibility, patience, active listening, a good memory, and strong inter-

personal communication skills to adapt quickly and manage the unpredictability of the interview situation (Brinkmann & Kvale, 2015; Seidman, 2013). Even in open interviews, the interviewer must prepare by determining the interview goals and objectives, recruiting and confirming interviewee participation, setting the meeting place and time, readying and testing the recording equipment, preparing the interview schedule, obtaining informed consent, recording, taking notes, and so forth. A reality is that not all researchers are equally effective in conducting interviews, and not all research participants prove equally willing, able, or consistent interviewees. Miscommunication and resistance are endemic to the activity (due to topic, time of day, length of interview, differential expectations, personality conflict, language and cultural differences, the presence of a recording device, and other factors).

## Conjoint and Group Interviews

When contemplating doing research interviews in applied linguistics, those involving one interviewer and one interviewee are typically what come to mind. But *conjoint interviews* (with two interviewees) and *group interviews* (with three or more interviewees) are also options. The former are common in bilingualism research, for example, when interviewing a couple or a parent and child to investigate language attitudes and family communication practices. Group interviews can be found in studies when the researcher seeks to elicit a range of responses to produce a more holistic picture of a group's or community's (shared and divergent) perspectives on their sociolinguistic experiences, perspectives, resources, practices, and so on. Interview groups can be formed from *pre-existing units* (e.g., classroom, family, workplace, team) or recruited based on specific characteristics directly relevant to the study (e.g., gender, social class, language, country of origin, occupation).

## Focus Group Interviews

Focus group research did not have a visible presence in the social sciences until the 1980s and 1990s (Brinkmann & Kvale, 2015; Morgan, 1997), where it was taken up in feminist research, communication and media studies, sociology, and social psychology (Wilkinson, 2004). In applied linguistics, studies employing focus groups appeared sporadically (e.g., Dushku, 2000; Hyland, 2002; Myers, 1998), with Ho (2006) being one of

the strongest advocates of this method. The focus group received almost no mention in applied linguistics research methods texts until Dörnyei (2007), who highlighted it as a specialized, group format, interviewing technique.

It is important to emphasize that focus groups are not simply group interviews. The focus group is a type of interview setting, usually semi-structured, but the researcher's role here is that of *moderator* or *facilitator* rather than interviewer in the traditional sense. To put focus group participants at ease, the setting is kept as informal and non-directive as possible. As with other interview formats, the moderator prepares an interview schedule, usually consisting of several prompts (e.g., around a topic, issue, open-ended question) designed to spark discussion. The moderator introduces a prompt and invites the focus group participants to enter into discussion. For example, in Parkinson and Crouch's (2011) focus group study on language and cultural identity of mother-tongue Zulu students in South Africa, one of the discussion prompts used was: "OK Let me ask you this thing ... Everyone knows that CSA [program name] only takes students from disadvantaged school ... Carrying that label with you from a disadvantaged school ... Does it impact on how people see you?" (p. 91).

In focus group research, the interaction among the participants is both the method and the data (Kitzinger, 1995). The goal is not for the group to reach a consensus or to respond to the moderator one by one, but to generate discussion (even disagreement) among one another and to bring out members' various viewpoints and experiences (Brinkmann & Kvale, 2015). The moderator is non-directive but active in keeping the discussion flowing, checking and clarifying, making sure that the group is "focused" on the topic, and ensuring all members participate.

Similar to recruitment for other interview studies, focus group participants are generally selected based on predetermined criteria. The size of a focus group varies, usually ranging from 6 to 12 (Dörnyei, 2007; Fern, 2001; Krueger & Casey, 2015). Stewart and Shamdasani (2014) caution that fewer than 8–12 members can lead to an overly narrow discussion; however, smaller groups of 4–8 are increasingly recognized as more effective (Kitzinger, 1995; Liamputtong, 2011), mainly because they are easier to recruit and host and overall more efficient because all participants are more likely to have an opportunity to speak.

As with other kinds of interviewers, focus group moderators require training and practice—perhaps even more so—because they must facilitate interaction among multiple participants at once. In addition to the basic preparation for the setting, essential are general interviewing skills, flexibility, self-control

(particularly to avoid monopolizing or over-directing the discussion), cross-cultural and pragmatic competence, and empathy (Morgan, 1997; Stewart & Shamdasani, 2014; Wilkinson, 2004). Time management is crucial. Participants have been recruited for a specific period (usually one to three hours), and the researcher must keep to that schedule (Stewart & Shamdasani, 2014). The ability to keep a discussion going requires a sensitivity to group dynamics and the ability to enable all members to participate. Sometimes in focus groups there is a tendency for some individuals (e.g., confident, louder, aggressive) to dominate the discussion; thus, their opinions may influence or inhibit those of others. It is also possible that personalities will clash, participants will disagree or fight, the discussion will go off topic, and participants will fragment into sub-groups (Litosseliti, 2007).

There are many benefits to using focus groups. Perhaps one of the main advantages is that they allow the moderator to probe and identify questions, issues, or concerns in an in-depth manner over a short period from a large data sample. The focus group interaction can prompt participants to describe their attitudes, priorities, frames of understanding, norms, values, and other things that may otherwise go unarticulated in other research methods (Kitzinger, 1995). Focus group interviews can be very useful for exploratory studies (Brinkmann & Kvale, 2015), especially as part of a larger research project. Some researchers find it useful to conduct multiple focus groups in a single study (Stewart & Shamdasani, 2014).

Focus group interviews can also be important equalizers. Unlike some other modes of research, the focus group does not discriminate against people who may have limited reading or writing skills. The group dynamic can also be facilitative for those who may be reluctant or unresponsive when interviewed in one-on-one interview settings (Kitzinger, 1995; Litosseliti, 2007).

Because the focus group involves multiple speakers, oftentimes speaking in overlap, it is essential that the discussion be audio or video recorded. This allows the moderator to focus on facilitating the discussion and not worry about writing everything down. The usual practice is to place an audiorecorder in the middle of the table. Participants generally get used to it quickly and ignore it. Video recording allows a much more detailed analysis of speakers and their verbal and nonverbal interaction. Multiple recording devices are recommended (but they should be kept as unobtrusive as possible), both for backup and to ensure all participants' contributions are captured. If the recording is audio only, then the researcher/moderator should transcribe it as soon as possible to be able to distinguish speakers while the discussion is still fresh in their memory.

Like interviews, focus groups have several limitations and concerns. Because focus groups rely on non-probability sampling (i.e., participants are recruited



from convenience samples by the interviewer or self-selected), the findings will not be generalizable to a larger population. There is also the danger that moderators may have biased or manipulated participants' responses, so they must take care to monitor their own verbal and nonverbal feedback and other contributions.

Another consideration is that in academic focus group research, as a matter of reciprocity, researchers generally offer an incentive or some form of compensation for participating (e.g., cash, gift certificate, even a language lesson). However, it can sometimes be awkward for the moderator to dispense the incentive—particularly when trying to establish friendly rapport or if a participant decides to leave the focus group early. It may therefore be helpful to assign another member of the research team to dispense the incentives at the end of the session. In some small-scale studies, such as those frequently conducted by graduate students in applied linguistics, people are often willing to participate for free; but this may raise some ethical questions (Litosseliti, 2007), possibly even of interviewee exploitation (Spradley, 1979).

### Sample Study 11.1

Ro, E., & Jung, H. (2016). Teacher belief in a focus group: A respecification study. *English Teaching, 1*(1), 119–145.

#### Research Problems/Gaps

Research on teacher cognition and teacher beliefs has focused largely on investigating the internal lives of teachers, but little attention has been given to the ways teachers talk about their beliefs with one other. To address this gap, this study takes a discursive approach that examines the concept of “teacher belief,” not as a stable object sitting within people’s heads, but as something that is co-constructed by participants within interaction.

#### Research Method

- *Type of research:* Focus group.
- *Setting and participants:* US university; three male participants were chosen based on their experience as EAP reading instructors and familiarity with the extensive reading approach.
- *Instruments/techniques:* One-time video-recorded, semi-structured focus group interview (100 minutes). A two-minute segment was selected for detailed analysis, and a rationale was provided.
- *Data analysis:* Discourse and interaction analysis (conversation analysis and discursive psychology) of selected sequences (e.g., disagreements, teasing) for insight into participants’ displayed “understandings” and other “psychological” business. Multimodal data were transcribed following modified CA conventions, and analysis included verbal and nonverbal actions. Interview prompts were included in an appendix.



**Key Findings**

Participants' professional beliefs were shown to evolve over the focus group session. Following friendly disagreement and teasing sequences, participants' articulated beliefs became more specific. Participants not only displayed their general beliefs about teaching, they displayed their professional competence and experience as English language teachers and academic reading experts. Findings show how examining speakers' co-construction of teacher beliefs in focus group interaction helps us understand how "collective thinking" shapes teacher learning and teacher practices.

**Comments**

This study demonstrates how interviews can be used to *collect* and *generate* data for analysis. The approach aligns closely with the *interview as social practice* perspective, where the data and meanings are collaboratively constructed. In their analysis and in the detailed transcripts, the researchers included the verbal and nonverbal actions of the participants (including the moderator). The published study was further enriched by the inclusion of video clips of participants, detailed transcripts, and a list of focus group prompts. However, more information on recruitment and a rationale for the small focus group size would have strengthened this study.

## Challenges and Controversial Issues

The "complex, multidimensional" (Gubrium et al., 2012, p. x) nature of interviewing creates many inherent challenges for interviewers and interviewees. I will now consider a few of these matters in more detail.

### Interviews as "Naturalistic" vs. "Contrived"

The matter of interviews as "naturalistic" or "contrived" (i.e., artificial) social occasions remains a point of contention among researchers. Codó (2008) describes the interview as "an authentic communicative situation in which naturally occurring talk is exchanged" (p. 158). Silverman (2014) characterizes interview and focus group studies as "researcher provoked data" (p. 455). Conversation analysts and other interaction researchers have often rejected interviews (but see, e.g., van den Berg, Wetherell, & Houtkoop-Steenstra, 2003), preferring instead to work with *naturally occurring* data<sup>3</sup> that is free of researcher influence (Goodman & Speer, 2016; Potter & Hepburn, 2005). However, Speer (2002) argues, "All data can be natural or contrived depending on what one wants to *do* with them" (p. 520, emphasis in original). Likewise, as many qualitative researchers (e.g., Holstein & Gubrium, 2004; Silverman, 2014) have pointed out, *no* data are ever free of the researcher's

influence, and “all methods have consequences” (Mishler, 1986, p. 120). Thus, although interviews are not “natural” if analyzed as *spontaneous conversation*, they can be considered “natural” as *interviews* (Goodman & Speer, 2016; Speer, 2002). The central issue for “naturalness” may not be the data—or even necessarily the method—but how the data are analyzed, contextualized, and represented.

## Rapport

Rapport is often cited in the literature as essential for prompting interviewees to talk freely and honestly (e.g., Atkinson, 1998; Brinkmann & Kvale, 2015). Interviewers are therefore advised to be open, sympathetic, and interested listeners (e.g., Rubin & Rubin, 2012; Seidman, 2013). They are also cautioned to guard against potential threats to rapport, such as their outsider status (Horowitz, 1986) or their negative verbal and nonverbal responses. Methods texts frequently gloss rapport as “attentive listening and engagement” (Potter & Hepburn, 2012, p. 566; see also Prior, 2017). Though rapport is often considered a kind of emotional closeness, Spradley (1979) argues it “can exist in the absence of fondness and affection” (p. 78). Some researchers have even induced rapport through “faking” friendship (e.g., Duncombe & Jessop, 2012). Because of the often intimate nature of in-depth and repeated interviews, there may also be a concern of “over-rapport” or “over-involvement” between interviewer and interviewee. This has been described as a feature of the general “emotionalist” or “romantic impulse” (Atkinson & Silverman, 1997; Gubrium & Holstein, 2002; Talmy, 2010), where the interviewer attaches so-called authenticity to the speaker’s “voice” and personal experience.

Because interviewers may, intentionally or not, use rapport to further their research aims, it has serious potential for participant coercion, exploitation, and abuse—especially when deception is part of the research design (e.g., when the researcher’s goals and actual interpersonal investment are masked behind the guise of cultivating “friendship” or “just doing conversation”). Because rapport is as much an ethical matter as it is a practical one, great care must be taken to ensure that it is not abused for the sake of getting data.

## Language and Interculturality

In applied linguistics research, interviews often take place between people with different L1s (first languages) and cultural backgrounds. Interactants may therefore be required to use a *lingua franca*, or shared language, to

communicate (e.g., L1-L2, L2-L2). Consequently, unexpected tensions may arise from, on the one hand, essentialist assumptions concerning “cultural differences” between interviewer and interviewee, and on the other, a hybrid or *interculture* that emerges in the interview situation. These matters of language and interculturality—part of the mode of communication and interactants’ collaborative identity work, as well as objects of researcher interest—are areas of study within the intercultural pragmatics paradigm (Kecskes, 2012; see also Kasper & Omori, 2010), and they raise many relevant questions surrounding the construction of meaning in interviews.

Some researchers (e.g., Chen, 2011; Winchitz, 2006) consider the interviewer’s L2 speaker status advantageous because it may encourage L1 interviewees to do more work to define concepts. Prior (2014) found being a cultural “outsider” helpful in eliciting extended narrative accounts from adult immigrant interviewees, although some unexpectedly took over the interviews, turning them into language lessons to instruct the interviewer. Menard-Warwick (2009) notes that despite her privileged social status and Anglo background, her ability and willingness to speak Spanish (her L2) allowed her to establish rapport with adult Latin American immigrant women interviewees in the US.

Pavlenko (2007), in a critical review of L2 autobiographic narrative research, questions the status quo of L2 interviews conducted in the primary language of the interviewer and makes the case that participants should be given an option to use their L1 (which may necessitate using a bilingual interviewer or interpreter; see also Prior, 2016). Code-switching is also an important expressive and symbolic resource in multilingual interviews. Canagarajah’s (2008) study on language shift in the Sri Lankan Tamil diaspora, for example, included fascinating instances of Tamil-English code-switching by the researcher and interview participants. Although language alternation in the interviews went largely unanalyzed, it nonetheless evoked the richness of interculturality as an analytic topic.

Language and interculturality are important issues, especially because they are *constructive of* and *constructed by* the interview interaction itself (see, e.g., Briggs, 1986; Pavlenko, 2007; Prior, 2014; Roulston, 2010). Moreover, ensuring a linguistic, and even cultural, match between interviewer and interviewee (or between moderator and focus group participants) may not always be enough to ensure their comfort and full participation (Miller, 2011). Social status or class, ethnicity, gender, age, skin color, and other personal characteristics are also important to consider, since they may index (rightly or wrongly) various norms, beliefs, competences, and other assumptions that can influence how interactants relate to one another (Chiu & Knight, 1999; Rubin & Rubin, 2012)—even when they are from the *same* language background (Kenney & Akita, 2008).

## Limitations and Future Directions

I have aimed in this chapter to highlight the dynamic nature and rich potential of interviewing in applied linguistics research. Many of the challenges researchers encounter can be addressed by developing a rigorous foundation in research theory and methodology, reading widely (within and outside the field), cultivating research apprenticeships and partnerships, planning carefully, piloting materials, developing reflective and ethical research practices, and most importantly, gaining practical experience.

But interviewing is not always the *right* or the *only* choice for a study. It will not allow the researcher to observe what people *do* in their daily lives. It cannot help predict the behavior of individuals or groups, nor will it reveal what people *really* think, believe, or feel (Brinkmann & Kvale, 2015). However, interviews can show how people articulate responses and meanings as they *describe* their “inner” (i.e., psychological) and “outer” (i.e., social) worlds. As interactions, interviews are rich analytical sources for examining how people do storytelling and use language and other communicative resources in generic as well as creative ways (Kasper & Prior, 2015; Roulston, 2010), but in a bounded context (i.e., with the interviewer or focus group members). Interviewing can also be an important resource for making the researcher aware of potentially researchable issues and concerns that may go unnoticed by other methods.

A growing area where interview researchers are enriching theory and practice is what Rapley (2012) has labeled “social studies of interview studies” (p. 552). Here interviewers return to their data (and the data of others) to examine, for example, research challenges, dilemmas, and “failed” or “uninformative” episodes (e.g., Nairn, Munro, & Smith, 2005; Prior, 2014, 2016; Roulston, 2010, 2011, 2014). Interviewers can use these studies to “reflect on and help them make sense of their own interviewing practice” (Rapley, 2012, p. 548). Reflection becomes a pedagogical tool (Roulston, 2012) or “therapeutic intervention” (Rapley, p. 548) for “refining interviewer technique...developing awareness...[and] improving analytical sensitivity” (Richards, 2011, pp. 108-109). These “therapeutic” benefits for the researcher extend also to studies where interviews are used when they should *not* be (e.g., as stand-ins for direct observation) (S. Talmy, personal communication, January 12, 2017).

Another positive outcome may be to sensitize interviewers to the ignored or taken-for-granted intersections of identity in the research process:

...interviewers must understand the social locations that they occupy as researchers—such as race, ethnicity, status, age, nationality, education, gender, language proficiency, and so forth—and how these may both limit and benefit the generation of interview data with research participants. (Roulston, 2012, p. 71)

This echoes Briggs' (1986) classic work on his "communicative blunders" as an Anglo-American learning to interview Spanish speakers in New Mexico. Roulston (2010) also considers possibilities for a "decolonizing conception of interviewing" (p. 68) that might help counter ways in which indigenous voices often get silenced by non-indigenous researchers (see also Griffin, 2015; Smith, 2012). Close reflection also forces us to recognize the "emotional labor in researcher–respondent interactions and the various ways in which it is incorporated into the research process and writing" (Lillrank, 2012, p. 281; Prior, 2016).

By critically questioning and reflecting upon the various "commonsense" assumptions that shape interview and focus group research in applied linguistics, we will be better able to advance theory and method—ultimately, leading to renewed insights and more rigorously grounded interviewing practices.

## Resources for Further Reading

Briggs, C. L. (1986). *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. New York: Cambridge.

Mishler, E. G. (1986). *Research interviewing: Context and narrative*. Cambridge, MA: Harvard University Press.

These classic works set the standard for conceptualizing the interview as an interactional speech event.

Brinkmann, S., & Kvale, S. (2015). *InterViews: Learning the craft of qualitative research interviewing* (3rd Ed.). London: SAGE.

This text offers a comprehensive and accessible introduction to interview theory and practice, particularly for novice researchers.

Gubrium, J. F., Holstein, J. A., Marvasti, A. B., & McKinney, K.D. (Eds.). (2012). *The SAGE handbook of interview research: The complexity of the craft* (2nd ed.). London: SAGE.

This is the most comprehensive collection representing the diversity of contemporary interview research written by scholars from across the social sciences.

Potter, J., & Hepburn, A. (2005). Qualitative interviews in psychology: Problems and possibilities. *Qualitative Research in Psychology*, 2, 281–307.

Potter, J., & Hepburn, A. (2012). Eight challenges for interview researchers. In J. F. Gubrium, J. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (2nd Ed., pp. 555–570). London: SAGE.

Potter and Hepburn's detailed and provocative critiques of qualitative interview research are regularly cited in the literature and are essential reading for all qualitative researchers.

Richards, K. (2003). *Qualitative inquiry in TESOL [CH2: Interviewing]*. New York: Palgrave Macmillan.

Written for TESOL professionals and graduate students, this chapter presents a step-by-step guide to conducting qualitative interviews.

Roulston, K. (2010). *Reflective interviewing: A guide to theory and practice*. London: SAGE.

Along with its thorough overview of the various theoretical and practical aspects of interviewing, this concise guide offers clear and compelling guidance for becoming a more self-aware interview researcher.

## Notes

1. I am primarily concerned here with face-to-face interviews. For other modes, see, for example, Gubrium et al. (2012).
2. This is a necessarily abbreviated history. Oral history interviews and anthropological interviews can be traced back much further. See, for example, Malinowski (1922) and Ritchie (2011).
3. See Rapley (2007) for another perspective on “naturally occurring” data.

## References

- Arksey, H., & Knight, P. (1999). *Interviewing for social scientists: An introductory resource with examples*. London: SAGE.
- Atkinson, P. (2015). *For ethnography*. London: SAGE.
- Atkinson, P., & Silverman, D. (1997). Kundera's Immortality: The interview society and the invention of the self. *Qualitative Inquiry*, 3(3), 304–325.

- Atkinson, R. (1998). *The life story interview*. London: SAGE.
- Baker, C. D. (2002). Ethnomethodological analyses of interviews. In J. Holstein & J. Gubrium (Eds.), *Inside interviewing: New lenses, new concerns* (pp. 395–412). London: SAGE.
- Barkhuizen, G., Benson, P., & Chik, A. (2014). *Narrative inquiry in language teaching and learning research*. New York: Routledge.
- van den Berg, H., Wetherell, M., & Houtkoop-Steenstra, H. (Eds.). (2003). *Analyzing race talk: Multidisciplinary approaches to the interview*. Cambridge: Cambridge University Press.
- Briggs, C. L. (1986). *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. New York: Cambridge University Press.
- Brinkmann, S., & Kvale, S. (2015). *InterViews: Learning the craft of qualitative research interviewing* (3rd ed.). London: SAGE.
- Canagarajah, A. S. (2008). Language shift and the family: Questions from the Sri Lankan Tamil diaspora. *Journal of Sociolinguistics*, 12(2), 143–176.
- Chase, S. (2011). Narrative inquiry: Still a field in the making. In N. Denzin & Y. Lincoln (Eds.), *The Sage handbook of qualitative research* (4th ed., pp. 421–434). London: SAGE.
- Chen, S.-H. (2011). Power relations between the researcher and the researched: An analysis of native and nonnative ethnographic interviews. *Field Methods*, 23(2), 119–135.
- Chiu, L.-F., & Knight, D. (1999). Developing focus group research: Politics, theory, and practice. In J. Kitzinger & R. Barbour (Eds.), *Developing focus group research: Politics, theory and practice* (pp. 99–112). London: SAGE.
- Codó, E. (2008). Interviews and questionnaires. In L. Wei & M. Moyer (Eds.), *Blackwell guide to research methods in bilingualism and multilingualism* (pp. 158–176). Oxford: Blackwell.
- Currihan, D. B. (2008). Conversational interviewing. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (p. 152). London: SAGE.
- De Fina, A., & Georgakopoulou, A. (2012). *Analyzing narrative: Discourse and sociolinguistic perspectives*. New York: Cambridge University Press.
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.
- Duncombe, J., & Jessop, J. (2012). Doing rapport and the ethics of “faking friendship”. In T. Miller, M. Birch, M. Mauthner, & J. Jessop (Eds.), *Ethics in qualitative research* (2nd ed., pp. 107–122). London: SAGE.
- Dushku, S. (2000). Conducting individual and focus group interviews in research in Albania. *TESOL Quarterly*, 34(4), 763–768.
- Edley, N., & Litosseliti, L. (2010). Contemplating interviews and focus groups. In L. Litosseliti (Ed.), *Research methods in linguistics* (pp. 155–179). London: Continuum.
- Fern, E. F. (2001). *Advanced focus group research*. London: SAGE.
- Flick, U. (2014). *An introduction to qualitative research*. London: SAGE.



- Foley, L. (2012). Constructing the respondent. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The Sage handbook of interview research: The complexity of the craft* (pp. 305–315). London: SAGE.
- Fulcher, G., & Davidson, F. (Eds.). (2013). *The Routledge handbook of language testing*. London: Routledge.
- Goodman, S., & Speer, S. (2016). Natural and contrived data. In C. Tileagă & E. Stokoe (Eds.), *Discursive psychology: Classic and contemporary issues* (pp. 57–69). London: Routledge.
- Goodwin, J. (2012). *Sage biographical research (Vol. 1: Biographical research: Starting points, debates and approaches)*. London: SAGE.
- Griffin, G. (2015). *Cross-cultural interviewing: Feminist experiences and reflections*. London: Routledge.
- Gubrium, J. F., & Holstein, J. A. (2002). From the individual to the interview society. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research* (pp. 3–32). London: SAGE.
- Gubrium, J. F., Holstein, J. A., Marvasti, A. B., & McKinney, K. D. (Eds.). (2012). *The SAGE handbook of interview research: The complexity of the craft* (2nd ed.). London: SAGE.
- Heyl, B. S. (2001). Ethnographic interviewing. In P. Atkinson, A. Coffey, S. Delamont, J. Lofland, & L. Lofland (Eds.), *Handbook of ethnography* (pp. 369–383). London: SAGE.
- Ho, D. (2006). The Focus group interview: Rising to the challenge in qualitative research methodology. *Australian Review of Applied Linguistics*, 29(1), 5.1–5.19.
- Holstein, J. A., & Gubrium, J. F. (2004). The active interview. In D. Silverman (Ed.), *Qualitative research: Theory, method, and practice* (pp. 140–161). London: SAGE.
- Horowitz, R. (1986). Remaining an outsider: Membership as a threat to research rapport. *Urban Life*, 14(4), 409–430.
- Hyland, K. (2002). Directives: Argument and engagement in academic writing. *Applied Linguistics in Academic Writing*, 23(2), 215–239.
- Kasper, G., & Omori, M. (2010). Language and culture. In N. H. Hornberger & S. L. McKay (Eds.), *Sociolinguistics and language education* (pp. 455–491). Bristol: Multilingual Matters.
- Kasper, G., & Prior, M. T. (2015). Analyzing storytelling in TESOL interview research. *TESOL Quarterly*, 49(2), 226–255.
- Kecskes, I. (2012). Interculturality and intercultural pragmatics. In J. Jackson (Ed.), *Routledge handbook of language and intercultural communication* (pp. 67–84). London: Routledge.
- Kenney, R., & Akita, K. (2008). When West writes East: In search of an ethic for cross-cultural interviewing. *Journal of Mass Media Ethics*, 23, 280–295.
- Kitzinger, J. (1995). Introducing focus groups. *British Medical Journal*, 311, 299–311.
- Kropf, M. E. (2004). Survey methods. In J. G. Greer (Ed.), *Public opinion and polling around the world: A historical encyclopedia* (Vol. 1, pp. 468–474). Oxford: ABC-CLIO.



- Krueger, R. A., & Casey, M. A. (2015). *Focus groups: A practical guide for applied research*. London: SAGE.
- Labov, W. (1972). *Language in the inner city: Studies in the Black English Vernacular*. Philadelphia: University of Philadelphia Press.
- Lavrakas, P. (Ed.). (2008). *Encyclopedia of survey methods*. London: SAGE.
- Liamputtong, P. (2011). *Focus group methodology: Principle and practice*. London: SAGE.
- Lillrank, A. (2012). Managing the interviewer self. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (2nd ed., pp. 281–294). London: SAGE.
- Litosseliti, L. (2007). *Using focus groups in research*. London: Continuum.
- Malinowski, B. (1922). *Argonauts of the Western Pacific*. London: Routledge.
- Maryns, K. (2012). Multilingualism in legal settings. In M. Martin-Jones, A. Blackledge, & A. Creese (Eds.), *Routledge handbook of multilingualism* (pp. 297–313). London: Routledge.
- Menard-Warwick, J. (2009). *Gendered identities and immigrant language learning*. Bristol: Multilingual Matters.
- Miller, E. (2011). Indeterminacy and interview research: Co-constructing ambiguity and clarity in interviews with an adult immigrant learner of English. *Applied Linguistics*, 32(1), 43–59.
- Miller, E. R. (2014). *The language of adult immigrants: Agency in the making*. Bristol: Multilingual Matters.
- Mishler, E. G. (1986). *Research interviewing: Context and narrative*. Cambridge, MA: Harvard University Press.
- Morgan, D. L. (1997). *Focus groups as qualitative research*. London: SAGE.
- Myers, G. (1998). Displaying opinions: Topics and disagreement in focus groups. *Language in Society*, 27, 85–111.
- Nairn, K., Munro, J., & Smith, A. B. (2005). A counter-narrative of a “failed” interview. *Qualitative Research*, 5(2), 221–244.
- Parkinson, J., & Crouch, A. (2011). Education, language, and identity amongst students at a South African university. *Journal of Language, Identity, and Education*, 10(2), 83–98.
- Pavlenko, A. (2007). Autobiographical narratives as data in applied linguistics. *Applied Linguistics*, 28(2), 163–188.
- Potter, J. (2004). Discourse analysis as a way of analysing naturally occurring talk. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 200–221). London: SAGE.
- Potter, J., & Hepburn, A. (2005). Qualitative interviews in psychology: Problems and possibilities. *Qualitative Research in Psychology*, 2, 281–307.
- Potter, J., & Hepburn, A. (2008). Discursive constructionism. In J. A. Holstein & J. F. Gubrium (Eds.), *Handbook of constructionist research* (pp. 275–293). New York: Guildford.
- Potter, J., & Hepburn, A. (2012). Eight challenges for interview researchers. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE*

- handbook of interview research: The complexity of the craft* (2nd ed., pp. 555–570). London: SAGE.
- Prior, M. T. (2014). Re-examining alignment in a ‘failed’ L2 autobiographic research interview. *Qualitative Inquiry*, 20(4), 495–508.
- Prior, M. T. (2016). *Emotion and discourse in L2 narrative research*. Bristol: Multilingual Matters.
- Prior, M. T. (2017). Accomplishing “rapport” in qualitative research interviews: Empathic moments in interaction. *Applied Linguistics Review*. <https://doi.org/10.1515/applirev-2017-0029>.
- Rapley, T. J. (2007). *Doing conversation, discourse, and document*. London: SAGE.
- Rapley, T. J. (2012). The (extra)ordinary practices of qualitative interviewing. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (2nd ed., pp. 541–554). London: SAGE.
- Reddy, M. J. (1979). The conduit metaphor—A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought* (pp. 284–310). Cambridge: Cambridge University Press.
- Richards, K. (2003). *Qualitative inquiry in TESOL*. New York: Palgrave Macmillan.
- Richards, K. (2009). Interviews. In J. Heigham & R. A. Croker (Eds.), *Qualitative research in applied linguistics: A practical introduction* (pp. 182–199). New York: Palgrave Macmillan.
- Richards, K. (2011). Using micro-analysis in interviewer training: ‘Continuers’ and interviewer positioning. *Applied Linguistics*, 32(1), 95–112.
- Riemann, G., & Schütze, E. (1991). “Trajectory” as a basic theoretical concept for analyzing suffering and disorderly social processes. In D. R. Maines (Ed.), *Social organization and social process: Essays in honor of Anselm Strauss* (pp. 333–357). New York: de Gruyter.
- Ritchie, D. A. (Ed.). (2011). *The Oxford handbook of oral history*. Oxford: Oxford University press.
- Roulston, K. (2010). *Reflective interviewing: A guide to theory and practice*. London: SAGE.
- Roulston, K. (2011). Working through challenges in doing interview research. *International Journal of Qualitative Methods*, 10(4), 348–366.
- Roulston, K. (2012). The pedagogy of interviewing. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (2nd ed., pp. 61–74). London: SAGE.
- Roulston, K. (2014). Interaction problems in research interviews. *Qualitative Research*, 14(3), 277–293.
- Rubin, H. R., & Rubin, I. S. (2012). *Qualitative interviewing: The art of hearing data* (3rd ed.). London: SAGE.
- Sapsford, R. (2007). *Survey research* (2nd ed.). London: SAGE.
- Schaeffer, N. C. (1991). Conversation with a purpose—Or conversation? Interaction in the standardized interview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A.

- Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 367–391). Hoboken, NJ: John Wiley & Sons.
- Seidman, I. (2013). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (4th ed.). New York: Teachers College Press.
- Silverman, D. (2014). *Interpreting qualitative data* (5th ed.). London: SAGE.
- Skinner, J. (2012). *The interview: An ethnographic approach*. London: Berg.
- Smith, L. T. (2012). *Decolonizing methodologies: Research and indigenous peoples* (2nd ed.). London: Zed Books.
- Speer, S. A. (2002). 'Natural' and 'contrived' data: A sustainable distinction? *Discourse Studies*, 4(4), 511–525.
- Spradley, J. P. (1979). *The ethnographic interview*. New York: Holt, Rinehart, & Winston.
- Stewart, D. W., & Shamdasani, P. N. (2014). *Focus groups: Theory and practice* (3rd ed.). London: SAGE.
- Talmy, S. (2010). Qualitative interviews in applied linguistics: From research instrument to social practice. *ARAL*, 30, 128–148.
- Talmy, S. (2011). The interview as collaborative achievement: Interaction, identity, and ideology in a speech event. *Applied Linguistics*, 32(1), 35–42.
- Talmy, S., & Richards, K. (Eds.). (2011). Qualitative interviews in applied linguistics: Discursive perspectives [Special issue]. *Applied Linguistics*, 32(1), 1–128.
- Watson-Gegeo, K. A. (1988). Ethnography in ESL: Defining the essentials. *TESOL Quarterly*, 22(4), 575–592.
- Weiss, R. S. (1995). *Learning from strangers: The art and method of qualitative interview studies*. New York: The Free Press.
- Wengraf, T. (2001). *Qualitative research interviewing*. London: SAGE.
- Wilkinson, S. (2004). Focus group research. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 177–199). London: SAGE.
- Winchatz, M. R. (2006). Fieldworker or foreigner? Ethnographic interviewing in nonnative languages. *Field Methods*, 18, 83–97.
- Zimmerman, D. H. (1998). Identity, context and interaction. In C. Antaki & S. Widdicombe (Eds.), *Identities in talk* (pp. 87–106). London: SAGE.



# 12

## Observation and Fieldnotes

Fiona Copland

### Introduction

This chapter introduces observation and fieldnotes in applied linguistics research. Traditionally, observation and fieldnotes have been central to fully fledged ethnographies and some applied linguistics studies have taken this approach (e.g., De Costa, 2014; Duff, 2002). However, it is fair to say that full ethnographies are rather rare in this research field: “topical” ethnographies, which focus on a particular aspect of a research site (see Shaw, Copland, & Snell, 2015), are more popular (see, e.g., Creese, Blackledge, & Kaur, 2014; King & Bigelow, 2012; Kubanyiova, 2015; Luk & Lin, 2010; Madsen & Karreboek, 2015). Nevertheless, a recent search of the journal *Applied Linguistics* using the term “fieldnotes” returned only six articles, which rose to eight when “field notes” was the search term, showing that even topical ethnographic studies are not generally reported in one of the leading applied linguistics journals. In contrast, the same search terms returned 38 articles in the *Journal of Sociolinguistics*, a journal which was first published at least 17 years after *Applied Linguistics*, suggesting that researchers taking a more qualitative, ethnographic approach are aligning themselves to sociolinguistics rather than the more mainstream applied linguistics, although for many,

---

F. Copland (✉)

Faculty of Social Sciences, University of Stirling, Stirling, Scotland

e-mail: [fiona.copland@stir.ac.uk](mailto:fiona.copland@stir.ac.uk)

sociolinguistics remains a strand of this larger field (see, e.g., statement by American Association for Applied Linguistics (AAAL) <http://www.aal.org/?page=AboutAAAL>). This observation in turn seems to be confirmed by examining research methods in a recent published issue of *Applied Linguistics* (June 2016, volume 37, issue 3). Out of six articles, two are experimental studies, two corpus linguistics studies and two are theoretically oriented studies of language classification.

Given this underrepresentation of ethnographically oriented studies in mainstream applied linguistics research, the purpose of this chapter is to outline the methodological warrant for ethnographic studies, and in particular to explain how to carry out observation and how to write and analyse fieldnotes. It is hoped that the discussion will contribute to the spirit of dialogue and mutual engagement between research paradigms that King and Mackey (2016) call for in their introductory article for the centenary edition of *The Modern Language Journal* and will go some way to ensuring that researchers are ready to respond to current, real-world research problems in applied linguistics that often require a range of research approaches.

## Current/Core Issues

Applied linguistics is a broad discipline which includes sub-disciplines, such as forensic linguistics, corpus linguistics, second language acquisition (SLA), translation, lexicography and TESOL (teaching English to speakers of other languages) (see Hall, Smith, & Wicaksono, 2011). While not common in all these sub-disciplines, observation has been employed as a data collection tool in translation, SLA and in TESOL. Often the researchers complete either a predetermined observation schedule or one designed by the researcher(s). In TESOL, for example, researchers might collect information about the organisation of learning, such as which students contribute to the class, what kinds of questions a teacher asks or what languages participants use during the lesson. Observation schedules, therefore, tend to guide the observer to particular features of the observed space in a structured way. Researchers may also audio or video record interactions with a view to applying an analytic framework post-observation (see Walsh, 2011).

Another approach to observation, both inside the classroom and in other contexts, is through ethnographic study. Ethnography, simply put, is the study of people in their own contexts with a view to seeing the context and the world as they see it (this is called taking an *emic* perspective). Ethnographers, therefore, generally focus on one (or a small number of) research sites, spend

time in those sites, observing, talking to people (“participants”) and taking part in local practices (this is called “participant observation”) and can take a number of different forms (see Copland & Creese, 2015).

Traditionally, the participants were not known to researchers and studying them meant travelling far from home, living for extended amounts of time with the people being studied and perhaps learning a new language (see e.g., Geertz, 1960; Scheper-Hughes, 2001). However, in recent times, the focus in ethnographic studies has been on understanding groups closer to home, and schools, neighbourhoods and institutions have all been researched. Ethnographers have focused on making the familiar strange (rather than the strange familiar, which was the concern of early ethnography), showing how taken-for-granted routines and behaviours result from and contribute to social structures and ideologies in play. Rock (2015), for example, uses an ethnographic approach to uncover how cautions are delivered to suspects in police custody in the UK and how suspects can be disadvantaged by how the caution is delivered in different contexts. She argues as the result of her findings that the wording of the caution should be reconsidered with a view to making it as transparent as possible.

In a number of cases, the researcher belongs (or has belonged) to the group being studied: this is called “native ethnography” or “practitioner ethnography” (Hammersley, 1992). Through living or working in a context, a researcher (or potential researcher) notices features of the research site that he/she finds puzzling, infuriating or interesting and decides to investigate them. Many PhD studies develop from premises such as these.

Fieldnotes are central to ethnography. The researcher not only observes and takes part in local practices, he/she also writes down what he/she sees. These fieldnotes, therefore, become the main data set from which the researcher draws findings. However, as noted in the Introduction, fieldnotes are not common in much applied linguistics research. This may be because in many applied linguistics sub-disciplines, empirical data seem relatively unassailable, coming as they do in the form of corpora of written texts or recorded conversations, for example. These texts and conversations exist beyond the researcher, who has not brought them into being. Fieldnotes, in contrast, are created by the researcher, often working alone in the field. They record the features that he/she wishes to note, particularly those that seem relevant to the research participants (see below for a detailed discussion). As such, they are clearly subjective. Indeed, if more than one researcher is present at the same event, the fieldnotes may well be very different (Creese, Bhatt, Bhojani, & Martin, 2008). Some applied linguistics researchers, therefore, can struggle with what they might consider the partiality, bias and subjectivity of fieldnotes.

Of course, researchers who work with fieldnotes would refute these criticisms. They argue that all research is subjective and biased and one of the strengths of ethnography is that it acknowledges these realities and focuses on research as interpretation. Fieldnotes provide accounts which record the “partialities and subjectivities that are necessarily part of any interpretative process” (Copland & Creese, 2015, p. 38). Copland and Creese (2015) argue, as a result, that fieldnote data should be open to scrutiny in the same way as other forms of data—concordance lines or transcriptions, for example. In this way, a reader can “formulate his or her own hunches about the perspectives of people who are being studied” (Bryman, 1988, p. 77). In addition, researchers would suggest that fieldnotes are more suitable as a data collection tool for describing the social world than methods such as surveys and even interviews. They allow the researcher to begin to understand and represent the insider’s perspective, providing situated, contextualised accounts of lived realities (see Taylor, 2002).

In the following section, I will outline the processes of writing fieldnotes, before considering core issues in the area currently.

## Essential Processes/Steps

As suggested above, fieldnotes are made by researchers as part of the observation process. However, before observation can begin, the researcher must instigate two important processes: gaining access to the research site and developing relationships with research participants. If the researcher is known in the site and has already developed relationships with the people in it, gaining access can be straightforward and may comprise a formal letter to the head of the site and gaining permission from the potential participants (although this situation may not be without ethical issues—see Copland & Creese, 2016, for a discussion of ethical issues in observational research). The process can be more difficult if the researcher is not known at the site. Potential participants may mistrust the researcher and be suspicious of his/her motives. In this case, the researcher may need to negotiate access with the head of the research site, through explaining the research in detail and developing relationships with the participants. Even then, access may not be granted or granted under certain conditions. Shaw (2015), for example, explains how she gained access to health think tanks through drawing on her experience of working in a well-known think tank in her communications with the executives she wished to interview, and through frequent and repeated contact. In one case, where the executive did not respond, Shaw carefully explained that



the think tank would not feature in the case study and therefore that the think tank's interests would not be represented. This approach was eventually successful and Shaw gained access. However, all the think tank executives insisted on seeing potential publications before going to print which resulted in some negotiation of content. This is quite a concession in any research, although increasingly common when research is funded by commercial companies or those with an overt political or ideological agenda.

The researcher must pay attention at all stages to developing good relationships with the research participants. It is particularly important at the beginning of the observation process so that the participants feel as comfortable as possible with the presence of the researcher. Agar (1996) calls this "early participant observation" and suggests "it is the bedrock on which the rest of your research will rest" (p. 166). It may be best at the beginning of the process not to take notes but to watch and listen in order to concentrate fully on the context without the distraction—for the researcher and the participants—of note taking.

Once the researcher feels comfortable in the site (and the participants feel comfortable having the researcher there), the researcher can begin to make fieldnotes. Fieldnotes are "productions and recordings of the researcher's noticings with the intent of describing the research participants' actions" (Creese, 2011, p. 44). Because fieldnotes are not guided by specific questions or structures (in contrast to observation schedules), the researcher is free to choose what to record. Fieldnotes, therefore, are always incomplete records and are always representative, reducing as they do "just-observed events, persons and places to written accounts" (Emerson, Fretz, & Shaw, 2001, p. 353). Furthermore, they are always evaluative as "whatever we write down, positions us in relation to what we observe in one way or another" (Copland & Creese, 2015, p. 44).

Fieldnotes are constructed accounts over time. When the researcher is observing at the research site, he/she will generally find it difficult to make full notes. Often, it is not possible to make any notes at all in which case the researcher must rely on memory. In other places, the researcher can note only "jottings" (Emerson et al., 2001, p. 356) or "scratch notes" (Sanjek, 1990, p. 95), words or phrases that remind the researcher of something he/she wishes to write about at a later date. It is usual, therefore, for the researcher to ensure that there is time to write up fieldnotes soon after observation, a process which can be both tiring and frustrating (Punch, 2012).

In Sample Study 12.1, you can see how Angela Creese's scratchings became considered, formal fieldnotes. This set of fieldnotes was just one of 42 field-note documents, with each set consisting of between two and nine pages of



### Sample Study 12.1

Creese, A. (2015). Case study one: Reflexivity, voice and representation in linguistic ethnography. In F. Copland & A. Creese with F. Rock & F. Shaw (Eds.), *Linguistic ethnography: Collecting, analysing and presenting data* (pp. 61–88). London: SAGE.

#### Research Background

The extract is from a longer case study which considers how a researcher collects, analyses and presents data, particularly in team ethnography. The author is drawing on a study which she carried out with other researchers which focuses on the links between bilingualism and social class.

#### Research Method

The author uses linguistic ethnography in which she collects both linguistic and ethnographic data, including fieldnotes, recordings and interviews.

- *Setting and participants*: The setting is a complementary school (i.e., a school which operates outside school times in the UK, often on a Saturday, to teach children the language and culture of their parents). Central participants are the teenage students at the school learning their heritage language, and Hema, the head teacher of the school.
- *Instruments/techniques*: Fieldnotes.

#### Key Results

Creese's first set of fieldnotes comprised rough jottings of first drafts made inside the classroom and added to immediately after class. An example of these can be seen in Fig. 12.1 from the original study. The fieldnotes show that at the time Creese was interested in aspiration and the relationship between social class and aspiration. When producing these first "scratchings," Creese would not have known if these particular observations would make an important contribution to research findings or if they would be afforded analytical attention.

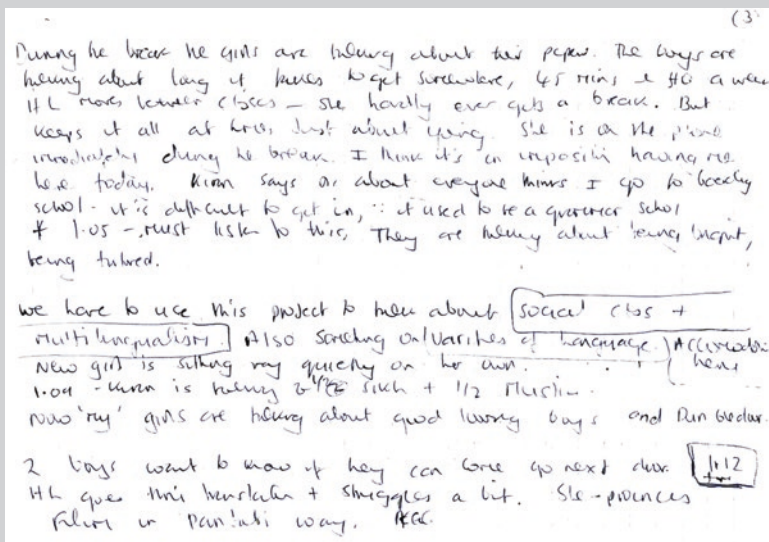


Fig. 12.1 Draft 1 of fieldnotes

Usually Creese's first drafts contain more abbreviations, shorthand text, half-finished sentences and jottings of words to remind her of things of interest to write up in more detail at second draft. These fieldnotes are fuller because the class she was observing was holding an examination and Angela was bored and had more time to write and reflect.

Below you can see how these scratches became a full typed-up set. Although few changes have been made between the original scratchings and these fully fledged fieldnotes, there are some differences, which you may wish to try to spot.

*19 March Fieldnotes (AC)*

*1pm a break is given. During the break the girls are talking about their papers. The boys are talking about money and time! Hema goes into the next class to make sure all is OK. She keeps it all going. She looks at her phone during the break and texts something.*

*\*Kareena says something about everyone thinks she goes to grammar school. She explains it used to be a boarding school and it's difficult to get in cos everyone wants to go there. Should be worth listening to this section of recording as they are all talking about "being bright," "being tutored."*

*My rough notes say we have to use this project to talk about social class and multilingualism. Also something on varieties of language and perhaps "accommodation theory." That should keep us going! New girl is not invited over to join in with the 4 girls who seem to be quite good friends now. She sits on her own. Later they include her.*

*\*Girls are also talking about "a half Sikh and half Muslim girl." Also talking about a good-looking boy and Dumbledore and when the next Harry Potter film is out. The 2 boys want to know if they can go next door but Hema says no, break is over.*

*Hema goes over the translation and appears to struggle a bit—either reading it or translating it. I can't tell. She reads "film" in Panjabi as "filim."*

**Comments**

In redrafting between rough notes and typed-up draft, Angela is making decisions about what is interesting and relevant and what gets carried forward. The first draft, for example, includes a comment about boarding schools which is also included in the second draft. In contrast, she omitted information about researcher imposition, so that her concern that her presence was disrupting the mock examination being conducted by the teacher was not included. In terms of what is added, further detail about grammar schools is included as Angela was interested in the girls' interest in different kinds of schools and she was also interested in the grammar school system herself.

One of the main differences between the two sets of fieldnotes are the phrases, "I write in my notes" and "that should keep us going." The first indicates that Angela would like the meta comment to be noted by others in the team, while the second comment is a light-hearted comment in which she acknowledges her own position as someone is going to be involved in the project for some time.

Whether these fieldnotes become central or not is not the issue at this point. A single entry about one theme in one set of fieldnotes does not constitute data. Rather fieldnotes must be analysed as a set so that themes and rich points can be identified and examined.

single-spaced A4 typed-up observations. As you read, consider how the fieldnotes move from scratchings to formal notes and to the presentation of analytical categories.

## What to Write in Fieldnotes

Writing fieldnotes can be a difficult task, particularly at the beginning of a project when it is difficult to decide what is relevant. Agar (1996) has useful advice in this regard:

At any given time during your early informal fieldwork, there will be a couple of topics you are focusing on. Centre your informal interviews, conversations, and observations on these topics.... When something interesting appears, note it. But don't lose the focus of the topics currently under consideration. (p. 162)

Focusing observations on overarching research questions, therefore, can reduce feelings of being overwhelmed by information and detail, a common response to observing in new research sites. Researchers, as Agar recognises, always have a focus or research questions that guide the study. If these are kept in mind in the beginning stages of fieldwork, they can support the researcher in looking. However, the researcher should be open to noting whatever appears interesting at the research site as new foci can emerge as studies develop.

More detailed advice about writing fieldnotes is provided by Emerson, Fretz, and Shaw (1995):

First jot down details of what you sense are key components of the observed scenes or interactions....Second avoid making statements characterising what people do that rely on generalisations (e.g. using opinionated words)... Third jot down concrete sensory details about actions and talk. (p. 32)

This advice urges the researcher to consider the vocabulary used to record fieldnotes, noting that careless language can affect the quality of the fieldnotes, particularly if the researcher uses "opinionated words" (Emerson et al., 1995, p. 32). Their advice, therefore, is to describe neutrally. Concrete sensory details, in addition, will be helpful in taking the researcher back to the site when reading and analysing the notes, as well as providing useful information when writing up the fieldnotes into research findings.

Richards (2003), drawing on Spradley (1980) provides a useful set of headings which can guide the observer in deciding what to note. Specifically, he suggests that observers record details of setting, people (including relationships

and interactions), systems (by which he means the procedures that participants follow in the research setting) and behaviour (including events and times). The advantage in systematically following a checklist of this nature is that the fieldnotes data can be reliably compared. Furthermore, changes and differences are quickly recognised. Importantly, the list of headings supports the novice researcher in making sense of the research site, which, as noted above, can be overwhelming when first encountered.

Each time a researcher carries out observation, a new set of fieldnotes is created. These fieldnotes become a “corpus” of data (Emerson et al., 2007, p. 353), which is available for analysis (see below). Levon (2013) explains that storing and organising fieldnotes requires some thought and that researchers “should come up with a standardised system for keeping” them (p. 76). Many researchers store their fieldnotes in multiple places, but it is important that these are treated as other research data and kept safely and securely in places which others cannot access.

## Reliability/Validity

For many ethnographic researchers, the notions of reliability and validity are not a concern. Starfield (2015) suggests this is because ethnographers do not want to be held accountable to concepts that have emerged from positivist research paradigms. Indeed, it is rare to see a mention of reliability or validity in texts which focus on ethnographic work. Hammersley (1992) suggests that rather than being concerned with validity and reliability, research of all types should focus on validity and relevance (p. 68). Validity is concerned with producing research findings that can be substantiated. Relevance should be concerned with the importance of the topic and whether the findings add to what we already know about the topic.

Validity answers the question, “How do I know that?”. It requires the researcher to provide explicit evidence to support claims. Validity can be achieved in a number of ways. The first is through convincing the reader that the researcher has carried out a “thorough, systematic iterative analysis” (Duff, 2002, p. 983). To do so, the researcher must provide details of the analytical process and examples of decisions made (see Copland & Creese, 2015, for a number of worked examples of this type). Another way is through providing a detailed description of the research site/participants that is recognisable to those who took part in the research. This can be achieved through member validation, that is, sharing findings with research participants (perhaps through vignettes (further illustrated in Sample Study 12.2)).

Related to this approach, Emerson et al. (2007) suggest that “by presenting herself as a participant in an event and witnessing insider actions, the ethnographer can convince by showing how she learned about a process” (p. 364). In this case, validity is achieved by the weight and authenticity of empirical evidence. A final way is to use more than one data set to provide a range of perspectives on the issue under consideration. While not originally developed with validity in mind, linguistic ethnography (Rampton, Maybin, & Roberts, 2015), which brings together ethnographic and linguistic data and data analysis, provides this affordance. Linguistic ethnography often conjoins fieldnotes with transcribed data from interactions in natural settings or with transcribed interviews, or with both data sets. Taken together, the data can show how a researcher has arrived at findings (see Snell et al., 2015, for a range of studies that take a linguistic ethnographic approach, and Copland & Creese, 2015, for a full discussion of collecting, analysing and presenting data in linguistic ethnography).

Relevance responds to the question, “so what?”. In an age where research impact is increasingly important, relevance cannot be overlooked. Relevance, of course, is a concern for the research project as a whole rather than for fieldnotes in particular. It forces researchers to consider whether the research project has a valuable purpose, in terms of either changing lives or developing theoretical perspectives. In this context, fieldnotes provide researchers with empirical evidence from which to draw findings to make their cases. Their analyses “produce sensitising concepts and models that allow people to see events in new ways” (Hammersley, 1992, p. 15). In that fieldnotes are generally written in a report or narrative form, it could be argued that they are more accessible to a range of readers than statistical data, ensuring that their relevance (or not) is more immediately apparent.

## Analysing Fieldnotes

Writing fieldnotes is the first level of data analysis in that it is an interpretive process (Emerson et al., 1995). As researchers work on their fieldnotes, adding to or refining them, they continue this process of analysis. Formal “observable” analysis begins when the researcher considers the fieldnotes as a data set from which to produce findings. Again, this process requires the researcher to go through a number of stages. First, the researcher must read the notes a number of times in order to become familiar with them. As Blommaert and Dong (2010) advise, “gradually you will start reading them as a source of ‘data’ which you can group, catalogue and convert into preliminary analysis” (p. 39). Often this preliminary analysis will be in the form of themes which begin to emerge from the data. Becoming familiar also means

that researchers will be able to locate particular sections easily for further analysis or publication.

During the reading stage, the researcher will often write analytical notes on the fieldnotes, either by hand or electronically. In Fig. 12.2, the analytical notes can be seen on the right-hand side. These analytical notes help the researcher to develop codes, the next stage in the process. Codes usually combine themes that have informed the research questions with those which emerge through reading and rereading the fieldnotes, often emerging from the notes or memos. For example, in Fig. 12.2, which comes from my data, the codes are given on the left-hand side. “FTA” (face-threatening act) came from a research question, which focused on face. However, the code “feedback structure” was developed as I read the fieldnotes and realised the importance of how feedback was organised to what was said (and allowed to be said).

The main purpose of codes is to provide evidence that themes were either important to participants or important to the researcher (see discussion of reflexivity below). For example, the code FTA appeared in every set of fieldnotes I wrote, often on more than one occasion. In discussing findings, I could be confident that FTAs were common in the feedback conferences I observed and that they deserved to be discussed. Empirically, the frequency of codes can be useful for demonstrating to readers not familiar with ethnography that the theme is important; the researcher, for example, might state that the code FTA appeared 38 times and at least three times in each set of fieldnotes.

Another purpose of coding is to identify the normal and regular from the unusual and noteworthy. It is often the irregular that alerts the researcher to what is taken for granted in the research site and what acceptable behaviour from the research participants’ perspectives looks like (Agar’s, 1996, “rich points”). Hammersley (1992) argues: “ethnographic work often seems to amount to a celebration of the richness and diversity of human social life, but at the same time seeks to identify generic features” (p. 17). Fieldnotes, therefore, provide evidence of both.

At this stage, researchers may produce memos. Drawn from grounded theory (e.g., Charmaz, 2000), memos integrate data from different fieldnotes

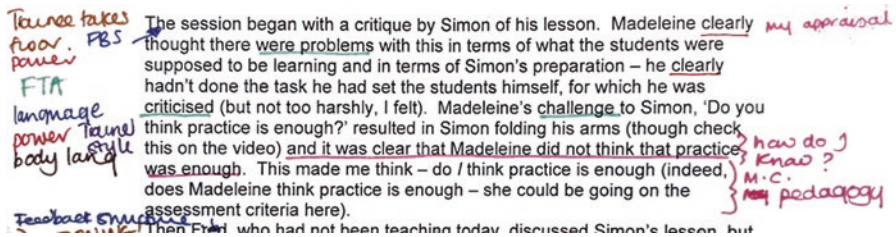


Fig. 12.2 Coded fieldnotes

(and potentially other data sources such as interviews or recorded interactions) into texts that link “analytic themes and categories” (Emerson et al., 1995, p. 143). According to Charmaz (2000), memoing is a process through which we explore and integrate our codes, linking analyses with empirical data. A memo therefore is the researcher’s written articulation of how theoretical insights are drawn from data and may include sections of fieldnotes to illustrate these insights.

Alternatively, or in addition, a researcher may use the observation and fieldnotes to produce vignettes. A vignette is “a focused description of a series of events taken to be representative, typical or emblematic” (Miles & Huberman, 1994, p. 81). Vignettes differ from memos, therefore, in being narrative accounts rather than (only) exploratory or analytical accounts. Their purpose is also different: rather than providing an articulation of the link between data and theory, a vignette attempts to capture something of the experience of being in the research site, including descriptive details of sights and sounds (see Flewitt, 2011). Vignettes, therefore, distil data (rather than including “raw” data such as fieldnotes) into a text which provides readers with an insight into findings of the research. In Sample Study 12.2, the reader can see how Rosie Flewitt achieved this in her vignette of the research participants’ literacy practices.

### Sample Study 12.2

Flewitt, R. (2011). Bringing ethnography to a multimodal investigation of early literacy in a digital age. *Qualitative Research*, 11(3), 293–310.

#### Research Background

The article explores how children develop literacy skills through interaction with technology, specifically the computer.

#### Research Question

In the article, Flewitt asks if ethnography and multimodality can be brought together effectively to provide “richly situated insights into the complexities of early literacy development in a digital age, and can inform socially and culturally sensitive theories of literacy as social practice” (Flewitt, 2001, p. 293).

#### Research Method

- *Setting and participants*: The setting is a pre-school classroom in which there is a computer. The two participants are pre-school children, Edward and Chrissie, who are 3–4 years old, and another young male child of the same age.
- *Instruments/techniques/analysis*: The author uses multimodal methods of data collection and analysis in her work. In this article she explains how ethnography and multimodality can combine effectively to provide insights into early literacy learning.



### Key Results

Flewitt produces a vignette of Edward's literacy play which focuses on an episode of semiosis to demonstrate to the reader how interactions between children and between the computer and children result in meaning making. Flewitt explains that the vignette is a "transduction of data collected in diverse media and presented anew in written format" (p. 302). The vignette, therefore, includes references to interviews, fieldnotes and the researcher's recollections while focusing for the most part on one section of video-recorded activity. Flewitt argues that a vignette cannot reproduce the field but "offers a way to distil observed practices into the portrayal of a scene that captures something of the rich sights, sounds and senses of the ethnographic research experience" (p. 302).

The vignette provides a detailed description of how three children negotiate their social and digital literacy interaction. In summary, Edward is on the computer playing a game and then Chrissie come over to join in, taking the mouse while Edward uses the space bar on the keyboard. After a brief tussle when Chrissie stops the game, the children decide to play a different game which she selects. They play on until a third child joins them and after a further tussle continue to play as a threesome. The full vignette provides a great deal more detail and a photograph supports the description.

### Comments

The value of the vignette is that it demonstrates effectively and efficiently how social practices, in this case developing literacy, are socially situated. In this vignette, the reader develops an understanding of the interactions that took place and of the atmosphere in which they happened. As Flewitt writes, "the reader begins to get a sense of how the interaction between the children was negotiated in ensembles of meaning created through multiple modes, primarily through embodied action, particularly pointing and the sensual use of touch, with some use of spoken language while their gaze remained largely fixed on the screen" (p. 303). Through this vignette, Flewitt clearly shows how ethnography and multimodality can work together to produce nuanced findings which help us to better understand children's literacy development.

## When to Write Fieldnotes

A challenge for researchers is when to write fieldnotes. If he/she is a participant, it is almost impossible to write fieldnotes contemporaneously with the actions being observed. There are other reasons why a researcher may want to postpone fieldnote writing until after the site visit. Participants may get anxious about what is being written. In classroom contexts in particular, observers writing notes are often equated with evaluative practices and so can make teachers uncomfortable. One PhD student of mine avoided this situation by writing his fieldnotes in his car which he parked a kilometre down the road from the school in which he had spent the morning.

Scott Jones and Watt (2010) describe how Scott Jones wrote fieldnotes in the toilet. Her reason for doing so was because her research was covert and she



did not want people in the research site to see what she was doing. In applied linguistics research, covert research is generally discouraged but the issue of when to write notes can provide similar challenges to researchers in different contexts. Likewise, overtly writing fieldnotes can attract attention. Scott Jones and Watt (2010) describe how participants in Watt's research "would often nod to my book as if to say 'you need to write this down'" (p. 165). The public nature of these fieldnotes meant that Watt "developed a strategy of leaving a space after the jotted note to remind her there was more to say on this topic in the privacy of her own space" (2010, p. 165). In this way, Watt could both satisfy the participants that she was attending to what was important to them at the same time as leaving herself some space in which to reflect on what she had seen and perhaps provide an interpretation of it. When to write fieldnotes, therefore, is contextually and motivationally contingent and requires sensitivity on the part of the researcher.

## What to Include in Fieldnotes

In the section above, guidance is given about how to focus fieldnotes and what kind of observations to make. However, content has become something of a controversial issue. Blommaert and Dong (2010) are clear that fieldnotes should include more than a description of what the researcher sees:

Do not attempt to be Cartesian in your fieldnotes: you can afford yourself to be subjective and impressionistic, emotional or poetic. Use the most appropriate way of expressing what you want to express, do not write for an audience, and do not feel constrained by any external pressure: your fieldnotes are private documents, and you will be the only one to decide what you release from them. (p. 38)

Their argument is twofold: first, fieldnotes are private and the researcher chooses what to make public so that choice belongs to the researcher. The second is that the emotional self is involved in the field and researchers should feel free to draw on it as our emotions are intrinsic to how we witness an event. However, the latter view in particular is not universally shared. Punch (2012) explains how in the field of sociology, fieldnotes and field diaries have traditionally been separate documents, with diaries being reserved for feelings, gripes, emotions and fieldnotes for "observations, descriptions of places, events, people and actions...reflections and analytical thoughts about what the observations may mean: emerging ideas, notes on themes and concepts, links to research questions and the wider literature" (p. 90). Drawing on

Blackman (2007), she suggests that this separation is because “incorporating emotions and personal challenges into discussions of the research process is still not accepted as a scholarly endeavour” (p. 87) and therefore should be relegated to a different, less academic document. Indeed, Palmer (2010) advocates separating notes into observational notes, theoretical notes and methodological notes (it is interesting that he does not also suggest “emotional notes”).

In applied linguistics and sociolinguistics research, however, the distinction between the observational self and the emotional self has not been as rigid. As Blommaert and Dong (2010) eloquently argue, how we witness tells a story about “an epistemic process: the way in which we try to make new information understandable for ourselves, using our own interpretive frames, concepts and categories” (p. 37). Writing how we feel, therefore, can support researchers in developing a reflexive approach in their work. Reflexivity involves the recognition that the researcher has an impact on what is being observed in terms of how participants behave, how the research is constructed and what gets reported (see Copland & Creese, 2015). This is not to say that our notes on feelings necessarily find their way into our research reports. Nevertheless, it is vital that we recognise how they influence what we do and see in order that we can account for our findings. Emotions and feelings, therefore, seem central to developing as a reflexive ethnographer (see Davies, 2008).

## Technology and Fieldnotes

In recent years, technology has played an increasingly important role in fieldnotes work. In the research site, tablets can now be used instead of notebooks, with all the electronic benefits of writing and storing at the same time. In terms of analysis, *NVIVO*, *Atlas.ti* and other data analysis programmes help researchers to code and sort fieldnotes data (see, e.g., Tusting, 2015). One of my PhD students has been using *Transana* to bring together classroom recordings, transcriptions of these recordings and fieldnotes. *Transana* provides a visual representation of all the data from a particular observation and has allowed the researcher to identify links between the data which might otherwise have been missed. Figure 12.3 is a screenshot of one such *Transana* page.

It is likely as technology continues to improve and researchers become increasingly adept at using it that much ethnographic data will be stored and analysed using programmes such as these.

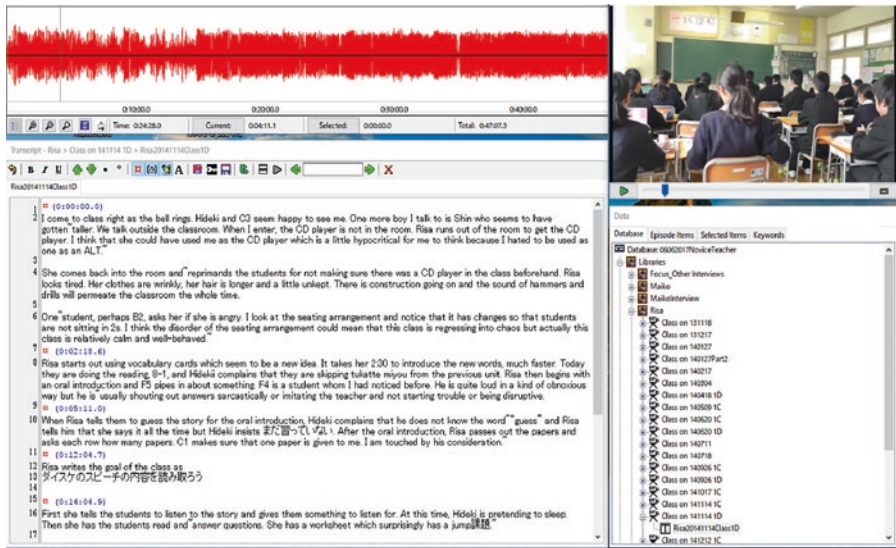


Fig. 12.3 Screenshot of Transana programme used to collate fieldnotes and recordings (Hall, personal data, 2015)

## Limitations and Future Directions

King and Mackey (2016) argue in their discussion of second language acquisition research that it needs to draw on a range of research approaches in order to answer the questions about language learning that are pertinent in today's world. My view is that this argument can be extrapolated to applied linguistics research in general. I can think of few articles I have read recently that would not benefit from taking a close-up look at language in its contexts of use particularly given the multi-layered complexity of many of the issues applied linguists are currently grappling with. While full-blown ethnographies of the kind produced by Duff (2002) are beyond the economic means of many researchers, topic-based ethnographic studies are far less onerous in terms of time and money and therefore within reach. Furthermore, as research funders increasingly request research to be interdisciplinary in nature, it is an opportune time for applied linguistics researchers to consider how an ethnographic element could support them in writing more relevant (in Hammersley's terms) research questions and in producing findings which tackle complexity and which are nuanced and contextualised. The purpose of this chapter is to persuade readers that such an enterprise is worth some investment in terms of understanding how to carry out observational research and how to write and analyse fieldnotes.

## Resources for Further Reading

Copland, F. (2011). Negotiating face in the feedback conference: A linguistic ethnographic approach. *Journal of Pragmatics*, 43(15), 3832–3843.

In this article, I use fieldnotes to develop the argument that linguistic data alone is not adequate in discussions of face threat. The context is English language teacher training but the linguistic ethnographic approach could be taken by anyone carrying out research in a limited number of research sites.

King, K. A., & Mackey, A. (2016). Research methodologies in second language studies: Trends, concerns and new directions. *Modern Language Journal*, 100(S1), 209–227.

In this article, King and Mackey explore current trends in applied linguistics research and suggest the qualitative/quantitative divide should be re-examined with a view to researchers developing research designs which are fit for purpose in terms of investigating complex linguistic problems.

Punch, S. (2012). Hidden struggles of fieldwork: Exploring the role and use of field diaries. *Emotion, Space and Society*, 5, 86–93.

In this article, Sam Punch explores the role of emotion in fieldwork. She describes the traditional approach in sociology to separate fieldnotes from field diaries: the former record observations in the field and the latter emotional responses and other personal responses. She challenges the notion that emotions should not be included in sociological reports and suggests that field diaries should become recognised sources of data.

## References

- Agar, M. H. (1996). *The professional stranger* (2nd ed.). Bingley: Emerald House Publishing.
- Blackman, S. J. (2007). 'Hidden ethnography': Crossing emotional borders in qualitative accounts of young people's lives. *Sociology*, 41(4), 699–716.
- Blommaert, J., & Dong, J. (2010). *Ethnographic fieldwork: A beginner's guide*. Bristol: Multilingual Matters.
- Bryman, A. (1988). *Quantity and quality in social research*. London: Unwin Hyman.

- Charmaz, K. (2000). Grounded theory: Objectivist and constructive methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 509–536). London: SAGE.
- Copland, F. (2011). Negotiating face in the feedback conference: A linguistic ethnographic approach. *Journal of Pragmatics*, 43(15), 3832–3843.
- Copland, F., & Creese, A. with Shaw, S., & Rock, F. (2015). *Linguistic ethnography: Collecting, analysing and presenting data*. London: SAGE.
- Copland, F., & Creese, A. (2016). Ethical issues in linguistic ethnography: Balancing the micro and the macro. In P. I. De Costa (Ed.), *Ethics in applied linguistics research* (pp. 161–178). London: Routledge.
- Creese, A. (2011). Making local practices globally relevant in researching multilingual education. In F. M. Hult & K. A. King (Eds.), *Educational linguistics in practice: Applying the local globally and the global locally* (pp. 41–55). Bristol: Multilingual Matters.
- Creese, A. (2015). Case study one: Reflexivity, voice and representation in linguistic ethnography. In F. Copland & A. Creese with F. Rock & F. Shaw (Eds.), *Linguistic ethnography: Collecting, analysing and presenting data* (pp. 61–88). London: SAGE.
- Creese, A., Bhatt, A., Bhojani, N., & Martin, P. (2008). Fieldnotes in team ethnography: Researching complementary schools. *Qualitative Research*, 8(2), 197–215.
- Creese, A., Blackledge, A., & Kaur, J. (2014). The ideal ‘native-speaker’ teacher: Negotiating authenticity and legitimacy in the language classroom. *Modern Language Journal*, 98(4), 937–951.
- Davies, C. A. (2008). *Reflexive ethnography: A guide to researching selves and others* (2nd ed.). Oxford: Routledge.
- De Costa, P. I. (2014). Making ethical decisions in an ethnographic study. *TESOL Quarterly*, 48(2), 413–422.
- Duff, P. (2002). The discursive co-construction of knowledge, identity, and difference: An ethnography of communication in the high school mainstream. *Applied Linguistics*, 23(3), 289–322.
- Emerson, R. M., Fretz, R. I., & Shaw, L. L. (1995). *Writing ethnographic fieldnotes*. Chicago: University of Chicago Press.
- Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2001). Participant observation and fieldnotes. In P. Atkinson, A. Coffey, S. Delamont, J. Lofland, & L. Lofland (Eds.), *Handbook of ethnography* (pp. 352–368). London: SAGE.
- Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2007). Participant observation and fieldnotes. In P. Atkinson, A. Coffey, S. Delamont, J. Lofland, & L. Lofland (Eds.), *Handbook of ethnography* (pp. 352–368). London: Sage.
- Flewitt, R. (2011). Bringing ethnography to a multimodal investigation of early literacy in a digital age. *Qualitative Research*, 11(3), 293–310.

- Geertz, C. (1960). *The religion of Java*. New York: Free Press.
- Hall, C., Smith, P. H., & Wicaksono, R. (2011). *Mapping applied linguistics*. London: Routledge.
- Hammersley, M. (1992). *What's wrong with ethnography?* London: Routledge.
- King, K. A., & Bigelow, M. (2012). Acquiring English and literacy while learning to do school: Resistance and accommodation. In P. Vinogradov & M. Bigelow (Eds.), *Low education second language and literacy acquisition, proceedings from the 7th symposium* (pp. 157–182). Minneapolis: University of Minnesota Press.
- King, K. A., & Mackey, A. (2016). Research methodologies in second language studies: Trends, concerns and new directions. *Modern Language Journal*, 100(S1), 209–227.
- Kubanyiova, M. (2015). The role of teachers' future self guides in creating L2 development opportunities in teacher-led classroom discourse: Reclaiming the relevance of language teacher cognition. *Modern Language Journal*, 99(3), 565–584.
- Levon, E. (2013). Ethnographic fieldwork. In C. Mallinson, B. Childs, & G. Herk (Eds.), *Data collection in sociolinguistics: Methods and applications* (pp. 69–79). London: Routledge.
- Luk, J. C. M., & Lin, A. (2010). *Classroom interactions as cross-cultural encounters: Native speakers in EFL lessons*. London: Routledge.
- Madsen, L., & Karreboek, M. (2015). Hip hop, education and polycentricity. In J. Snell, S. Shaw, & F. Copland (Eds.), *Linguistic ethnography: Interdisciplinary explorations* (pp. 246–265). London: Palgrave Macmillan.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). London: SAGE.
- Palmer, C. (2010). Observing with a focus: Fieldnotes and data recording. In J. Scott Jones & S. Watt (Eds.), *Ethnography in social science practice* (pp. 141–156). London: Routledge.
- Punch, S. (2012). Hidden struggles of fieldwork: Exploring the role and use and of field diaries. *Emotion, Space and Society*, 3(2), 86–93.
- Rampton, B., Maybin, J., & Roberts, C. (2015). Theory and method in linguistic ethnography. In J. Snell, S. Shaw, & F. Copland (Eds.), *Linguistic ethnography: Interdisciplinary explorations* (pp. 14–50). London: Palgrave Macmillan.
- Richards, K. (2003). *Qualitative inquiry in TESOL*. Basingstoke: Palgrave Macmillan.
- Rock, F. (2015). Case study three: Ethnography and the workplace. In F. Copland & A. Creese (Eds.), *Linguistic ethnography: Collecting, analysing and presenting data* (pp. 117–142). London: Sage.
- Sanjek, R. (1990). A vocabulary for fieldnotes. In R. Sanjek (Ed.), *Fieldnotes: The makings of anthropology* (pp. 92–138). Ithaca and London: Cornell University Press.
- Scheper-Hughes, N. (2001). *Saints, scholars, and schizophrenics* (20th anniversary edition). London: University of California Press.

- Scott Jones, J., & Watt, S. (2010). Making sense of it all: Analysing ethnographic data. In J. Scott Jones & S. Watt (Eds.), *Ethnography in social science practice* (pp. 157–172). London: Routledge.
- Shaw, S. (2015). Case study four: Ethnography, language and healthcare planning. In F. Copland & A. Creese (Eds.), *Linguistic ethnography: Collecting, analysing and presenting data* (pp. 143–170). London: Sage.
- Shaw, S., Copland, F., & Snell, J. (2015). An introduction to linguistic ethnography: Interdisciplinary explorations. In J. Snell, S. Shaw, & F. Copland (Eds.), *Linguistic ethnography: Interdisciplinary Explorations* (pp. 1–13). London: Palgrave Macmillan.
- Snell, J., Shaw, S., & Copland, F. (2015). *Linguistic ethnography: Interdisciplinary explorations*. London: Sage.
- Spradley, J. P. (1980). *Participant observation*. New York: Holt, Rinehart and Winston.
- Starfield, S. (2015). Ethnographic research. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics* (pp. 137–152). London: Bloomsbury.
- Taylor, S. (2002). Researching the social: An introduction to ethnographic research. In S. Taylor (Ed.), *Ethnographic research: A reader* (pp. 1–12). London: The Open University and SAGE.
- Tusting, K. (2015). Workplace literacies and workplace society. In J. Snell, S. Shaw, & F. Copland (Eds.), *Linguistic ethnography: Interdisciplinary explorations* (pp. 51–70). London: Palgrave Macmillan.
- Walsh, S. (2011). *Exploring classroom discourse: Language in action*. Abingdon: Routledge.





# 13

## Online Questionnaires

Jean-Marc Dewaele

### Introduction

This chapter considers issues linked to questionnaires and to web-based research, more specifically within second language acquisition (SLA) and multilingualism research. Referring to my own experiences with online questionnaires, I consider the extent to which sampling and the increased heterogeneity of samples can affect findings. I also address the question of sample size and of the various pitfalls that researchers can tumble into. I mention issues of validity, suggest future directions for research and remind readers of some important issues of etiquette, such as gratitude to participants and respect with regard to ethical issues. The present chapter takes Wilson and Dewaele's (2010) study of the use of web questionnaires in SLA and bilingualism research as a starting point.

Dörnyei and Taguchi (2010) explain that because the essence of scientific research “is trying to find answers to questions in a systematic manner, it is no wonder that the questionnaire has become one of the most popular research instruments in the social sciences” (p. 1). They define “questionnaire” as “the self-completed, written questionnaire that respondents fill in by themselves” (p. 3). Questionnaires are employed “as research instruments for measurement purposes to collect valid and reliable data” (p. 3). The authors do lament the

---

J.-M. Dewaele (✉)

Department of Applied Linguistics and Communication, Birkbeck, University of London, London, UK

e-mail: [j.dewaele@bbk.ac.uk](mailto:j.dewaele@bbk.ac.uk)



fact that, despite a growing methodological awareness in applied linguistics, “the practice of questionnaire design/use has remained largely uninformed by the principles of survey research accumulated in the social sciences” (p. 2).

Applied linguists have been using questionnaires—both paper-based and online versions—for many years. Paper-based questionnaires also have broad applications for other types of research designs including experimental research, case study, action research and ethnographic research. Paper-based questionnaires guarantee a higher response rate—but not necessarily a higher quality of response—when the researcher is on site. Though there are plenty of publications on paper-based questionnaires (Dörnyei & Taguchi, 2010), where online questionnaires might get a passing mention, there is a distinct lack of texts focused more specifically on online questionnaires, a gap which merits to be filled.

The omnipresence of the Internet has greatly facilitated the development of web-based instruments for researchers at all levels: ranging from undergraduate students collecting data for a small research project to Masters’ students collecting information for their dissertation, to PhD students collecting substantial quantities of quantitative and/or qualitative data for their thesis, or established researchers reaching out to the general public in order to collect data that could allow them to answer intriguing research questions.

The danger is that it has become almost too easy to create an online questionnaire, and there is a danger that (budding) researchers may underestimate the amount of work and thought that must go into the research design, into the criteria for participation, into the formulation of the questions and the items, and the balance between conciseness and completeness, while keeping an eye on the overall length of the questionnaire and countless other things that could make a big difference at the time of the analysis of the feedback (Dörnyei & Taguchi, 2010).

The reactions in the field of second language acquisition (SLA) have been sufficiently positive for the use of online questionnaires to become firmly established (Dörnyei, 2007). Dörnyei points to the advantages and disadvantages of web-based research. The most important advantage that outweighs the disadvantages by far is the low cost of setting up an online questionnaire. One popular platform, *SurveyMonkey*, is free but only ten questions can be included in the questionnaire and there is a cap of 100 responses. For a small monthly fee, it allows up to 1000 responses, and for a slightly higher annual amount the number of responses is unlimited and statistical assistance is included (<https://www.surveymonkey.com>). *Google Forms* are free but offer fewer services (<https://docs.google.com/forms/u/0/>). A growing number of academic institutions have purchased an online authoring tool/survey tool

licence that academic staff, students and its employees can use for free. A survey online authoring tool, such as *Lime* is common and user friendly and a URL for the survey can be generated and hosted by an academic institute (<https://www.limesurvey.org/>). Researchers in applied linguistics and psychology have also started using *Mechanical Turk* to gain immediate (online) access to and data from large numbers of participants (<https://www.mturk.com/>).

Once online questionnaires are up, they run themselves without further input from the researcher, keeping count of the number of participants, until the point where sufficient data have been collected and the harvest can start. This amounts to downloading the data into a spreadsheet.

Another advantage, according to Dörnyei (2007) is anonymity, as there is no face-to-face interaction between researcher and participants, and no pressure on the latter to participate, which enhances the level of honesty in responses. Indeed, because there are no social consequences to participation, there is less chance that participants may be tempted to exaggerate or distort their responses to please the authors of the questionnaire. Dörnyei points out that another crucial advantage is the potential access to much larger and more diverse populations all over the world, or, on the other hand, the possibility to reach “small, scattered, or specialised populations” (p. 121). This is particularly important for applied linguists who try to collect data from learners with different language backgrounds engaged in the acquisition of a variety of target languages in different school systems around the world. The diversity of contexts boosts the ecological validity of the data, in other words, the richness of the materials and settings approximate the real world. Patterns between variables that are unique to one context, and that researchers may interpret as being universal, may well be absent in another context. The multi-context perspective thus protects researchers from prematurely jumping to sweeping conclusions about the phenomena they are interested in.

One major limitation of questionnaires, and online questionnaires in particular, which will be highlighted further, is the inevitable self-selection bias. Indeed, in “Internet-based research (...) it is not possible to apply a systematic, purposive sampling strategy, as all participants are self-selected” (Dörnyei, 2007, p. 122). Even with paper-based questionnaires, potential participants can decline to fill out the questionnaire or participate without much enthusiasm and leave some questions unanswered or start to answer at random.

Several of the studies that I have carried out using online questionnaires were based on data collected from between 1000 and over 2000 participants (see Sample Study 13.1). These numbers would have been unattainable to a lone researcher in the past or to a researcher using paper-based questionnaires. I remember the days when I handed out printed questionnaires to my students,

or to colleagues, and hovered around to make sure they completed the questionnaire and did not promise to hand it to me the following week—which was quite unlikely to occur. It took a lot of time and persuasive effort, and getting 150 properly completed questionnaires was an achievement. Then came the tedious transcription of Likert scale responses onto Excel files and the struggle to decipher bad handwriting on open questions. These days, the data can be imported straight into an Excel file, and while it does typically take a full day to clean up the data, the effort required for preparing the data of 1500 participants for statistical analysis is probably less than 10 per cent of the effort required to do the same thing with the data of 150 printed questionnaires.

## Sampling

### Types of Sampling

Participant selection has traditionally been a thorny issue among social scientists, well before the advent of online questionnaires (Bass & Firestone, 1980; Ness Evans & Rooney, 2013).

There are two main types of sampling strategies and subcategories within these. The first type is probability sampling, which aims to constitute a “representative sample” of the general population. One procedure is random sampling: “whereby a sample is drawn such that each member of the population has an equal probability of being included in that sample” (Ness Evans & Rooney, 2013, p. 126). Ness Evans and Rooney point out that random sampling rarely happens in the social sciences but that it is not really a problem because social scientists “are typically testing theories, not generalizing to entire populations” (p. 127). However, random assignment of participants to groups is crucial as it is “an important assumption of several statistical procedures” (p. 127).

Ness Evans and Rooney (2013) urge researchers to be cautious “in generalizing the results to populations that may differ from our sampled population” (p. 132). Dörnyei adds that the generalizability of findings based on data gathered via non-random sampling is “often negligible” (2007, p. 99). This is a common question too for academics who use convenience sampling (i.e. collecting data from their own students). These students represent a captive participant pool, who might be gently coerced into participating in order to earn some pocket money or to obtain a credit for their participation in the experiment. These participants are smart, accessible, willing and cheap. Depending on the topic of research, the results are not generalizable to the

whole population. A sample of students is fine for research on students' attitudes or performance, but it would be inadequate for broader research, for example, on the political views in a certain region or country (Ness Evans & Rooney, 2013). Sampling strategies are thus dictated by practical considerations: "it is in sampling, perhaps more than anywhere else in research, that theory meets the hard realities of time and resources" (Kemper, Stringfield, & Teddlie, 2003, pp. 273–274).

One specific problem is self-selection bias: only people who are interested in a topic and feel strongly about it, whether positively or negatively, will be willing to spend 20 minutes filling out an online questionnaire on it. The bias does not invalidate the research, but it does require careful interpretation of the results. In Dewaele and MacIntyre (2014, 2016), for example, we collected data from 1746 participants who were learning a foreign language focusing on their enjoyment and anxiety in the foreign language class. A paired *t*-test showed that the mean scores for foreign language enjoyment were significantly higher than for foreign language anxiety. It would have been easy but erroneous to jump to the conclusion that foreign language learners report experiencing more enjoyment than anxiety in class. Despite the size of the sample, it was very likely that learners who did not like foreign language classes much were less likely to fill out the questionnaire. Hence, our participants were not a representative sample of the whole foreign language learner population (it is unlikely that such a perfectly balanced sample could ever be constituted).

### **The Difficulty of Attaining Balance in Terms of Age, Education Level and Gender**

Participants in online questionnaires in the field of SLA and multilingualism are typically older adults (the average age of the 1573 participants in the Bilingualism and Emotion Questionnaire (BEQ) was 34 (Dewaele, 2013; Dewaele & Pavlenko, 2001–2003); the average age of the 2324 participants who filled out a questionnaire on swearing in English was 32 (Dewaele, 2016). The participants in Dewaele and MacIntyre (2014, 2016) were younger on average (24 years old), partly because of the unexpected recruitment of 11, 12, 13 and 14-year-olds.

Gender imbalance is another issue that I have encountered in all the open online questionnaires that I have run. No matter how big the sample, the majority of participants have been female. It is possible that this partly reflects the proportion of female learners engaged in foreign language learning or of

learners willing to spend 20 minutes in an exercise of self-reflection about their language use and their emotions when learning and using the foreign language. Statistical analysis of gender differences in the data is possible despite the imbalance. However, I wonder then if the minority of male participants willing to fill out the questionnaire may differ from those male learners who do not participate. In other words, to what extent are the male participants representative of all male learners?

Feldman Barrett said that Internet-based samples may be more representative of the general population's age but pointed out that they tend to over-represent participants with an above-average educational and socio-economic status who can afford to access a computer with an Internet connection (Feldman Barrett, cited in Jaffe, 2005). This was certainly the case in the previously mentioned corpora. In the BEQ, 90 per cent of participants had university degrees, including Bachelors, Masters and doctoral degrees (Dewaele, 2013, p. 43). Similar proportions emerged in Dewaele (2016).

I have speculated that the high proportion of participants with university education is probably linked to the fact that filling out these questionnaires requires practice, self-confidence, metalinguistic and metapragmatic awareness of one's language practices and a genuine interest in the topic. Filling out questionnaires is a particular literacy skill that university students in the social sciences acquire quickly but that can leave lay people dumbfounded. It requires an ability to condense complex experiences, feelings and language choices on highly abstract Likert scales. It also implies an awareness of the convention that answers in these numerical formats are at best an approximation of a complex truth. A typical comment on the BEQ, which asked about language preference for the communication of emotions with various categories of interlocutor, including "friends," was that it depended on the specific friend being addressed (Dewaele, 2013). The first exposure to such a questionnaire must be unsettling, comparable to one's first encounter with a dictionary or a grammar. However, no PhD is needed to be able to reflect on one's experiences in multiple languages, and a good number of our participants had lower education levels. Interestingly, their responses were generally comparable to those with higher levels of education (Dewaele, 2013).

Reflecting on the difficulty of recruiting more participants with lower education levels (Dewaele, 2013), I included a story from my research assistant at the time, Dr Benedetta Bassetti, about her unsuccessful attempt to collect data for the BEQ from working class, bilingual Italians:

I finally managed to get hold of a bunch of very uneducated Italian-English bilingual males, and tried to administer your questionnaire. Given the very low levels of literacy, I did it orally and recorded the answers. (...) The first problem

was of course getting the message across, these people have a down-to-earth approach that is simply devastating. ‘This research is about language and emotions.’ ‘About what?’ Language and—it’s about how you express your emotions in Italian and in English, like, y’know, when you’re in love, do you prefer to express this feeling in English or in Italian? You are THICK, of course if I’m in love with an Italian girl I say it in Italian, if I’m in love with an English I say it in English. Ok (sigh!), yes, of course, but which one do you prefer, like, in which language do you feel more comfortable to express this emotion, or in which language is it more meaningful?—the informant didn’t get this abstract idea, and the discussion shifted to whether he preferred an Italian or English girlfriend. The second problem was suspicion. Why do they want to know this? E’ un professore dell’università di Londra, he’s doing some research. Why does a professore want to know my opinion? Well you know, he has to do research to get his salary. —He gets paid to do this?!? (...) Finally, I would say that the level of abstraction was too high. At the question ‘if you had some unpleasant experiences in the past’, he answered ‘I didn’t have any unpleasant experiences’, so I decided to change the question into a concrete image ‘for instance, if your mother dies, would it be easier to talk about this in English or in Italian?’ He looked at me thoughtfully and thoroughly scratched his testicles for two minutes (Italian version of touching wood) (personal communication, 2003) (Dewaele, 2013, pp. 43–44)

This episode clearly shows that these Italian participants would have been unlikely to fill out either a paper-based or an online questionnaire, even if they had had Internet access. Being put under pressure to fill out the paper version of the questionnaire would probably have been equally unsuccessful, as information would have been incomplete. I did try to recruit multilingual participants with lower socio-economic status using a paper version of the BEQ. I asked parents from various ethnic and linguistic backgrounds at my daughter’s primary school to fill out the questionnaires. I managed to collect about 20 questionnaires in this way, but I did meet considerable resistance and unwillingness.

There is a serious point about all this: sitting in their comfy chairs, reviewers and examiners can complain about a sample being too highly educated, too imbalanced in terms of age and gender and therefore not representative of the general population. It’s only when the researcher has ventured on the ground that the gap between theory and reality becomes clear. Research instruments such as questionnaires with Likert scales and open questions on language choices probably look intimidating and probably even threatening to people who are not used to them. It may bring back unhappy school memories, it may provoke a fear of looking stupid in the eyes of the researcher as they may think that there are “right” and “wrong” answers. Explaining to them

that this is not the case might be insufficient to allay their fear. To conclude, we have to accept that it is unlikely for us to ever have a sample of participants that is representative of the general population. This should not discourage us from doing research, as long as we acknowledge the limitations of the research design and sample, and we remain careful in drawing conclusions from our findings.

## Reaching Vulnerable or Closed Niche Groups

To obtain ethics approval for research on human subjects, researchers must typically explain the aim of the research, the design, the target group and the instruments they will use. When the target group is foreign language learners or multilinguals over the age of 16, obtaining ethics approval is relatively easy in my institution. It becomes much harder when the target group are foreign language learners who are younger than 16 and when consent needs to be obtained from the school, from the teachers, from the parents and from the students themselves. This cascade of levels of consent makes the research more time-consuming, and the sample size will tend to be smaller. A different type of difficulty occurs when the target group is psychologically vulnerable, such as current or former (multilingual) psychotherapy patients. Firstly, it is impossible to obtain patient lists from therapists or official mental health organisations because of confidentiality issues. A consequence of this is that very little research has been carried out on larger samples of (former) patients. Most studies investigating these issues use a case study approach, with typically up to five participants. Beverley Costa and myself decided that such small samples and the evidence that emerged from these studies would never yield the kind of generalisable result that is needed to influence policy. In this particular project we wanted to attract attention to the fact that monolingualism is the norm in health services and that the fact that a patient is relatively fluent in the language of the host country does not mean s/he can be treated as a monolingual. Hence the need to adapt the training of future psychotherapists by introducing notions on individual multilingualism, the meanings of code-switching and its consequences for psychotherapy. In Costa and Dewaele (2012), we used an anonymous online questionnaire that attracted the feedback of 101 therapists using snowball sampling. In Dewaele and Costa (2013) we used the same approach to reach 182 multilingual clients who had experienced therapeutic approaches in different countries.

Some communities have their groups and forums where potential participants can be recruited. In Simpson and Dewaele (2019), we focused on self-



misgendering (i.e. making a gender error in referring to the self) among multilingual transgender speakers. Transgender people are sparsely distributed and often hidden, which means that the Internet is the ideal place to make contact with them. Being transgender, the first author knew the social-networking sites and forums used by the transgender community (over 30 groups and forums, based in various countries) where the call for participation could be issued. Being a member of the group meant that her request was also met with less suspicion than if an outsider had issued the call. We highlighted the fact that the questionnaire was completely anonymous. A total of 132 people participated.

## Sample Size

### How Much Is Enough?

Another important issue in questionnaire-based research is sample size. In other words, how many participants are needed to answer the research questions? Barkhuizen (2014) writes that when he is asked this question, his answer is “Well, it depends” (p. 5). Many factors need to be considered when deciding on the number of participants. Some factors are outside the control of the researcher: “the specific requirements of the research design and methods, the availability of the participants, constraints of time and human resources, and organizational structures within research sites, such as class size and timetabling” (p. 7). Other factors are under the control of researchers: “determining the purposes and goals of the study, planning for and monitoring access to participants and research sites, and gauging feasibility in terms of scale of the project, time constraints, and one’s own research knowledge and skills” (p. 7). Barkhuizen also advises researchers to consult published research literature in the same field and discuss the issue with experienced colleagues.

Ness Evans and Rooney (2013) take a more statistical point of view on the question “how many is enough?”. They explain that for quantitative research, the number of participants needed “depends on the power of [the] statistic. Parametric statistics such as the *t*-test and analysis of variance (ANOVA) have a great deal of power, but chi-square, a nonparametric procedure, has relatively low” (p. 134). They explain further that the research design determines the sample size. Larger samples are needed when multiple statistical comparisons are planned, smaller samples are fine for tightly controlled experiments where there is a strong relation “between the manipulated variable and the behaviour”



(p. 135). Larger samples are thus needed in field research, where researchers have less control than their colleagues in the laboratory. Having more participants can compensate for greater variability. However, it is worth noting here that more does not automatically mean better. In very large-scale studies, because the analyses are very highly powered (large  $N$ ), correlations and mean differences are more likely to be statistically significant. In other words, there is a danger of getting an inflated view of the relationships in question. To avoid this, researchers have to be sure to pay attention to the actual size of the relationship as measured not by  $p$  but the  $r$  or  $r^2$  value or the  $d$  value for mean differences (see Plonsky & Oswald, 2014).

### Sample Study 13.1

Dewaele, J.-M. (2015). British 'Bollocks' versus American 'Jerk': Do native British English speakers swear more—or differently—compared to American English speakers? *Applied Linguistic Review*, 6(3), 309–339.

Dewaele, J.-M. (2016). Thirty shades of offensiveness: L1 and LX English users' understanding, perception and self-reported use of negative emotion-laden words. *Journal of Pragmatics*, 94, 112–127.

#### Background

Previous work on swearing in a foreign language has shown that swearwords are perceived to be more powerful, and preferred, in the L1 than in the LX, even among bilinguals who feel equally proficient in both languages.

#### Method

An online questionnaire attracted 2347 participants (1636 females, 664 males). This included 1159 English first language users and 1165 English foreign language users. They were asked to rate a list of 30 emotion-laden words and expressions ranging in emotional valence from mildly negative to extremely negative. More specifically they were asked about how sure they were about the meaning of the word, the offensiveness and their frequency of use. The 2015 study was a comparison of the same variables for the 414 British and 556 L1 users of American English.

#### Research Questions

1. Do LX users of English report being as comfortable in their understanding of the meaning of the 30 words as L1 users of English?
2. Do LX users of English have the same perception of offensiveness of the 30 words as L1 users of English?
3. Do LX users of English report comparable frequencies of use of the 30 words as L1 users of English?
4. What is the effect of having lived in an English-speaking environment on LX users' understanding of the meaning, perception of offensiveness and reported frequencies of use of the 30 words?

5. What is the effect of context of acquisition on LX users' understanding of the meaning, perception of offensiveness and reported frequencies of use of the 30 words?
6. What is the relationship between self-reported level of oral proficiency in English and LX users' understanding of the meaning, perception of offensiveness and reported frequencies of use of the 30 words?

### Results

The LX users were found to overestimate the offensiveness of most words. They were significantly less sure about the exact meaning of most words compared to the L1 users and reported a preference for relatively less offensive words, while the L1 users enjoyed using the taboo words. Among the LX users more contact and exposure to English was linked to a better understanding of the meaning of the words, a better calibration of offensiveness and frequency of use.

The comparison of British and American L1 users showed that the British participants gave significantly higher offensiveness and frequency of use scores to four words (including "bollocks") while the American English L1 participants rated a third of words as significantly more offensive (including "jerk"), which they also reported using more frequently.

### Comments

These two papers relied on a corpus that was large enough to allow multiple studies, comparing different groups and subgroups. One of the typical questions from reviewers was how reliable self-report is. This is a tricky question because it is impossible to claim that it is entirely reliable. The main defence against this type of critical observation is that since the questionnaire was anonymous there was no reason for participants to lie. Also, the participant is a pretty reliable source of his/her own linguistic behaviour. Everybody has a pretty clear idea about the frequency with which they use certain swearwords. Using a 5-point Likert scale is both vague and precise enough for the purpose of the research. It allows for precise comparisons of scores for different words. Actual production data are obviously better but other methodological problems arise. Wiring somebody up in order to record daily interactions for a length of time (typically never more than two days) might yield interesting data about swearing if that person happens to find itself in the right situation but if no such situation is encountered or if the person swears infrequently, no usable data would be collected. Moreover, it would be impossible to collect data from more than 1000 participants over a sufficient long period (not even mentioning the ethical issues).

## Getting the Ball Rolling

As some of my students have found, putting a questionnaire online is no guarantee of mass participation. Snowball sampling implies that the researcher throws the ball in the direction of fresh snow that will adhere and that multiple little pulls and pushes are needed to keep it rolling. I typically spend about two days contacting students, friends, colleagues asking them nicely to

fill out the questionnaire and to forward the call for participation to their students, colleagues and anyone who fits the selection criteria. Posting the call on *Linguist List* (<http://linguistlist.org>) or professional listservs also helps. It is good to have a wide network of socio-professional contacts and friends. My guess is that I probably send out about 3000 emails when launching a new project. I have learned not to include more than 500 addresses per message in order to avoid the account being blocked for spamming. I do not see calls for participation in studies as spam, but I agree that when it happens too frequently it can be perceived as (friendly) harassment. I often help my students in sending out their calls for participation because I realise that those who receive a call are more likely to participate and forward it when they recognise the name of the sender. As a rule, I also respond positively to calls for participation and forward these to my students. In the academic world, we need to help each other out. I do get the odd angry reaction, with an addressee's asking me how I got his or her address. I have generally no idea how the person ended up in my list of email contacts, but having been editor of the *International Journal of Bilingual Education and Bilingualism* since 2013 has certainly boosted my list of contacts. Once somebody asked for proof of the ethics clearance for the study, which I forwarded to him, as it was a legitimate question.

A delicate decision is at what point to withdraw the online questionnaire from the web. This depends firstly on the schedule of the researcher and whether, for example, a date has been set to start analysing the data in order to write a dissertation or to prepare a paper for a conference. The researcher can also decide to analyse the first batch of data while the questionnaire remains active on the web. *Google Forms* allows the researcher to see how the recruitment is going. If the number of new participants tails off to fewer than one or two a day, it might be the right time to haul in the data. A fitting metaphor is that of the fisherman/woman hauling in the net, hoping that the catch will be glittering, abundant and rich, and will contain the fish that were targeted. It is important at that point not to be discouraged by a smaller than expected haul. Good research can be carried out on small datasets, but it may need some tweaking of the research questions.

## **Does the Questionnaire Look Attractive and Professional?**

When the request to participate is sent out, it is crucial that the questionnaire looks good and professional. This means that it cannot have any spelling mistakes, ungrammatical constructions, ambiguous questions, irritating questions

and too many open questions that require too much effort on the part of the participants. Convincing someone to fill out a questionnaire for free is only going to work if the layout is nice. If it is sloppy, or the first paragraph sounds pretentious, or there are mistakes, people will not bother. This is also why it is important to pilot test the questionnaire with a small group of critically minded participants who can point out what is wrong or ambiguous or boring. I recommend avoiding questions in the style of “Do you do A more than B?” as the results are hard to analyse statistically. It is much better to ask “How often do you do A?” with a Likert scale, followed by “How often do you do B?” with a Likert scale. This will allow the calculation of means and a simple t-test will say whether the difference is statistically significant.

Another crucial aspect of an attractive questionnaire is having a neutral, concise, clear and interest-arousing title that accurately reflects the content of the questionnaire but does not give away too much to avoid potential bias in the responses. In other words, the title should not contain words such as “the benefits of...” nor “the dangers of...” which might attract more people who agree with the message and the agenda it betrays or which might influence the responses of participants to please the author of the questionnaire. The same principle applies to the inclusion of a sufficient number of items and questions that reflect different possible views and do not steer participants in a particular direction. It is also a good idea to include a final open question about observations the participant wishes to make about the questionnaire. The observations can include requests for being contacted once the findings are published, suggestions about missing elements in the questionnaires, congratulations or complaints about the research project. Obviously, the title of the paper or book based on the findings can have a clear positive or negative direction, possibly with a question mark at the end. It informs the reader that a certain hypothesis is being tested. I recently co-authored a paper (Dewaele, MacIntyre, Boudreau, & Dewaele, 2016) entitled “Do girls have all the fun? Anxiety and enjoyment in the foreign language classroom.” The question was deliberately provocative and a reference to the famous 1983 Cyndi Lauper song, “Girls just want to have fun.” We felt that the question part of the title was both catchy and appropriate as our female participants reported significantly more enjoyment but also more anxiety in the FL classroom (effect sizes were small though).

A final point needs to be made on the current trend of filling out online questionnaires using a mobile phone. A distraught PhD student recently told me that her pilot project participants had only selected the first four boxes on 9-point Likert scales after viewing video fragments. She then discovered that the phone screen gave a restricted view and that her participants had been unable to see the right side of her questionnaire, which was perfectly visible on a laptop.

## Limitations and Future Directions

The present chapter has discussed some of the limitations inherent to the use of paper-based and online questionnaires. What has not yet been mentioned is the fact that “it is all too easy to produce unreliable and invalid data by means of an ill-constructed questionnaire” (Dörnyei, 2007, p. 115). My advice is thus to consult a good source like Dörnyei and Taguchi (2010) before attempting the creation of a questionnaire. Questionnaires developed haphazardly on the kitchen table without proper pilot testing will lead to tears and disappointment. Budding authors of questionnaires need to remember to keep questions simple in order to elicit simple short responses so that participants do not spend too long on filling out the questionnaire. As a consequence, the investigation cannot go very deep and “questionnaire surveys usually provide a rather “thin” description of the target phenomena” (p. 115). One possible way to avoid shallowness is a mixed-method approach which involves gathering both quantitative and qualitative data in a single study, allowing the researcher to identify significant trends in the quantitative data, the possible causes of which can be pursued in interviews with a smaller number of participants (see Creswell & Plano Clark, 2010). This approach is particularly useful for a comparison between “typical” participants and outliers. Interviews often reveal totally unexpected reasons for specific behaviour or attitudes. Interviews are also an excellent way to complement questionnaires. Benedetta Bassetti helped me in interviewing 20 multilinguals after they had filled out the BEQ (Dewaele, 2013). Noticing that a participant had indicated that she had never used swearwords in her English L2, Benedetta challenged her in the interview. The participant admitted that she had indeed started using mild English swearwords, but only when having tea with her Chinese friends in London. This is a nice example of the fact that the questionnaire is a passive and relatively bland receptacle of data; it is good if it can be complemented by the quick thinking of an inquisitive researcher in an interview.

The future of online questionnaires is bright because of the continuous growth of social media and technological advances. Questionnaires can include video clips on *YouTube* about which specific questions can be asked in the questionnaire (see Sample Study 13.2); webcams can catch facial expressions of participants watching videos, which can be analysed with facial recognition software.

### Sample Study 13.2

Lorette, P., & Dewaele, J.-M. (2015). Emotion recognition ability in English among L1 and LX users of English. *International Journal of Language and Culture*, 2(1), 62–86.

#### Background

To what extent do foreign language users struggle to recognise emotions in that language compared to first language users? Rintell (1984) showed that EFL students performed less well than a control group of native speaker students in recognising emotions in audio recordings. Proficiency and cultural distance were the main independent variables to have an effect.

#### Method

An online questionnaire was used to test individual differences in the emotion recognition ability (ERA) of 356 first language and 564 foreign language users of English. Participants were shown six short English videos of an English L1 British actress displaying an improvisation of six basic emotions. Participants were asked to identify the emotion portrayed by the actress (<http://www.youtube.com/embed/8VcoNbk3HVE>). A comparison of the performance of native and foreign language users of English revealed no difference between the two groups despite the foreign language users scoring significantly lower on a lexical decision task (LEXTALE) which was integrated in the questionnaire ([www.lextale.com](http://www.lextale.com)), which was used as an indicator of English proficiency.

#### Research Questions

1. Are there differences between L1 and LX users of English in their ability to recognise a (basic) emotion conveyed by an L1 British English speaker?
2. Is proficiency in English linked to the ability to recognise a (basic) emotion conveyed by an L1 British English speaker?
3. Is a participant's L1 culture linked to their ability to recognise a (basic) emotion conveyed by an L1 British English speaker?

#### Results

LX users of English can generally recognise basic emotions in English video clips as accurately as L1 users of English despite lower levels of proficiency. LX users' proficiency in English is related to their ERA in English, but the threshold for the successful recognition in an LX is probably lower than has been assumed so far. English L1 users with high proficiency scores tended to score higher on the ERA, which could hint at variation on some underlying linguistic or psychological dimension. A link exists between cultural distance and ERA as the Asian participants scored significantly lower on ERA than other groups, despite having English proficiency levels comparable to those of Continental European participants.

#### Comments

We experienced some difficulty in convincing the reviewers that our video clips were appropriate stimuli. We argued that the use of the picture of a face—a common method in emotion recognition research—might be more straightforward but that our spontaneous stimuli were more likely to reflect the messy reality we face every day, where we observe somebody during a stretch of time and form an opinion on the emotions that the speaker experiences based on the weighing of verbal, vocal and non-verbal cues.

## Conclusion

Using online questionnaires in applied linguistics and multilingualism research is a great way to obtain rich and abundant data while sitting behind one's desk. It is crucial to remain aware of the possibilities and the limitations of the data gathered in this way. At some point, the desk will have to be left behind in order to face some of the anonymous participants, and their input has the potential to create major surprises and revelations about possible causes behind the observed phenomena. It is equally important to express gratitude towards the many participants who spent some precious time answering the many questions. This can happen in the acknowledgement at the end of a published paper or book or at the start or end of a conference presentation. Since many of the participants may be friends and colleagues, or their students, it's crucial to acknowledge that without their input, we would be standing there with empty hands. It is equally important to behave ethically, meaning that if we promised anonymity, nobody should recognise themselves or somebody they know in the research. As academics, we have to respect proper etiquette just like actors in their traditional tearful speech at the Oscars.

## Resources for Further Reading

Creswell, J., & Plano Clark, V. (2010). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

This is the “bible” for mixed-method research. It offers practical and step-by-step guidance on how to develop a research design—a crucial step that precedes the creation of a questionnaire. It discusses the formulation of research questions, the collection of data and the interpretation of results.

Dörnyei, Z., & Taguchi, N. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). London: Routledge.

This revised edition is again the prime source of information for anyone wishing to design questionnaires. It is very hands-on and no-nonsense. Like “celebrity” chefs, the authors share their recipes and provide advice on how to create reliable, valid and successful questionnaires.

## References

- Barkhuizen, G. (2014). Number of participants. *Language Teaching Research*, 18(1), 5–7.
- Bass, A. R., & Firestone, I. J. (1980). Implications of representativeness for generalizability of field and laboratory research findings. *American Psychologist*, 35(2), 141–150.
- Costa, B., & Dewaele, J.-M. (2012). Psychotherapy across languages: Beliefs, attitudes and practices of monolingual and multilingual therapists with their multilingual patients. *Language and Psychoanalysis*, 1, 18–40; revised version reprinted in (2014) *Counselling and Psychotherapy Research*, 14(3), 235–244.
- Creswell, J., & Plano Clark, V. (2010). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: SAGE.
- Dewaele, J.-M. (2013). *Emotions in multiple languages* (2nd ed.). Basingstoke: Palgrave Macmillan.
- Dewaele, J.-M. (2015). British ‘Bollocks’ versus American ‘Jerk’: Do native British English speakers swear more—or differently—compared to American English speakers? *Applied Linguistic Review*, 6(3), 309–339.
- Dewaele, J.-M. (2016). Thirty shades of offensiveness: L1 and LX English users’ understanding, perception and self-reported use of negative emotion-laden words. *Journal of Pragmatics*, 94, 112–127.
- Dewaele, J.-M., & Costa, B. (2013). Multilingual clients’ experience of psychotherapy. *Language and Psychoanalysis*, 2(2), 31–50.
- Dewaele, J.-M., & MacIntyre, P. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, 4(2), 237–274.
- Dewaele, J.-M., & MacIntyre, P. (2016). Foreign language enjoyment and foreign language classroom anxiety. The right and left feet of FL learning? In P. MacIntyre, T. Gregersen, & S. Mercer (Eds.), *Positive psychology in SLA* (pp. 215–236). Bristol: Multilingual Matters.
- Dewaele, J.-M., MacIntyre, P., Boudreau, C., & Dewaele, L. (2016). Do girls have all the fun? Anxiety and enjoyment in the foreign language classroom. *Theory and Practice of Second Language Acquisition*, 2(1), 41–63.
- Dewaele, J.-M., & Pavlenko, A. (2001–2003). *Web questionnaire on bilingualism and emotion*. Unpublished manuscript, University of London.
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.
- Dörnyei, Z., & Taguchi, N. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). London: Routledge.
- Jaffe, E. (2005). How random is that. *American Psychological Society*, 9, 17–30. Retrieved from <http://www.psychologicalscience.org/observer/how-random-is-that#.WI68i3qvjvo>



- Kemper, E. A., Stringfield, S., & Teddlie, C. (2003). Mixed methods sampling strategies in social research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 273–296). Thousand Oaks, CA: SAGE.
- Ness Evans, A., & Rooney, B. J. (2013). *Methods in psychological research* (3rd ed.). New York: Sage Publications.
- Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912.
- Simpson, L., & Dewaele, J.-M. (2019). Self-misgendering among multilingual transgender speakers. *International Journal of the Sociology of Language*.
- Wilson, R., & Dewaele, J.-M. (2010). The use of web questionnaires in second language acquisition and bilingualism research. *Second Language Research*, 26(1), 103–123.



# 14

## Psycholinguistic Methods

Sarah Grey and Kaitlyn M. Tagarelli

### Introduction

Psycholinguistics is dedicated to studying how the human mind learns, represents, and processes language (Harley, 2013). In recent decades, overlapping interests in psychology and linguistics have resulted in a surge in knowledge about language processing, and the psycholinguistic methods borne out of this research are increasingly of interest to applied linguistics research.

Like other language research methods—such as surveys, interviews, or proficiency tests—psycholinguistic methods can uncover information about language learning and use. Crucially, through experimental and theoretical rigor, psycholinguistic methods also reveal the psychological representations and processes underlying language learning and use. Although research conducted in naturalistic settings—such as in foreign language classrooms—provides important perspectives, such work is inherently confounded with extraneous variables. For example, even if one were to assume that all students receive the exact same input during a Spanish class, there is likely a large range in how much and in what ways (e.g., music, television, friends) each student is exposed to Spanish outside the classroom. This lack of empirical control

---

S. Grey (✉)

Department of Modern Languages and Literatures, Fordham University,  
Bronx, NY, USA

e-mail: [sgrey4@fordham.edu](mailto:sgrey4@fordham.edu)

K. M. Tagarelli

Department of Neuroscience, Georgetown University, Washington, DC, USA

e-mail: [kaitlyn.tagarelli@gmail.com](mailto:kaitlyn.tagarelli@gmail.com)

severely limits the precision and clarity with which researchers can measure and interpret language-related cognitive processes, which has consequences for the theories and applications informed by these interpretations. Psycholinguistic experiments, which are carried out in laboratory settings, employ carefully controlled designs that enable researchers to target factors of interest (e.g., the effects of phonological similarity on word recognition) while limiting, or controlling for, potentially confounding variables (e.g., word frequency). Their capacity for collecting detailed and well-controlled information is perhaps the strongest advantage of using these methods in applied linguistics research. Furthermore, psycholinguistic methods can reveal information about language learning and processing where other methods cannot. As we will discuss in more detail, traditional measures of performance or proficiency are often too coarse to distinguish different levels of language processing, but psycholinguistic measures like reaction time and brain recordings are more sensitive and thus can reveal effects where behavioral methods cannot.

In this chapter, we review psycholinguistic methods. We do not cover the entire methodological corpus of psycholinguistics but instead focus on methods that are particularly relevant for addressing compelling questions in applied linguistics. In the first part of the chapter, we discuss psycholinguistic measures and tasks, and in the second part we describe common experimental paradigms.

## Tasks and Measures

Psycholinguistic methods are especially useful for studying the cognitive processes about language learning and use, from phonetics and phonology to discourse-level pragmatics. This section reviews common psycholinguistic measures and tasks that are useful for applied linguistics researchers, together with strengths and limitations of these methods. The section begins with behavioral measures—including decision tasks, reaction time, mouse-tracking, and eye-tracking—and ends with a discussion of a popular neurocognitive measure, event-related potentials.

### Decision Tasks

We conceptualize decision tasks as experimental tasks that instruct participants to make a decision in response to a stimulus, from which researchers infer some underlying psycholinguistic process or representation. Such tasks are incredibly common in psycholinguistics and can be employed along all language domains. For example, a phoneme discrimination task requires

participants to decide whether a pair of speech sounds (e.g., [pa] and [ba]) are the same or different and provides insight into phonological representations. A lexical decision task (LDT) requires participants to decide whether a string of letters or sounds constitutes a real word (*cucumber*) or not (*nutolon*). LDTs reveal information about the organization of and access to the mental lexicon via a variety of experimental manipulations that tap domains such as orthography, phonology, morphology, and semantics.

At the sentence level, a widely used decision task elicits acceptability judgments, where participants might decide whether a sentence is grammatically well-formed, as in (1) or perhaps whether a sentence makes semantic sense, as in (2).

1. a. The winner of the big trophy *has* proud parents.  
 b. The winner of the big trophy *ʃhave* proud parents. (Tanner & Van Hell, 2014)
2. a. Kaitlyn traveled across the ocean in a *plane* to attend the conference.  
 b. Kaitlyn traveled across the ocean in a *ʃcactus* to attend the conference. (Grey & Van Hell, 2017)

Acceptability judgment tasks are popular in second language acquisition research because researchers can design materials to assess knowledge of specific target structures, such as word order (Tagarelli, Ruiz, Moreno Vega, & Rebuschat, 2016) or grammatical gender agreement (Grey, Cox, Serafini, & Sanz, 2015). Additionally, acceptability judgments that impose a time limit on the decision have been shown to reliably tap implicit linguistic knowledge, while those that are untimed tap explicit knowledge (Ellis, 2005).

A limitation of decision tasks is that they require participants to make an explicit and oftentimes unnatural evaluation about the target content: “Are these two sounds the same or different? Is this sentence correct?”. This may fundamentally change the way the given linguistic information is processed, which has consequences regarding the extent to which the inferred psychological representations can be said to underlie naturalistic language processing. However, this limitation does not outweigh the benefits: decision tasks are easy and efficient to administer and most are supported by cross-validation through decades of research. Also, in addition to using decision tasks to gather data on a target variable (e.g., accuracy in LDT) researchers can use them to keep participants attentive during an experiment or to mask the goals of the study while other less explicit measures, such as eye movements, are gathered (see Sample Study 14.1).

## Latency Measures

Questions about language processing are often associated with the notions of “time” or “efficiency,” and as such many psycholinguistic tasks include a measure of latency: the time between a stimulus and a response. Perhaps the most common psycholinguistic measure of latency is reaction time (RT): a measure of the time it takes, in milliseconds, to respond to an external stimulus, usually via button press (or by speech onset in naming tasks). In applied linguistics, RTs gained early popularity in studies of L2 automaticity (Segalowitz & Segalowitz, 1993) and are increasingly employed to study L2 development, for example, in study abroad research (Sunderman & Kroll, 2009) and in testing different pedagogical techniques (Stafford, Bowden, & Sanz, 2012). RT in fact has been critical in clarifying the efficacy of different pedagogical techniques, since latency differences can be observed in the absence of accuracy differences (e.g., Robinson, 1997), thereby revealing finer-grained details about language learning than can be captured with accuracy alone (for a review, see Sanz & Grey, 2015).

RTs are used for button press and similar responses, whereas voice onset time (VOT) is a production-based measure of latency that can be used to examine the timing and characteristics of articulation. VOT is the time, in milliseconds, between the onset of vocal fold vibrations and the release of the articulators. VOT exhibits well-identified cross-linguistic patterns (Lisker & Abramson, 1964), making it an attractive metric for studying the psycholinguistic processes underlying articulation in different language groups. For example, VOT has been used to study speech planning in L2 learners (Flege, 1991) and bilingual code-switching (Balukas & Koops, 2015).

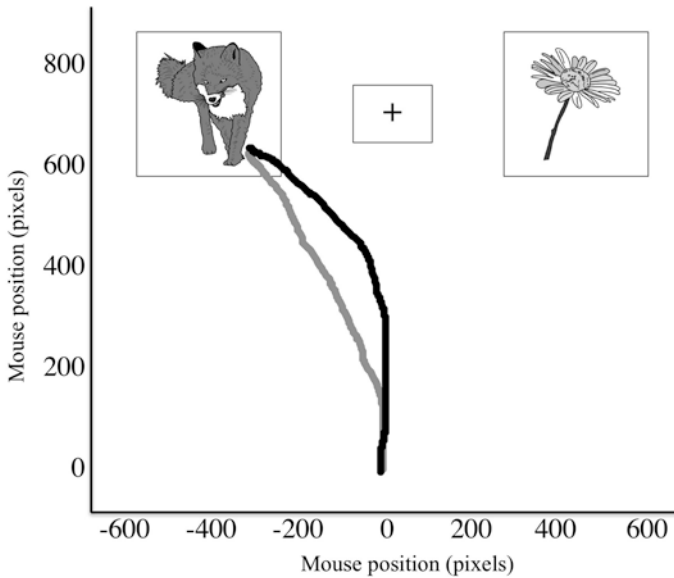
A third latency measure is eye-tracking. Eye-tracking data, like VOT, are ecologically appealing because they measure naturally occurring behavior as it unfolds in time, for example, as participants view scenes or read sentences. Eye movements are measured in terms of fixations, saccades (Rayner, 1998), or proportion of looks to regions of interest (ROIs, e.g., a word or picture; Huettig, Rommers, & Meyer, 2011). Fixations refer to the time spent on a location and can be divided into earlier measures, such as first-fixation duration, and later measures, such as total time in an ROI. Saccades refer to quick eye movements from one location to another and provide information on forward movements in reading (forward saccades, which are rightward movements in most languages) as well as returns to a location, or regressive saccades (Rayner, 1998). Such measures have recently been used to study attention in L2 learning (Godfroid, Housen, & Boers, 2013) and the use of subtitles in foreign language films (Bisson, Van Heuven, Conklin, & Tunney, 2012).



**Fig. 14.1** Sample visual world. Note: In this example, “cat” is the target, “caterpillar” is an onset competitor, “bat” is a rhyme competitor, and “hedgehog” is an unrelated distractor. Images are from the Multipic database (Duñabeitia et al., 2017)

Proportion of looks to a target ROI (compared to competitors) is another metric for eye-movement data and is often used to measure language processing in visual world paradigms, where both linguistic and visual information are presented (Fig. 14.1; Sample Study 14.1).

A more recent latency measure is mouse-tracking, which records the trajectory of the computer mouse on the screen by sampling the movement many dozen times a second (Freeman & Ambady, 2010). This technique allows researchers to examine competition and selection between multiple response options with detailed latency information. The continuous trajectory of the mouse as it moves toward a target (see Fig. 14.2) is the core metric, which is often analyzed by calculating its maximum deviation (MD) or area under the curve (AUC). MD refers to the largest perpendicular deviation between the ideal mouse trajectory (straight line to the target) and the observed trajectory, and AUC is the area between the ideal and observed trajectories. Although



**Fig. 14.2** Sample data from mouse-tracking language experiment. Note: The black line represents a competitor trajectory; the gray line represents a target trajectory. Images are from the Multipic database (Duñabeitia et al., 2017)

mouse-tracking methodology is rather new, it is gaining momentum in language research and, unlike many types of psycholinguistic testing software, mouse-tracking software for experimental research is freely available (*MouseTracker*, Freeman & Ambady, 2010) which makes it especially appealing. Of interest to applied linguistics, mouse-tracking has recently been employed to study pragmatic intent (Roche, Peters, & Dale, 2015), syntactic transfer in bilinguals (Morett & Macwhinney, 2013), and lexical competition in monolinguals and bilinguals (Bartolotti & Marian, 2012).

## Language Production

Using measures such as VOT, psycholinguistic methods are fruitful in revealing insights about production. One well-studied production task is picture naming, whereby participants are presented with pictures and asked to name them aloud. This task reveals information on lexical access and the organization of the mental lexicon more generally and has been employed across many different populations and languages (Bates et al., 2003). Researchers usually measure naming accuracy and can gather naming latency, as well as VOT. This makes picture naming versatile in the information it provides; it has been used

to test questions with implications for applied linguistics, including those pertaining to the effects of semantics on L2 word learning (Finkbeiner & Nicol, 2003) as well as conceptual/lexical representations during lexical access (for a review, see Kroll, Van Hell, Tokowicz, & Green, 2010).

Elicited imitation (EI) is another production task that is impressively simple. Participants listen to a sentence and are asked to repeat it verbatim. The premise of EI is that if participants can repeat a sentence quickly and precisely, they possess the grammatical knowledge contained in that sentence. Under this premise, EI has been established as a reliable measure of language proficiency, both globally (Wu & Ortega, 2013) and on specific linguistic structures (Rassaei, Moinzadeh, & Youhannaee, 2012). Additionally, EI can assess implicit linguistic knowledge (Ellis, 2005; Spada, Shiu, & Tomita, 2015), the attainment of which is a central research area in applied linguistics. EI has also been the subject of increased methodological rigor, spurring a recent meta-analysis of the technique (Yan, Maeda, Lv, & Ginther, 2016) and the extension of EI to assess proficiency in heritage languages (Bowden, 2016).

The main critique of EI is in fact its main design element. Because participants are repeating predetermined sentences outside of a larger discourse context, the language produced is not representative of naturalistic speech. This critique is stronger for designs that include ungrammatical sentences (e.g., Erlam, 2006), which are quite common in L2 research, but by their very nature do not represent genuine language.

The study of speech disfluencies offers a more naturalistic psycholinguistic perspective. Disfluencies occur very frequently in natural language production (Fox Tree, 1995) and include editing terms (e.g., *uh* and *um*), pauses, repetitions, restarts, and repairs. They have long been used as a window into the effects of cognitive load on speech planning and production. For example, disfluencies are more frequent at the beginning of utterances, preceding longer utterances, and for unfamiliar conversational topics (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Oviatt, 1995), where cognitive demand is relatively high.

An increasing amount of research examines the effects of speaker disfluencies on language comprehension. Interestingly, these effects appear to be facilitative. For example, disfluencies speed up word recognition (Corley & Hartsuiker, 2011) and serve as cues to new information in discourse (Arnold, Fagnano, & Tanenhaus, 2003), though this may depend on speaker-specific factors such as native versus non-native speech (Bosker, Quené, Sanders, & de Jong, 2014). Growing psycholinguistic interest in disfluencies is promising for future research, in part because they represent naturalistic phenomena with implications for comprehension and production.



## Event-Related Potentials

In recent years, psycholinguistic interests have expanded to consider questions on how the human brain acquires, represents, and processes language. One approach for considering such questions is the event-related potential (ERP) technique. Many of the methods discussed above can be coupled with this technique, such as acceptability judgments (Tanner & Van Hell, 2014) and LDTs (Barber, Otten, Kousta, & Vigliocco, 2013). Compared to other neuroimaging techniques like PET, MEG, and MRI, ERP equipment is generally much less expensive, and there are even excellent lower-cost ERP equipment options on the market (e.g., actiCHamp from Brain Products, Germany). As such, ERPs provide applied linguists with a cost-effective method for applying neuroscience technology to timely issues in applied linguistics research (for a review, see Morgan-Short, 2014).

ERPs are derived through recording naturally occurring electroencephalogram data, which consist of changes in the brain's electrical activity measured from electrodes on the scalp. Using ERPs, researchers investigate neurocognitive processes with millisecond precision (see Luck, 2014, for detailed information), which are elicited in response to a time-locked event (e.g., the onset of the words “plane” or “cactus” in (2) above). ERP language studies often use violation paradigms, which are characterized by measuring the neural activations of correct/standard stimuli (e.g., 1a, 2a above; or “cucumber” in LDTs) compared to matched “violation” stimuli (e.g., 1b, 2b; “nutolon” in LDTs). A key ecological advantage of ERPs is that the ERPs themselves serve as automatic, implicit responses to such stimuli, so researchers do not necessarily *have* to elicit an explicit decision from their subjects, which, as discussed above, could result in less naturalistic language processing.

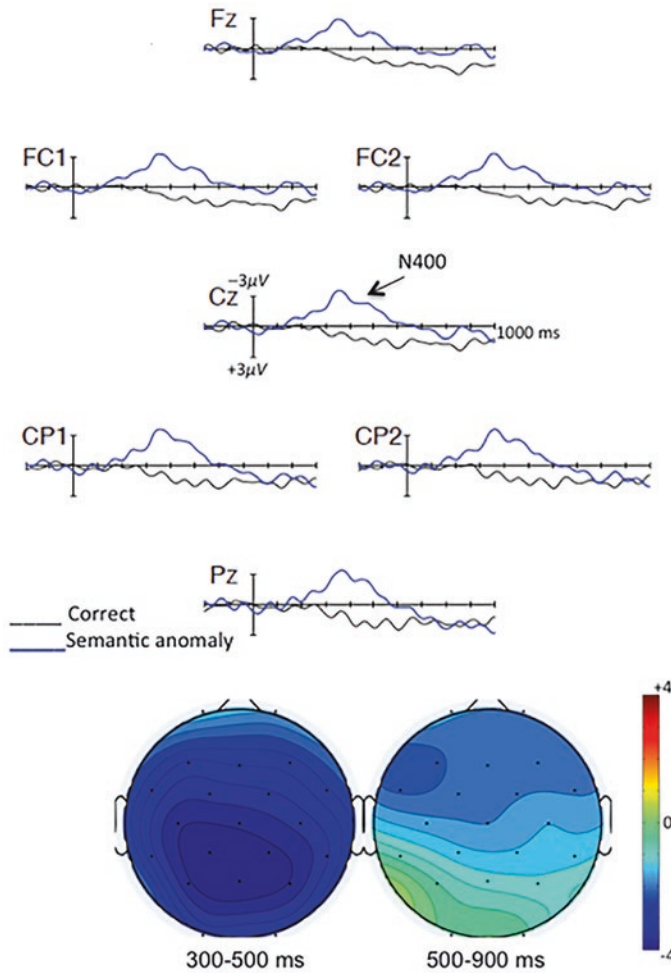
ERP language research has revealed a set of well-studied neural responses, or ERP effects, that are considered to reflect different neurocognitive processes. Table 14.1 summarizes some of these effects and Fig. 14.3 illustrates a sample ERP pattern for semantic processing. In the last 15 years, ERPs have become popular for applied linguistics interests (Morgan-Short, 2014), with studies testing the role of L1/L2 proficiency (Newman, Tremblay, Nichols, Neville, & Ullman, 2012), different L2 training conditions (Batterink & Neville, 2013), cross-linguistic transfer (Gillon Dowens, Guo, Guo, Barber, & Carreiras, 2011), and prediction during L2 reading (Martin et al., 2013), among others. Notably, changes in ERP responses as a result of language development have been observed even when proficiency changes are not apparent at the behavioral level (McLaughlin, Osterhout, & Kim, 2004).

Table 14.1 Common language-related ERP effects

ERP effect	Language domain	Sample stimuli	^Representative study
N400	Semantics	They wanted to make the hotel look more like a tropical resort, so along the driveway they planted rows of <i>palms</i> <i>† pines</i> <i>†† tulips</i> (Federmeier & Kutas, 1999)	McLaughlin et al. (2004)
P600	(Morpho)syntax	The man in the restaurant doesn't like the <i>hamburgers</i> that <i>are</i> on his plate. The man in the restaurant doesn't like the <i>hamburger</i> that <i>tare</i> on his plate. (Kaan & Swaab, 2003)	Bowden, Steinhauer, Sanz, and Ullman (2013)
<sup>1</sup> AN	(Morpho)syntax	The scientists criticized Max's <i>proof of</i> the theorem. The scientists criticized Max's <i>tof proof</i> the theorem. (Neville, Nicol, Barss, Forster, & Garrett, 1991)	Batterink and Neville (2013)
<sup>2</sup> PMN	Pre-lexical phonetics/phonology	snap – nap snap – <i>†tap</i> (Newman & Connolly, 2009)	Goslin, Duffy, and Floccia (2012)
Frontal positivity	Lexical prediction	The bakery did not accept credit cards so Peter would have to write <i>a check</i> <i>†an apology</i> to the owner (DeLong, Groppe, Urbach, & Kutas, 2012)	Martin et al. (2013)

†Represents violation/unexpected/mismatch item. ^Related to interests in applied linguistics. <sup>1</sup>ANs (anterior negativities) have also been referred to as LANs (left anterior negativities). When present they generally appear as a biphasic AN-P600 response. <sup>2</sup>PMN, phonological mapping negativity

ERPs represent a highly sensitive temporal measure of neural processing, which is important for understanding language. They do not, however, reflect the spatial location within the brain that gives rise to the observed effects. That is, observing an ERP effect in posterior (parietal) scalp locations does not mean the activity originated in the parietal cortex. For this level of spatial detail, researchers use fMRI or MEG, though their high costs and lower temporal resolution make these techniques less common in psycholinguistics.



**Fig. 14.3** Sample ERP waves and scalp topography maps of the standard ERP correlate of semantic processing (*N400*). Note: Each tick mark on y-axis represents 100 ms; x-axis represents voltage in microvolts,  $\pm 3\mu V$ ; negative is plotted up. The black line represents brain activity to correct items, such as *plane* in example 2a. The blue line represents brain activity to a semantic anomaly, such as *cactus* in example 2b. The topographic scalp maps show the distribution of activity in the anomaly minus correct conditions with a calibration scale of  $\pm 4\mu V$ . From data reported in Grey and Van Hell (2017)

The use of decision tasks, latency measures, production measures, and brain wave recordings reveals information about the cognitive bases of language and is fundamental to psycholinguistic research. By employing these measures, the psycholinguistic paradigms—or standard experimental designs—described in the following section allow researchers to understand language learning and

processing in a way that complements and expands on traditional methods in applied linguistics.

## Psycholinguistic Paradigms

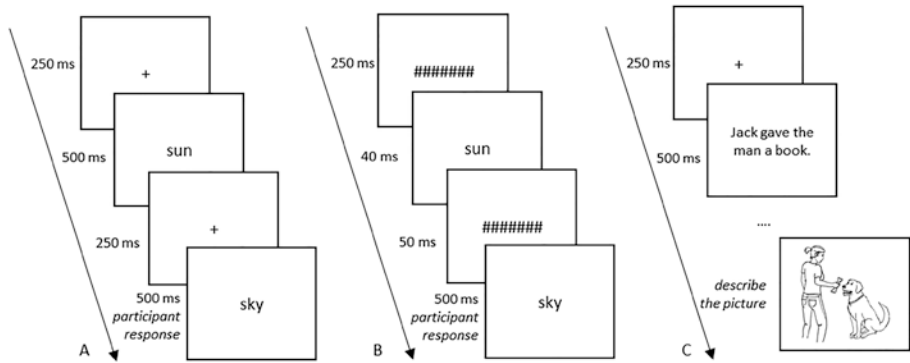
Over the last few decades, several reliable psycholinguistic paradigms have been established to investigate the mental underpinnings of language. This section reviews common experimental approaches that are relevant for interests in applied linguistics: priming, visual world, and language learning paradigms.

### Priming

Priming refers to the cognitive process whereby exposure to some language (the “prime”) influences processing of subsequently presented language (the “target”). This effect can be facilitative, wherein processing of the target speeds up as a result of the prime, or inhibitory, wherein processing of the target slows down. Priming effects generally reflect implicit processes—individuals are sensitive to language details without being aware of how such sensitivities influence subsequent language processing (McDonough & Trofimovich, 2009).

In a standard priming paradigm, the participant is briefly (for <1000 milliseconds) presented with a prime. After a short delay, they are presented with a target and make a decision about it or say something out loud (Fig. 14.4(a)). The properties of primes and targets, as well as the tasks and outcome measures, vary depending on the research question. In masked priming, which aims to isolate automatic, implicit processes, the prime is presented for a very brief time (~40–60 milliseconds) and “masked” so that it is not consciously perceptible. This mask is typically achieved by adding a string of random letters or symbols preceding and/or following the prime (Fig. 14.4(b)). RTs are the most common measure used in priming and are often gathered via decision tasks, especially LDT and semantic categorization (e.g., “Is it an animal?”), though production tasks, like naming, are also used.

Of interest to applied linguists, priming paradigms have been used in bilingualism and L2 research to examine two main questions: “Is priming the same in L1 and L2?” and “Do bilinguals have shared or separate stores for various aspects of language?” (McDonough & Trofimovich, 2009). These questions have been investigated using priming paradigms that tap phonetics, phonology, morphology, semantics, and syntax. Here, we focus on the latter two.



**Fig. 14.4** Examples of (a) semantic priming using lexical decision, (b) masked semantic priming, and (c) syntactic priming using a picture description task. Note: Drawing credit: Kyle Brimacombe

Semantic priming examines the extent to which processing of a target is facilitated by a prime with similar meaning. The underlying assumption is that the more semantically related two words are the more one will facilitate the other in a prime-target relationship. In within-language designs, researchers have tested semantic priming in L1 (e.g., *sun-sky*) and comparable L2 (e.g., *soleil-ciel*, sun-sky in French) prime-target pairs (e.g., Crinion et al., 2006). Findings suggest that in bilinguals and high-proficiency L2 speakers, priming effects tend to be similar in both languages, but for low-proficiency speakers, priming effects are stronger in L1 than L2 (Frenck-Mestre & Prince, 1997).

In between-language or cross-language semantic priming, participants are presented with prime-target pairs in the L1-to-L2 or L2-to-L1 direction (e.g., Basnight-Brown & Altarriba, 2007). These can be semantically related (*sun-ciel*) or translation equivalents (e.g., *sky-ciel*). If a bilingual's languages share a lexicon, an L1 prime should facilitate an L2 target, and vice versa. Findings from between-language semantic priming studies have been variable, with L2 proficiency and direction of priming influencing the effects (see Altarriba & Basnight-Brown, 2007, for a review).

Syntactic priming examines the extent to which the processing of a target syntactic structure is facilitated by a prime of similar structure. The focus of syntactic priming is on form. This paradigm tests whether the prime structure increases the likelihood that an individual will produce a sentence using that structure, compared to an equally acceptable structure. This is classically tested using dative constructions (e.g., Bock, 1986; for a review, see Pickering & Ferreira, 2008). For example, if a participant is presented with a double-object

dative prime, as in “Jack gave the man the book,” syntactic priming would be evidenced by the participant later producing the sentence “Ellen gave the dog the bone,” rather than the equivalent sentence with the prepositional dative, “Ellen gave the bone to the dog.” In syntactic priming, tasks tend to focus on production, such as picture description (see Fig. 14.4(c)) and sentence completion. Scripted interaction in which a participant has a conversation with a “confederate” who uses the structure(s) of interest (Kootstra, van Hell, & Dijkstra, 2010) is also used. Syntactic priming has been productive in examining the mental representation of grammatical structures, including in cross-linguistic research (Hartsuiker, Pickering, & Veltkamp, 2004) and work on L2 development (McDonough & Mackey, 2008).

Priming studies are limited in the questions they can address. For example, syntactic priming can only test a limited set of structures that have at least one other equivalent form. Additionally, priming studies often focus on very small RT differences, which can be influenced by many linguistic aspects outside the variables of interest. For instance, word frequency and length can impact observed effects. Therefore, stimuli in priming studies must be painstakingly designed to limit effects of confounding variables.

## Visual World

The visual world (VW) paradigm allows researchers to investigate the processing of linguistic and visual information together. In a standard VW paradigm, participants are presented with a visual display that might consist of a semi-realistic scene (e.g., a child surrounded by objects in a playroom) or a set of objects displayed on a computer screen (see Fig. 14.1). In a comprehension study, the participant hears an utterance (word or sentence), and in a production study they describe what they see. The dependent variables often consist of the participant’s eye movements within the display while listening or speaking. Similar to eye-tracking studies of reading (see Latency Measures section above), analyses in the VW paradigm focus on fixation proportions within and the number of saccades toward ROIs, such as target objects or experimental distractors. The properties of targets and distractors, as well as how they relate to the linguistic information, vary depending on the research questions. This flexibility in the research applications of the VW paradigm is one of its key benefits. Studies using this method have examined a range of questions, including the effects of cognitive biases (Kamide, Altmann, & Haywood, 2003) and linguistic information (Altmann & Kamide, 2007) on our expectations about language, effects of speaker reliability on pragmatic inferences

(Grodner & Sedivy, 2011), and effects of semantic and syntactic context on word recognition (Dahan & Tanenhaus, 2004), among others. For an illustrative study that employed the VW paradigm to study L2 grammatical processing, see Sample Study 14.1.

### Sample Study 14.1

Dussias, P. E., Kroff, J. R. V., Tamargo, R. E. G., & Gerfen, C. (2013). When gender and looking go hand in hand. *Studies in Second Language Acquisition*, 35(2), 353–387.

#### Background

Native-like morphosyntactic processing in L2 is difficult to attain and a central area of investigation in SLA. Regarding the online processing of grammatical gender, L1 work demonstrates that prenominal gender information on the article (*la*, feminine; *the*) tends to facilitate subsequent processing of the matching noun (*mesa*, feminine; *table*) whereas mismatching nouns (e.g., *escritorio*, masculine; *desk*) slow processing. Grammatical gender is studied less in L2, so it is not as well understood.

#### Research Questions

1. Can L1 speakers whose language does not instantiate grammatical gender show effects of prenominal gender-marking for L2 processing of the subsequent noun?
2. For speakers of an L1 whose language instantiates gender, does L1-L2 overlap in the gender system affect the degree to which the learners show native-like effects of prenominal gender-marking?

#### Methods

Three groups of participants were tested: 16 monolingual Spanish L1 speakers; 16 Italian L1-Spanish L2 learners; 18 English L1-Spanish L2 learners.

- Italian and Spanish have extensive overlap in their grammatical gender systems; English and Spanish have no overlap.
- Spanish proficiency was assessed in L2 groups with a standardized proficiency test and a picture-naming task. This resulted in two proficiency groups for the English L1-Spanish L2 learners: 9 lower proficiency, 9 higher proficiency.

The experiment presented Spanish sentences, auditorily and one-at-a-time, while participants viewed two pictures of common objects on a computer screen. Participants' task was to click the picture mentioned in the sentence and, after each sentence, provide an acceptability judgment on sentence plausibility.

- Clicking kept participants attentive and the acceptability judgment task concealed the real experimental purpose.

Eye movement data were recorded throughout the experiment. The dependent measure was proportion of gaze shifts to pictures (target and distractor).

#### Results

- Monolingual Spanish speakers showed facilitative effects of prenominal gender-marking for processing of the subsequent noun.
- Effects of prenominal gender-marking for Spanish noun processing in the English L1-Spanish L2 learner group depended on proficiency.

- Higher-proficiency English-Spanish learners were similar to the Spanish monolinguals.
- Lower-proficiency learners did not show sensitivity to feminine noun targets and showed, contrary to expectations, facilitation effects when both targets were masculine (facilitation is normally found when targets differ in gender).
- Italian L1-Spanish L2 learners showed sensitivity to feminine noun contexts, similar to the Spanish monolinguals and higher-proficiency English-Spanish learners. No effects were found for masculine noun contexts.

### Conclusions

The study provides further evidence that native-like processing of gender can be achieved by learners whose L1 does not instantiate this feature, at least for higher-proficiency learners. Further, the study shows that for L2 learners with a similar L1 gender system, native-like processing is not guaranteed.

While this paradigm is versatile, studies employing it are limited to relatively concrete aspects of language, as the linguistic information has to relate to the visual display. Additionally, the necessary relationship between the visual and linguistic information, and the fact that they are temporally aligned, means that language is likely processed differently in VW paradigms than in other paradigms lacking visual information. Relatedly, the eye movements can reveal what kinds of visual information listeners use, but only within the highly constrained options available. Nevertheless, humans often do process language with some sort of visual scene, which makes this paradigm ecologically appealing. Overall, the VW paradigm is, as Huettig et al. (2011) note, a useful way to investigate the many cognitive processes, including language, vision, and attention, that are implicated in language processing but rarely studied together.

## Language Learning Paradigms

Within psycholinguistics, language learning paradigms involve training a group of participants on a specific aspect of language or a language model over a relatively short period of time (i.e., minutes, hours, or days). Unlike more naturalistic language learning, which takes years, these paradigms allow researchers to examine learning at a smaller scale and in a highly controlled way. There are five main types of systems commonly used in language learning paradigms: novel word sets, statistically governed speech streams (word segmentation), artificial grammars, artificial languages, and mini-languages. Figure 14.5 depicts sample experimental designs for these language learning paradigms.



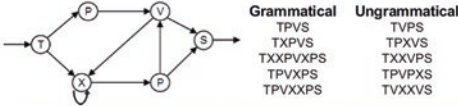
**Novel Word Learning**

Participants can be trained on unknown words with pictures, translations/definitions, or within sets of highly constrained sentences, as in the examples from Batterink & Neville (2011) shown below.

*Several white fluffy meeves spotted the clear blue sky.  
Thunder rumbled and low gray meeves gathered over the horizon.*

**Artificial Grammars**

Shown below is an example of an artificial grammar rule system and some grammatical and ungrammatical strings.



**Speech Segmentation**

The example below from Saffran et al. (1996), demonstrates how probabilities between syllables differ within and across nonwords.

**Speech stream:** bidakupadotigolabubidaku  
**Nonwords:** bidaku | padoti | golabu | bidaku

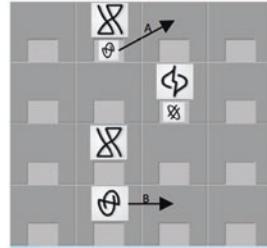
**Transitional probabilities between syllables:**



**Artificial Languages**

Artificial languages consist of words and at least some grammatical structures that are novel to participants. In Brocanto2 (Morgan-Short et al., 2010), participants hear sentences that correspond to board game moves.

Syntactic Structures	Word Categories
S	NP VP NP NP VP
	N (nouns) pleck <sub>masc</sub> ♁ necp <sub>masc</sub> ♁ vode <sub>fem</sub> ♀ blom <sub>fem</sub> ♁
NP	N d (articles) l-i <sub>masc</sub> -u <sub>fem</sub>
VP	v v m M (adjectives) trois -c <sub>masc</sub> -o <sub>fem</sub> = round neim -c <sub>masc</sub> -o <sub>fem</sub> = square m (adverbs) noyka = vertically zayma = horizontally V (verbs) praz = swap klin = move yab = release nim = capture

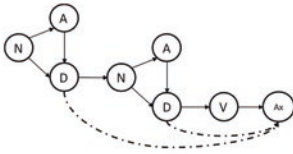


The following sentences describe moves A & B:  
A: Neep neime li vode neimo lu yab zayma.  
Neep square-m<sub>masc</sub> vode square-f<sub>fem</sub> release horizontally.  
The square neep releases the square vode.  
A: Vode neimo lu klin zayma.  
Vode square-f<sub>fem</sub> moves horizontally  
The square vode moves horizontally.

Note: NP = noun phrase, VP = verb phrase, *masc/fem* indicate masculine and feminine grammatical gender

**Mini Languages**

Mini languages, like Mini-Basque (Tagarelli, 2014) contain real words and sentences from natural languages.



Note: N = noun, A = adjective, D = determiner, V = main verb, Ax = auxiliary verb; dashed lines indicate polypersonal agreement between determiners and auxiliary verb

Hartzek zaldi urdina bultzatu dute.  
Bear(S)-DET(pl/erg) have(S) blue(ad) (S)-DET(erg/acc) pushed(V) horse(S)  
The bears have pushed the blue horse.

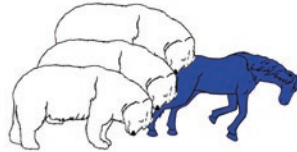


Fig. 14.5 Artificial linguistic systems in language learning paradigms (based on Morgan-Short et al., 2010; Saffran et al., 1996; Tagarelli, 2014). Note: Drawing credit: Kyle Brimacombe

## Novel Word Learning

Word learning has been regularly examined in experimental settings since the 1970s, when Carey and Bartlett (1978, p. 17) coined the term “fast mapping.” This term refers to the phenomenon by which individuals can learn the meaning of a new word after as little as one exposure to the word with its meaning. Most word learning paradigms involve a short training period in which learners are presented (visually or aurally) with each new word in a set and its meaning, as an image or translation equivalent (e.g., Breitenstein et al., 2007). More active training paradigms include forced-choice matching, whereby participants are presented with a novel word and multiple concepts or translations and must choose the correct match (Tagarelli, 2014). Participants can also be trained on words in sentential contexts, with pictures or in highly constrained sentences (e.g., Batterink & Neville, 2011). These approaches inform lexical-level learning, but language acquisition also involves recognizing word boundaries and learning grammatical structures, which other paradigms can elucidate.

## Speech Segmentation

An interesting challenge in language acquisition is that speech is a continuous stream that listeners must break up into words during comprehension—we are not presented with individual words in conversation. Speech segmentation paradigms are designed to investigate how humans parse this noisy speech stream, from infancy (Saffran, Aslin, & Newport, 1996) through adulthood (Karuza et al., 2013). Such paradigms typically present learners with continuous speech composed of randomly placed multisyllabic nonwords (e.g., *pabikugolatudaropitibudo* is a stream of the nonwords *pabiku*, *golatu*, *daropi*, *tibudo*). Transitional probabilities of syllables within a word are high and transitional probabilities across word boundaries are lower because of the random word order. Using the example above, “pa” is followed by “bi” 100% of the time, but “ku” has a 25% chance of being followed by either “pa,” “go,” “da,” or “ti” (see Fig. 14.5). After exposure of as little as two minutes, listeners distinguish words from nonwords or partwords (e.g., *bikugo*) with greater-than-chance ability.

Until recently, speech segmentation paradigms focused on single-language input. However, statistical regularities in speech vary across languages, so bilingual speakers or L2 learners must form two distinct statistical representations in order to successfully segment their two languages. Recent research has

begun to explore such issues (Weiss, Poepsel, & Gerfen, 2015), and further research will continue to reveal how bilinguals and L2 speakers learn and represent multiple sets of statistical regularities.

### **Artificial Grammars**

Artificial grammars (Reber, 1967) are structured systems of elements (e.g., letters) whose order is determined by complex rules (see Fig. 14.5). Although artificial grammars look different from natural languages on the surface, the relationships between elements in the grammars tend to rely on structural dependencies similar to those found in natural languages. Indeed, artificial grammar processing has been shown to elicit similar neural responses as natural language syntactic processing (Bahlmann, Gunter, & Friederici, 2006).

Artificial grammar research has demonstrated that complex structural information can be learned under both implicit and explicit training, though there is a longstanding debate regarding whether the acquired knowledge is implicit (Reber, 1990) or explicit (Perruchet & Pacteau, 1990). In reality, it is probably a bit of both. Note also that learning outcomes following implicit training seem to be related to different underlying cognitive abilities for artificial grammar and natural language learning (Robinson, 2005), which has implications for the generalizability of this paradigm to questions in applied linguistics.

### **Artificial, Semi-Artificial, and Mini-Languages**

Word learning and grammar learning paradigms allow researchers to develop detailed knowledge about these aspects of language in isolation, which is impossible to do with natural languages. However, natural language combines lexical semantics and grammar, as well as phonology, prosody, and discourse-level information. For this reason, it is useful to examine how people learn languages using paradigms that more closely approximate natural languages. Over the past 25 years, there has been an explosion of applied linguistics research using paradigms—namely, artificial, semi-artificial, and mini-languages—that do just that (see Fig. 14.5).

Artificial and semi-artificial languages are typically composed of a few grammatical structures that are consistent with natural language structures. Also like natural languages, they can be spoken and understood. Semi-artificial languages contain L1 lexical items and grammatical structures borrowed from

another language. For example, the sentence in (3) uses English words with German word-order rules (Rebuschat & Williams, 2012).

(3) When his wife in the afternoon the office left, prepared Jim dinner for the entire family.

Because semi-artificial languages use L1 vocabulary, they are easy to understand, which frees up cognitive resources for grammar learning. However, this design might also be confusing for learners, who need to inhibit what they know about their L1 grammar in the face of a new grammatical system associated with known L1 words. Artificial languages, on the other hand, contain novel words (typically between 10 and 100) in addition to grammatical rules. Learners can be trained on (semi-)artificial languages within hours, allowing researchers to more quickly examine how learners acquire target grammatical structures such as word order (Rebuschat & Williams, 2012), case marking (Grey, Williams, & Rebuschat, 2014), and morphological agreement (Morgan-Short, Sanz, Steinhauer, & Ullman, 2010). Notably, artificial language learning has been shown to correlate with L2 abilities (Ettliger, Morgan-Short, Faretta-Stutenberg, & Wong, 2016), which adds ecological validity to using artificial languages for informing questions in SLA.

A mini-language is an actual subset of a real, natural language. Mini-languages maintain the grammar, lexicon, and phonotactics of their source languages, and are comprised of sentences that native speakers of the source language would understand. When designing a mini-language, it is important to choose a source language that naturally contains the linguistic features of interest. For example, Mini-Nihongo, a miniature version of Japanese, includes variable word order, case marking, and classifiers. The first two features exist in German, but classifiers do not, so this mini-language allowed the researchers to examine L2 processing of familiar and unfamiliar linguistic targets for L1 German speakers (Mueller, Hahne, Fujii, & Friederici, 2005). In another study, Mini-Basque, derived from Basque, was used to test L2 acquisition of entirely unfamiliar linguistic targets, so it was important that it have a lexicon and grammatical features that were new to the L1 English speakers learning the mini-language (Tagarelli, 2014). Because mini-languages rely heavily on natural languages but focus on small, controlled aspects of those languages, they serve as a way to bridge the gap between artificial and natural languages in SLA research. Indeed, by using ERP (Mueller et al., 2005) and fMRI (Tagarelli, 2014) techniques, research has demonstrated that mini-language learning and processing relies on many of the same neural correlates as L1 and L2 processing.

While these learning paradigms approximate natural languages in many ways, they still employ very small language models that cannot fully generalize to the enormity of natural language learning. Additionally, while there has been extensive evidence of quick learning in a variety of training conditions, studies rarely include untrained controls, which makes it difficult to assess whether high performance is a result of training, or rather a testing effect (for discussion, see Hamrick & Sachs, 2017). Nevertheless, behavioral and neural evidence support their efficacy in informing the processes and outcomes involved in language learning.

## Conclusions

In this chapter, we have reviewed a diverse body of psycholinguistic methods that can be utilized to inform work in applied linguistics. We have discussed well-established tasks, measures, and paradigms that provide researchers with detailed information about the underlying psychological (and neural) representations and processes involved in language learning and processing. Throughout, we considered the appeal of these methods to researchers as well as their limitations.

Future research will continue to benefit from methodological bridges across psycholinguistics and applied linguistics. For instance, debates on explicit/implicit L2 learning and knowledge may be elucidated by syntactic priming research in L1 (Fine & Jaeger, 2013), and applied interests regarding L2 learning in the digital age can also be informed by psycholinguistic approaches (Lan, Fang, Legault, & Li, 2015). We hope we have illustrated that merging psycholinguistic methods with interests in applied linguistics opens new avenues of research and offers novel insights into classic questions.

## Resources for Further Reading

Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

This book provides a thorough foundation in the science behind the ERP technique and important methodological considerations for its use in research. The text is appropriate for novice through advanced ERP researchers and is a critical resource for scholars in applied linguistics who are interested in incorporating electrophysiological methods to answer their research questions.

McDonough, K., & Trofimovich, P. (2009). *Using priming methods in second language research*. New York: Taylor & Francis.

This book provides detailed information on theory, experimental design, and data analysis for priming paradigms in second language processing and acquisition research. It is a comprehensive resource for independent researchers and is also appropriate as a text for courses on research methods, second language acquisition, and psycholinguistics.

## References

- Altarriba, J., & Basnight-Brown, D. M. (2007). Methodological considerations in performing semantic- and translation-priming experiments across languages. *Behavior Research Methods*, *39*(1), 1–18.
- Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.
- Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research*, *32*(1), 25–36.
- Bahlmann, J., Gunter, T. C., & Friederici, A. D. (2006). Hierarchical and linear sequence processing: An electrophysiological exploration of two different grammar types. *Journal of Cognitive Neuroscience*, *18*(11), 1829–1842.
- Balukas, C., & Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, *19*(4), 423–443.
- Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53.
- Bartolotti, J., & Marian, V. (2012). Language learning and control in monolinguals and bilinguals. *Cognitive Science*, *36*(6), 1129–1147.
- Basnight-Brown, D. M., & Altarriba, J. (2007). Differences in semantic and translation priming across languages: The role of language direction and language dominance. *Memory & Cognition*, *35*(5), 953–965.
- Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., & Pléh, C. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, *10*(2), 344–380.
- Batterink, L., & Neville, H. (2011). Implicit and explicit mechanisms of word learning in a narrative context: An event-related potential study. *Journal of Cognitive Neuroscience*, *23*(11), 3181–3196.
- Batterink, L., & Neville, H. (2013). Implicit and explicit second language training recruit common neural mechanisms for syntactic processing. *Journal of Cognitive Neuroscience*, *25*(6), 936–951.



- Bisson, M.-J., Van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2012). Processing of native and foreign language subtitles in films: An eye-tracking study. *Applied Psycholinguistics*, 35(2), 399–418.
- Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3), 355–387.
- Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2), 123–147.
- Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014). Native ‘um’s elicit prediction of low-frequency referents, but non-native ‘um’s do not. *Journal of Memory and Language*, 75, 104–116.
- Bowden, H. W. (2016). Assessing second language oral proficiency for research. *Studies in Second Language Acquisition*, 38(4), 647–675.
- Bowden, H. W., Steinhauer, K., Sanz, C., & Ullman, M. T. (2013). Native-like brain processing of syntax can be attained by university foreign language learners. *Neuropsychologia*, 51(13), 2492–2511.
- Breitenstein, C., Zwitserlood, P., de Vries, M. H., Feldhues, C., Knecht, S., & Dobel, C. (2007). Five days versus a lifetime: Intense associative vocabulary training generates lexically integrated words. *Restorative Neurology and Neuroscience*, 25(5&6), 493–500.
- Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, 15, 17–29.
- Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS ONE*, 6(5), e19792.
- Crinion, J., Turner, R., Grogan, A., Hanakawa, T., Noppeney, U., Devlin, J. T., ... Usui, K. (2006). Language control in the bilingual brain. *Science*, 312(5779), 1537–1540.
- Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 498.
- DeLong, K. A., Groppe, D. M., Urbach, T. P., & Kutas, M. (2012). Thinking ahead or not? Natural aging and anticipation during reading. *Brain and Language*, 121(3), 226–239.
- Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2017). MultiPic: A standardized set of 750 drawings with norms for six European languages. *The Quarterly Journal of Experimental Psychology*, 22, 1–24.
- Dussias, P. E., Kroff, J. R. V., Tamargo, R. E. G., & Gerfen, C. (2013). When gender and looking go hand in hand. *Studies in Second Language Acquisition*, 35(2), 353.
- Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, 27(2), 141–172.
- Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27(3), 464–491.

- Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. (2016). The relationship between artificial and second language learning. *Cognitive Science*, 40(4), 822–847.
- Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4), 469–495.
- Fine, A. B., & Jaeger, F. T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, 37(3), 578–591.
- Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied Psycholinguistics*, 24(03), 369–383.
- Flege, J. E. (1991). Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *The Journal of the Acoustical Society of America*, 89(1), 395–411.
- Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6), 709–738.
- Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, 42(1), 226–241.
- French-Mestre, C., & Prince, P. (1997). Second language autonomy. *Journal of Memory and Language*, 37(4), 481–501.
- Gillon Dowens, M., Guo, T., Guo, J., Barber, H., & Carreiras, M. (2011). Gender and number processing in Chinese learners of Spanish—Evidence from event related potentials. *Neuropsychologia*, 49, 1651–1659.
- Godfroid, A., Housen, A., & Boers, F. (2013). An eye for words: Gauging the role of attention in L2 vocabulary acquisition by means of eye tracking. *Studies in Second Language Acquisition*, 35(3), 483–517.
- Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language*, 122(2), 92–102.
- Grey, S., Cox, J. C., Serafini, E. J., & Sanz, C. (2015). The role of individual differences in the study abroad context: Cognitive capacity and language development during short-term intensive language exposure. *Modern Language Journal*, 99(1), 137–157.
- Grey, S., & Van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of sentence comprehension. *Journal of Neurolinguistics*, 42, 93–108.
- Grey, S., Williams, J. N., & Rebuschat, P. (2014). Incidental exposure and L3 learning of morphosyntax. *Studies in Second Language Acquisition*, 36(4), 611–645.
- Grodner, D., & Sedivy, J. (2011). The effect of speaker-specific information on pragmatic inferences. *The processing and acquisition of reference*, 2327, 239–272.
- Hamrick, P., & Sachs, R. (2017). Establishing evidence of learning in experiments employing artificial linguistic systems. *Studies in Second Language Acquisition*, 1–17.
- Harley, T. A. (2013). *The psychology of language: From data to theory*. East Sussex: Psychology Press.



- Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science, 15*(6), 409–414.
- Huetig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151–171.
- Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic reanalysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience, 15*, 98–110.
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133–156.
- Karuz, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language, 127*(1), 46–54.
- Kootstra, G. J., van Hell, J. G., & Dijkstra, T. (2010). Syntactic alignment and shared word order in code-switched sentence production: Evidence from bilingual monologue and dialogue. *Journal of Memory and Language, 63*(2), 210–231.
- Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The revised hierarchical model: A critical review and assessment. *Bilingualism: Language and Cognition, 13*(03), 373–381.
- Lan, Y.-J., Fang, S.-Y., Legault, J., & Li, P. (2015). Second language acquisition of Mandarin Chinese vocabulary: Context of learning effects. *Educational Technology Research and Development, 63*(5), 671–690.
- Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384–422.
- Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, MA: MIT press.
- Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language, 69*(4), 574–588.
- McDonough, K., & Mackey, A. (2008). Syntactic priming and ESL question development. *Studies in Second Language Acquisition, 30*(1), 31–47.
- McDonough, K., & Trofimovich, P. (2009). *Using priming methods in second language research*. New York, NY: Taylor & Francis.
- McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Reviews Neuroscience, 7*, 703–704.
- Morett, L. M., & Macwhinney, B. (2013). Syntactic transfer in English-speaking Spanish learners. *Bilingualism: Language and Cognition, 16*(1), 132–151.
- Morgan-Short, K. (2014). Electrophysiological approaches to understanding second language acquisition: A field reaching its potential. *Annual Review of Applied Linguistics, 34*, 15–36.

- Morgan-Short, K., Sanz, C., Steinhauer, K., & Ullman, M. T. (2010). Second language acquisition of gender agreement in explicit and implicit training conditions: An event-related potential study. *Language Learning, 60*(1), 154–193.
- Mueller, J. L., Hahne, A., Fujii, Y., & Friederici, A. D. (2005). Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. *Journal of Cognitive Neuroscience, 17*(8), 1229–1244.
- Neville, H. J., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience, 3*(2), 151–165.
- Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. T. (2012). The influence of language proficiency on lexical-semantic processing in native and late learners of English. *Journal of Cognitive Neuroscience, 24*, 1205–1223.
- Newman, R. L., & Connolly, J. F. (2009). Electrophysiological markers of pre-lexical speech processing: Evidence for bottom-up and top-down effects on spoken word processing. *Biological Psychology, 80*(1), 114–121.
- Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech & Language, 9*(1), 19–35.
- Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General, 119*(3), 264.
- Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin, 134*(3), 427.
- Rassaei, E., Moïnzadeh, A., & Youhannaee, M. (2012). Effects of recasts and meta-linguistic corrective feedback on the acquisition of implicit and explicit L2 knowledge. *The Journal of Language Teaching and Learning, 2*(1), 59–75.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855–863.
- Reber, A. S. (1990). On the primacy of the implicit: Comment on Perruchet and Pacteau. *Journal of Experimental Psychology, 119*(3), 340–342.
- Rebuschat, P., & Williams, J. N. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics, 33*(4), 829–856.
- Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition, 19*(02), 223–247.
- Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition, 27*(2), 235–268.
- Roche, J. M., Peters, B., & Dale, R. (2015). “Your tone says it all”: The processing and interpretation of affective language. *Speech Communication, 66*, 47–64.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

- Sanz, C., & Grey, S. (2015). Effects of conditions on L2 development: Beyond accuracy. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 301–324). Amsterdam: John Benjamins.
- Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*, 369–369.
- Spada, N., Shiu, J. L. J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*(3), 723–751.
- Stafford, C. A., Bowden, H. W., & Sanz, C. (2012). Optimizing language instruction: Matters of explicitness, practice, and cue learning. *Language Learning*, *62*(3), 741–768.
- Sunderman, G., & Kroll, J. F. (2009). When study abroad fails to deliver: The internal resources threshold effect. *Applied Psycholinguistics*, *30*, 79–99.
- Tagarelli, K. M. (2014). *The neurocognition of adult second language learning: An fMRI study*. (Doctoral Dissertation), Georgetown University.
- Tagarelli, K. M., Ruiz, S., Moreno Vega, J. L., & Rebuschat, P. (2016). Variability in second language learning: The roles of individual differences, learning conditions, and linguistic complexity. *Studies in Second Language Acquisition*, *38*(2), 293–316.
- Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morpho-syntactic processing. *Neuropsychologia*, *56*, 289–301.
- Weiss, D. J., Poepsel, T., & Gerfen, C. (2015). Tracking multiple inputs. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 167–190). Amsterdam: John Benjamins.
- Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*(4), 680–704.
- Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497–528.



# 15

## SLA Elicitation Tasks

Susan Gass

### Introduction

Ways of eliciting data to understand (1) how languages are learned and (2) the type and extent of second language<sup>1</sup> (L2) knowledge are limited to some extent only by the imagination of the researcher. This does not imply that one can use elicitation techniques randomly because a basic premise of solid research is to match the elicitation task to what is being investigated.<sup>2</sup> When determining an appropriate elicitation task, it is crucial, therefore, that the task itself measure the construct under investigation. It is also important to note that no construct has a unique elicitation technique associated with it. There are generally multiple approaches, and in fact, it might be appropriate, if not more revealing, to triangulate data from multiple elicitation techniques to investigate a particular construct.

This chapter reviews two common elicitation tasks: judgments and elicited imitation. Each section will deal with the historical use, conceptual motivations, and practical aspects of the technique, including benefits and limitations. Both elicitation measures have had a long history of use, most likely due to a number of common advantages of data collection and analysis. For example, both are relatively easy to administer, and this ease has continued from days of paper and pencil administration

---

S. Gass (✉)

English Language Center, Michigan State University, East Lansing, MI, USA  
e-mail: [gass@msu.edu](mailto:gass@msu.edu)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_15](https://doi.org/10.1057/978-1-137-59900-1_15)

(judgments) and/or tape recorders for delivery and data capturing (elicited imitation) to more sophisticated measures of computer delivery. Second, because of the standard mode of administration, a large amount of data can be captured from participants in a relatively short amount of time and comparisons across populations can be made. Finally and perhaps most important is the fact that other types of data, in particular spontaneous production data, do not reliably yield a sufficient number of exemplars of a particular linguistic structure necessary for analysis and interpretation; both judgment data and elicited imitation data can fill that gap by targeting a specific linguistic structure.

## Judgment Tasks

In this section, I deal with four types of judgment tasks: (1) grammaticality judgments, (2) magnitude estimation, (3) preference judgments, and (4) truth-value judgments.

### Grammaticality Judgments

The most common judgment task is what is generally referred to as grammaticality judgments, although the more appropriate term, acceptability judgments, is used as well.<sup>3</sup> Generally, elicitation, in its simplest form, takes place by asking individuals to judge something (usually a sentence) as being either grammatical or not. What is inferred from learner responses is the nature of their interlanguage (Selinker, 1972) grammar. In other words, if they say that a given sentence is grammatical, researchers infer that that sentence and the different linguistic features therein are part of the learners' second language (L2) grammar. An example of a grammaticality judgment task with a brief instruction is seen in (1). Instructions can vary in many respects, something we turn to later.

#### 1. Example of task

Instruction: You will see sentences and your task is to judge each one as to whether it is a grammatical sentence or not (or correct or not). Indicate with a C for *correct* or NC for *not correct*; if it is not correct, indicate what needs to be changed to make it correct.

- a. Last year the man has flown to Europe three times.

Response: NC

flew

Last year the man ~~has flown~~ to Europe three times.

- b. Last year the man flew to Europe three times.

Response: C

Grammaticality judgments represent one of the earliest forms of data elicitation in L2 research, in large part, due to the early dependence of the field on linguistically-focused research where understanding grammars of the world's languages was often determined by asking speakers of those languages whether specific sentences are acceptable or not. They have remained a cornerstone of empirical work in linguistics and in second language acquisition (SLA).

The adoption of grammaticality judgments into L2 research has often been without scrutiny and without the recognition of what such judgments actually reflect. Making judgments about what is and is not possible in one's native language (NL) is quite different from asking about the target language (TL), the language being learned. In the former case, one is asking about the NL and inferences are made about the grammar of that language. In the case of second language research, a third language comes into the picture, namely, the interlanguage (Selinker, 1972). One asks about judgments of the second language (the target language) and then makes inferences about another system (the interlanguage). Thus, the data elicited in L2 research must be interpreted with caution and with the understanding that the relationship between judgment and language may be even more circumspect than it is when conducting research on native languages.

Over the years, perhaps prompted by some of these issues, there has been considerable debate about grammaticality judgment data. Selinker (1972) was an early researcher to cast doubt on the reasonableness of using such data in understanding what learners know and how they use language. He explicitly claimed that judgment data were inappropriate for second language research: because researchers "will gain information about another system, the one the learner is struggling with, i.e., the TL" (1972, p. 213). (See Gass & Polio, 2014, for further discussion on this issue.)

In the 1960s, 1970s, and beyond, there were arguments on both sides of the controversy. Nonetheless, against the backdrop of these arguments, grammaticality judgments continue to be an important data source to measure knowledge and learning. During the early years of L2 research, researchers often felt the need to justify the use of grammaticality judgments. In today's research world, justifications occur with much less frequency, suggesting more or less acceptance as a standard elicitation technique.

The restriction of L2 data to produced utterances, as suggested by Selinker, went against the view of those with traditional linguistic training. For example, Schachter, Tyson, and Diffley (1976) emphasized the need to characterize “the learner’s interlanguage, at different stages, in the acquisition of the target language” (p. 67). Most important was their argument that “[w]e are interested in characterizing learner *knowledge* of his language not simply learner production” (p. 67). Results from their study showed that learners could distinguish between sentences about which they had no intuitions and those about which they did have intuitions, the former the authors referred to as indeterminate sentences. Thus, through the use of judgment data, a theoretical construct of indeterminacy emerged.

Other early advocates for the value of judgment data were Hyltenstam (1977) and Gass (1979). Both were proponents of validation via multiple data sources. In the case of Hyltenstam’s research, the main issue was the need for large sample sizes and, importantly, the need to understand “the linguistic competence of the learner” (p. 385). The problem, then, was how to collect a large enough sample without forced production. Hyltenstam (and Corder, 1973) had made the argument for sequential data sources, whereas Gass used grammaticality judgments collected simultaneously with other types of data with each data source used to either support findings of other data sources or recognize the role of task differences on outcomes.

As noted earlier, controversy has always surrounded the use of this elicitation tool and continues to do so today (cf. review of metalinguistic judgments by Chaudron, 1983). There are many reasons for this, many of which will be discussed below. To this point, the implementation of grammaticality judgments can vary across a number of features, each of which may also impact study outcomes: (1) number of sentences, (2) timed versus untimed, (3) modality, (4) scale used, (5) corrections, (6) instructions, and (7) scoring. But, despite the fact that opinions are strong and researchers are adamant on both sides (e.g., Cowart, 1997; Marantz, 2005; Myers, 2009; Schütze, 1996), they are still a common source of data in numerous domains of L2 research. The reader is referred to Spinner and Gass ([forthcoming](#)) and Marsden, Plonsky, Gass, Spinner, and Crowther (2016) for more complete coverage of issues related to the use and interpretation of grammaticality judgment data. The latter study provides a synthesis of approximately 400 studies that used grammaticality judgments, focusing on a variety of parameters such as mode, timing, and response types, as well as questions asked and theoretical frameworks used.<sup>4</sup>



## What Do Judgment Data Reflect?

Because it is crucial to match technique to construct, any discussion of grammaticality judgments must include an understanding of what grammaticality judgment data reflect. One of the main controversies in the field of SLA concerns knowledge types, namely, what is referred to as implicit and explicit knowledge. Implicit knowledge of language is that knowledge that we are not aware of, whereas explicit knowledge reflects knowledge that we are aware of (see Ellis, 2005 and Bowles, 2011) for further distinctions between these two constructs). Numerous studies have considered the role of grammaticality judgments in this debate and have focused on modes of elicitation, in particular whether there is a time limit for a response or not. Consider a sentence such as “Yesterday the man walk from his house to his office,” a typical utterance of a second language learner. If this sentence is presented to a learner, she/he might have explicit knowledge of the ungrammaticality of the sentence and so indicate that on a response sheet. Yet, that same learner might utter sentences with lack of past tense agreement. Thus, one must question whether the response truly represents that individual’s underlying knowledge of English, and, if so, how does one reconcile the difference between response and production?

A number of studies have addressed this issue (e.g., Ellis, 2005; Gutiérrez, 2013; Ionin & Zyzik, 2014; Vafaei, Suzuki, & Kachisnke, 2016). The central questions are: What is the underlying construct that grammaticality judgments measure? Does including a time limit on responses alter the type of knowledge being reflected? One possible conclusion from these validation studies is that time pressure alters the knowledge reflected (Ellis, 2005); another is that time pressure does not alter the underlying knowledge source. Rather, grammaticality judgments reflect different levels of explicit knowledge (Vafaei et al., 2016). Sample Study 15.1 provides a description of a study that used both grammaticality judgment tests (GJTs) and a production test to verify a theory prevalent in SLA.

### Sample Study 15.1

Spinner, P. (2013). Language production and perception: A Processability Theory study. *Language Learning*, 63, 704–739.

#### Background

Processability Theory (PT) (Pienemann 1998, 2005) claims that there is a predictable order of emergence of morphosyntactic forms in L2 acquisition. This finding has been documented for production data, but has not been explored for reception data.



### Research Question

To what extent are the PT procedures, which have been empirically well established for production, shared by the receptive system?

### Participants

#### Study 1

- Number: 51
- Native Language: Asian, African, Slavic, and Romance languages
- Target Language: English
- Control: 40 native speakers (NSs) of English
- Proficiency Level: Not lowest level in a community ESL program

#### Study 2

- Number: 12 (subset of Study 1)
- Native Language: Chinese, Korean, Arabic, Spanish, Portuguese
- Target Language: English

#### Study 3

- Number: 63

### Instruments

Three separate studies, two using GJTs and one using a production test

### GJTs

#### Study 1 (Audio)

#### Materials

- 150 sentences (75 grammatical and 75 ungrammatical, each 6–8 words in length).
- 15 different grammatical structures representing 5 different predicted stages of acquisition: 3 representing stages 2, 3, 4, and 5, and 2 representing stage 6.
- Two versions were created with individuals taking one version or the other. No one saw both the grammatical and the ungrammatical version of a sentence.

#### Instructions

- In this test, you will hear English sentences. Some sentences have correct English grammar and some sentences have incorrect English grammar. If a sentence has correct grammar, check *correct* on the sheet. If a sentence has incorrect grammar, check *incorrect* on the sheet. If you don't know or are not sure, check *I don't know*.

#### Administration

- Items presented aurally (timed)
- After 25 items (each sentence was 6–8 words in length), 1-minute break
- After 75 items, longer break

#### Study 3

- Similar to Study 1
- *Materials*: As with Study 1 with six sentences per structure and with some structures eliminated. Only one version.

### Other Instrument

#### *Study 2*

Production data gathered from one-to-one conversation with advanced NNS of English with specific tasks designed to elicit specific PT structures.

#### Statistical Tools

- Study 1: Implicational Scaling
- Study 2: Implicational Scaling
- Study 3: Implicational Scaling

#### Results

For the productive test (Study 2), the PT predicted order was verified. For the two receptive tests (Studies 1 and 3), the predicted order of emergence was not found. Spinner concludes that “the acquisition of processing procedures may proceed differently in production than in reception” (p. 734).

## Preparing a GJT

The *sine qua non* of any study using a particular elicitation tool is that the research question is such that that tool is appropriate and can provide an answer to the research question. Grammaticality judgments are no exception, but in addition, there are numerous factors that one must take into account when employing GJTs in a study.

### *Number of Sentences*

Sentence numbers vary with as many as 240 (Batterink & Neville, 2013), 282 (Johnson & Newport, 1989), or 300 (Montrul, Davidson, De La Fuente, & Foote, 2014). There is no absolute answer to the question of how many. What is important is reliability and the not unrelated issue of participant fatigue. Cowan and Hatasa (1994) recommend 60–72 sentences (see Godfroid et al., 2015, who used 68) for purposes of reliability. Trying to balance between the need to avoid fatigue while maintaining reliability, Gass and Mackey (2007) recommend between 50 and 60. However, there are times when more sentences are needed because (1) one has to include fillers so that the particular structure being examined will not be recognized or (2) many structures are being examined (Spinner, 2013: see Sample Study 15.1). The latter situation may obviate the need for fillers because one structure serves as a “filler” for others. Given the need for reliability, the need to avoid fatigue becomes a serious issue. Taking this into account, one could make multiple versions so that each person responds to a subset, or the researcher could provide breaks between blocks or groups of sentences.

### *Timing*

Should there be a time limit for responses, and if so, how much? In previous studies, timed GJTs have been taken to reflect implicit knowledge, whereas untimed GJTs have been seen to reflect explicit knowledge (see discussion of controversies about this interpretation in Gutiérrez, 2013; Vafae et al., 2016). Once one has determined that a timed GJT is appropriate, the important question is how much time should be granted. There is no definitive answer. However, most studies do some variation of average NS response time. For example, Vafae et al. (2016), in a computer-based study, left sentences on the screen 1.2 times as long as NSs took to respond to the sentences. Godfroid et al. (2015) added 20% to NS judgments, as did Ellis (2005) and Loewen (2009).

### *Modality*

Generally, GJTs are given in written format, either via computer (e.g., Gass & Alvarez Torres, 2005; Vafae et al., 2016) or in a paper and pencil format (see Murphy, 1997, for a review). Sophistication of computer delivery for purposes of timing of delivery of responses has allowed researchers to focus more on computer delivery of GJTs and reduce the use of paper and pencil tests. Examples abound of GJTs given aurally (e.g., Johnson & Newport, 1989; Spinner, 2013). Aurally delivered stimuli are advisable if one wants a quick response from learners because there is little time for them to ponder and draw on something “a teacher might have said.” On the other hand, aurally delivered stimuli have the disadvantage that participants must be matched on listening abilities; the oral mode also makes it more difficult to include follow-up tasks such as identifying the location of the error in ungrammatical items. These topics will be dealt with below when we address the elicitation tool referred to as *elicited imitation*.

### *Scale*

There are many ways to construct a scale for GJTs. In some studies, there is a dichotomous judgment asked of participants (sentence is good or not). However, most studies use a Likert scale ranging from 4 to 7 points. Two important issues need to be considered: what to name the points on a scale and whether there is a midpoint/not sure category. Often scales have numerical values, with some having a negative value; others have words, such as *completely acceptable/completely unacceptable* (Yuan & Dugarova, 2012) or

*impossible, probably impossible, probably possible, perfectly possible* (Montrul, Dias, & Santos, 2011).

There are a number of questions that need to be addressed: When there is an odd number and scales go from a minus to a plus value, what does zero stand for? One needs to carefully consider how participants will interpret that option. Does it mean that the respondent is truly on the fence and has no idea or does it mean that the respondent believes that it is one of those sentences that could be acceptable in some situations, but not in others? Instead of a zero value, some label the scale with a midpoint as *I don't know*. But what does the *I don't know* option mean? Does one even want an *I don't know* option? All of these possibilities and many others are found in the literature, but each one must be carefully considered when designing a study using judgment data. One possible solution is to use a separate *I don't know* option and/or a scale with an even number of points. When selecting this possibility, one has to make sure that individuals do not overuse the *I don't know* choice and, if so (to be determined statistically), eliminate those participants' responses from analysis. One final consideration has to do with analysis and the need to keep in mind that non-parametric tests might be most appropriate. Even when there is a numerical scale that appears to be an interval scale, the scale itself may not actually be an interval scale. That is, the distance between points may not be equidistant. One way to create an interval judgment task is through magnitude estimation, to be discussed below.

Confidence ratings are another parameter that some studies use although they are not always analyzed (see Godfroid et al., 2015). Rebuschat (2013) provides an overview of confidence ratings with guidelines on how to use them. In his discussion, he addresses the limitations of participant bias as well as a lack of understanding of what the scale means. Each participant has his/her own "criterion for reporting knowledge: More conservative subjects may indicate that they are guessing on their grammaticality judgments unless they are absolutely sure, while more liberal subjects may consistently report high levels of confidence at the slightest intuition" (p. 610).

### *Corrections*

Corrections to sentences judged ungrammatical are important to an appropriate interpretation of results. If one is focusing on a specific grammatical category, one must be sure that responses address the acceptability of that category. For example, if one is investigating relative clauses and particularly the use of resumptive pronouns, one might present learners with the sentence in (2):

2. That's the man whom I went to school with him.

The researcher is targeting *him*, but the learner may focus on the difference between *who* and *whom* and may mark this sentence incorrect because she believes *whom* should be *who* and further believes that *with him* is acceptable. In sum, we need to have an understanding whether or not the researcher's focus and the learners' focus are the same.

When and how to make corrections can be a difficult decision. With paper and pencil tests, it is relatively easy because corrections can be made by crossing out the perceived error and the correct form can be inserted in the right place, as in Example 1. With computer delivery or with aurally delivered materials, corrections are more difficult. Gass and Alvarez Torres (2005) devised a way of delivering sentences that need to be corrected using a computer to obtain judgment data and a printout following judgments for purposes of correction (see Sample Study 15.2 for a description). Not asking for corrections, however, can undermine the validity of results.

### Sample Study 15.2

Gass, S., & Alvarez Torres, M. (2005). Attention when? An investigation of the ordering effect of input and interaction. *Studies in Second Language Acquisition*, 27(1), 1–31.

#### Background

This study was designed to test how input and interaction as individual constructs and together led to the learning by English native speakers of three aspects of Spanish: (1) gender agreement, (2) *estar* + locative, and (3) lexical items.

#### Research Hypotheses

1. Given the differential functions of input and interaction, we predict that performance will differ following treatment involving one or the other and performance will also differ following treatment in which the order of presentation differs.
2. Given that interaction is said to be an attention-drawing device, the three experimental groups with interaction will perform better than the group with no interaction.
3. Because input and interaction serve different yet important functions, when there is a combination of conditions (input followed by interaction or interaction followed by input), performance will be better than when only one type of presentation is available.
4. Given Gass's (1997) assumption that interaction serves as a priming device that readies learners to utilize follow-up input, the learners who experience interaction followed by input will perform the best in the experimental measures.
5. Because learners behave differently with regard to different language areas, each of the three language areas investigated will yield distinct results. Moreover, hypotheses 1–4 will be applicable for each language area.

### Participants

- Number: 102
- Proficiency: Enrolled in the third semester of a university-based Spanish class

### Treatment Groups

1. Input only ( $n = 23$ )
2. Interaction only ( $n = 26$ )
3. Input + interaction ( $n = 19$ )
4. Interaction + input ( $n = 18$ )
5. Control ( $n = 16$ )

### Instruments

An acceptability test was used for pre- and posttests. Each test (pretreatment and posttreatment) consisted of 24 sentences: 6 distractors, 6 agreement sentences containing nouns ending  $-o$  or  $-a$ , 6 agreement sentences with nouns ending in consonant or vowel other than  $-o$  or  $-a$ , and 6 sentences with *estar* + locative. Sentences appeared on the screen for a maximum of 18 seconds or until the participant indicated whether the sentence was correct or not (dichotomous choice).

From the pool of 48 sentences, the computer generated randomly half from each category for the pretest. The remainder were used for that individual on the posttest. At the end of the session, all sentences to which the participant had responded “incorrect” were printed out so that corrections could be made.

### Statistical Tool

ANOVA and *t*-tests

### Results

Learning was determined by gain scores from pre- to posttest scores. The group that had interaction and input (regardless of order) had higher gain scores than the control group or either the input or the interaction group. Gender agreement and *estar* + location constructions benefitted from interaction followed by input.

### *Instructions*

As with any data elicitation measure, instructions are an important part of the task. With judgment tasks, it is particularly significant because for most participants, the idea of judging a sentence as “grammatical” is not an everyday occurrence and “grammaticality” may often be confused with something that makes sense (see also Lim & Godfroid, 2015, for issues related to plausibility). To take an example from Lim and Godfroid, the sentence *my mother was cooked by a cake* is semantically anomalous, but grammatically correct. Thus, participants need to be made familiar with the construct of grammaticality and be given examples of what is intended. Bley-Vroman, Felix, and Ioup (1988, p. 2) provide one of the most thorough sets of instructions for participants. These are given below.

### Sample Instructions

Speakers of a language seem to develop a “feel” for what is a possible sentence, even in the many cases where they have never been taught any particular rule.

For example, in Korean you may feel that sentences 1–3 below sound like possible Korean sentences, while sentence 4 doesn’t. [The sentences below were actually presented in Korean.]

1. Young Hee’s eyes are big.
2. Young Hee has big eyes.
3. Young Hee’s book is big.
4. Young Hee has a big book.

Although sentences 2 and 4 are of the same structure, one can judge without depending on any rule that sentence 4 is impossible in Korean.

Likewise, in English, you might feel that the first sentence below sounds like it is a possible English sentence, while the second one does not.

1. John is likely to win the race.
2. John is probably to win the race.

On the following pages is a list of sentences. We want you to tell us for each one whether you think it sounds possible in English. Even native speakers have different intuitions about what is possible. Therefore, these sentences cannot serve the purpose of establishing one’s level of proficiency in English. We want you to concentrate on how you feel about these sentences.

For the following sentences please tell us whether you feel they sound like *possible* sentences of English for you, or whether they sound like *impossible* English sentences for you. Perhaps you have no clear feeling for whether they are possible or not. In this case mark *not sure*.

Read each sentence carefully before you answer. Concentrate on the structure of the sentence. Ignore any problems with spelling, punctuation, and so on. Please mark only one answer for each sentence. Make sure you have answered all 32 questions.

In sum, instructions must be carefully crafted so that participants understand what is being asked for. Otherwise, one risks a mismatch between responses and interpretations.

### Scoring

Scoring will depend on the scale that is used. For example, a dichotomous scale is relatively easy to score. However, there are certain prerequisites that need to be considered. First, if a correction is made to something other than the target item, the response is most likely “correct” even when the participant marked it “incorrect.” Let’s assume a study in which correct responses were given 1 point and incorrect responses 0. If a student marked as incorrect the sentence in (3),

3. That's the man whom I went to school with him.

but changed it to the sentence in (4),

who

4. That's the man ~~whom~~ I went to school with him.

the response should not be given a value of (1) even though the sentence was marked incorrect, because it was incorrect for the wrong reasons. A study by Falk and Bardel (2011) provides an example of a study where correct responses for ungrammatical sentences had to meet two conditions: (1) the sentence was marked *incorrect* and (2) the appropriate correction was made. Second, when scoring, it is important to separate TL grammatical sentences from TL ungrammatical sentences regardless of what scoring decision is to be made.

There are times when partial credit might be appropriate. For example, Gass, Svetics, and Lemelin (2003) gave partial credit when the respondent knew what was wrong, but couldn't make the change. This might be interpreted as incipient knowledge.

More difficult decisions have to be made when a Likert scale is used. If using an even number scale with gradations of confidence (e.g., definitely correct versus correct), does one give a different value to each category, or should the two "correct" categories be collapsed? One should make this decision before collecting data because if collapsing is to take place, there may be no need to present to participants anything more than a dichotomous scale.

With an odd number scale, other decisions have to be made. What does one do with a zero or middle point? What does zero mean? Is that the same as *I don't know*? As mentioned earlier, some studies separate out the *I don't know* from the actual scale. As further mentioned, it might be advisable to eliminate a participant's responses when there is a preponderance of *I don't know* or zero options.

## Magnitude Estimation

Earlier, I discussed the problems involved in interval rather than ordinal scales. Less common than standard grammaticality judgments is magnitude estimation (Bard, Robertson, & Sorace, 1996). Using this version of acceptability puts the rating scale in the hands of the rater and not the researcher. As Sorace (2010) points out, linguistic judgments "are gradient, i.e., they come in varying degrees of acceptability" (p. 58). With magnitude estimation, par-



ticipants are asked to give a numerical value to a sentence as a starting point, although the researcher could present a numerical starting point. Subsequent sentences are rated with relation to that number (e.g., is one sentence twice as acceptable as the previous one or half as acceptable as the previous one). As Bard et al. note, through this method,

[r]esearchers can observe meaningful differences that directly reflect differences in the participants' impressions of the property being investigated. This is so because magnitude estimation allows researchers to subtract the score on one sentence from that of another and be confident about the magnitude of difference (p. 41).

Sorace (2010) provides information on the type of grammatical features that magnitude estimation might be useful for. "Because Magnitude Estimation typically tends to yield more fine-grained distinctions than conventional rating or ordinal ranking scales, it is particularly suitable for the investigation of structures that exhibit gradience and, more generally, developmental optionality in ... L2 acquisition..." (p. 61). Hopp (2009) adds that fine-grained judgments are "essential in probing subtle acceptability differences according to discourse context" (p. 470).

Studies using magnitude estimation have been mostly limited to the acquisition of syntactic phenomena (e.g., Ayoun, 2005; Hopp, 2009; Kraš, 2011; Papp, 2000; Parodi & Tsimpli, 2005), but it has also been used within other frameworks (e.g., Gass, Mackey, Alvarez-Torres, & Fernandez-Garcia, 1999, who conducted work on task repetition).

As with standard judgment tasks, instructions take on a central role. Sorace (2010) presents instructions that may be appropriate for initial calibration. They include numerous examples including line length as a way of orienting participants to comparison judgments as well as specific examples of sentences. Preparing participants for the task at hand is essential to ensuring that conclusions drawn are appropriate.

## Preference Judgments

Of all judgment tasks, preferences are among the least used. Essentially, a preference task requires that a learner compares sentences, often asking which one is preferable. Trahey and White (1993) used a preference task in a study investigating the input necessary to modify second language grammars. Students were given two sentences (5) and (6):

5. Anna carefully drives her new car.
6. Anna drives carefully her new car.



## Elicited Imitation

Elicited imitation is a technique which, in its most typical administration, takes place in a lab setting and requires participants to listen to and repeat a sentence. The underlying assumption is that when someone hears a sentence and is asked to repeat it, he or she decodes the sentence to form a meaning representation following which is a reconstruction of the stimulus sentence within the boundaries of one's interlanguage grammar (see also Bley-Vroman & Chaudron, 1994). If the sentence corresponds to an individual's grammar, it will be relatively easy to imitate; if it is not, learners are likely to modify the sentence to reflect their current grammar. I deal below with some of the challenges that are inherent in using this technique because the entire process may be confounded by other factors.

An early discussion of elicited imitation comes from Slobin and Welsh (1973). As they state, "sentence recognition and imitation are filtered through the individual's productive linguistic systems" (p. 496). If a sentence cannot be maintained in short-term memory (e.g., it exceeds an individual's capacity), a participant cannot match the stimulus sentence. Instead, she/he decodes the sentence and re-encodes it in a way that conforms with his/her interlanguage grammar. More proficient speakers (and native speakers) are better able to match the input with their grammars with the resulting output (imitation) corresponding to the form of the input stimulus. As Spada, Shiu, and Tomita (2015) point out, "learners without any prior experience with the stimulus they are being asked to recall cannot do so because they have not built up long-term memory representations for the information encoded in the stimulus" (p. 726).

### History and Purposes of Use

An original use for elicited imitation was to determine the limits of short-term memory/working memory, with a focus on the number of units that can be held in memory (see reviews in West, 2012, and Jessop, Suzuki, & Tomita, 2007). In the past half century, its use has been prevalent in studies of child language learning (see Fraser, Bellugi, & Brown, 1963, for an early use in which the major goal was a comparison of production and comprehension). The elicitation technique extended to second language use initially in the 1970s (e.g., Hamayan, Saegert, & Larudee, 1977; Naiman, 1974) and has continued in the decades since (e.g., Munnich, Flynn, & Martohardjono, 1994) to this day. The primary uses of elicited imitation in second language research are as a way to determine (1) the nature of L2 grammars, (2) a learner's proficiency, and (3) L2 implicit knowledge.

Since early L2 adoption of elicited imitation, it has been used determine the nature of a person's grammar. For example, Schimke (2011) uses this tool

to determine the extent to which native-like syntax is acquirable. Is it the case that L2 underlying representations are fundamentally different from those of native speakers or is it the case that underlying representations are not different and the problems that frequently occur are superficial and relate to morphological surface realizations? Using elicited imitation as an elicitation measure (along with production data), Schimke sought to determine support for one or the other of these perspectives. The data were not able to support either position unequivocally, but did support a view that allowed for developmental phases with grammars that overlapped stages (see Vainikka & Young-Scholten, 1994, 1996a, 1996b).

Another use for elicited imitation is in a study by Trofimovich and Baker (2006) in which the researchers investigated the impact of time in a second language environment on the production of suprasegmentals. The task was a slight modification of a standard elicited imitation task in that the participants heard a question followed by a response. This was followed by the original question to which the participant was expected to repeat the response. The results showed that the learning of L2 suprasegmentals was similar to the learning trajectory of L1 suprasegmentals.

Elicited imitation has been used in recent years as a way of measuring proficiency. Such use is predicated on the assumption that processing efficiency is a reflection of proficiency mediated by working memory capacity (Gaillard & Tremblay, 2016; Van Moere, 2012). In studies by Wu and Ortega (2013) and Gaillard and Tremblay (2016), elicited imitation was shown to be a good measure of general proficiency. This was supported by a meta-analysis by Yan, Maeda, Lv, and Ginther (2016) who considered 21 studies over a period of nearly 45 years. Their results suggest that elicited imitation does discriminate across proficiency levels.

## Validation

As discussed earlier, an issue in L2 research concerns the representation of implicit and explicit knowledge. Elicited imitation has been used as a tool to get at this distinction. In recent years, there have been studies in which the goal was to validate this tool as a measure of implicit knowledge. Erlam (2006) used an elicited imitation task with 95 L2 learners of English (mostly Chinese native speakers). Seventeen structures were targeted with a meaning component prior to imitation (see also Hsieh & Lee, 2014, who used a picture identification task to determine comprehension). Participants were asked to state whether a statement was true/not true/not sure. Her results suggest that in order to claim that an elicited imitation task is reconstructive (i.e., does not involve rote repetition), it must require “participants to first process state-

ments for meaning and then to repeat those statements that are grammatically correct and correct those statements that are ungrammatical” (p. 488).

## Considerations

As with most instruments that purport to tap internal linguistic knowledge or processing, elicited imitation is not without controversy. In particular, one doesn't always know whether failure to accurately repeat a stimulus is due to lack of comprehension or lack of grammatical knowledge (see Erlam, 2006; Hsieh & Lee, 2014; Vinther, 2002). To compensate for this and other issues, I present considerations that should be taken into account before undertaking a study using elicited imitation.

### 1. Length and structure

- a. Should there be a single target structure (Spada et al., 2015) or multiple structures (Erlam, 2006)?
- b. To avoid participants' reliance on rote memory, length must be carefully controlled. There is no single best approach, but pilot testing to ensure appropriate length is recommended.

West (2012) suggests that what is important is the number of morphemes and less so the number of syllables. Keeping syllables constant and varying morphemes provided greater likelihood of discrimination between proficiency levels (e.g., *limpia la ventana antigua* [“he/she cleans the old window”] versus *limpias las ventanas antiguas* [“you {fam} clean the old windows”]; both sentences have 9 syllables, whereas they have 8 and 12 morphemes, respectively).

Whatever is decided upon, the decision should keep in mind that one always wants to go beyond short-term memory if it is expected that reconstruction of a sentence is taking place.

2. Timing. Should responses be immediate or should there be a delay? A delay allows rehearsal by participant. Sometimes participants are asked to count to some number before responding. If there is some sort of comprehension measure before repetition, the issue of rehearsal is most likely moot. The question of time pressure is another consideration that may alter the outcome (e.g., Erlam, 2006). Spada et al. controlled for time pressure by controlling time after the power point presentation of the stimulus (truth-value question remained on screen for six sentences). This was followed by a beep after which participants had eight seconds to repeat the stimulus sentence.

3. Should there be grammatical and ungrammatical tokens or only grammatical tokens (see Erlam, 2006, for a justification for using ungrammatical stimuli)?
4. Production versus comprehension. The issue of what elicited imitations tap is controversial (Bley-Vroman & Chaudron, 1994). Comprehension measures help to alleviate this concern.
5. Instructions vary and are often not made explicit in research articles. They can be as simple as “Repeat as best you can,” “Repeat in correct English,” or “Repeat exactly what you heard and as much as you heard.” For reporting purposes, it is important to state what the instructions were.
6. Scoring. Studies use different scoring procedures. For example, Erlam (2006) had a 3-point scale (correct, used the target structure, but not correctly, didn't use the target structure). Ortega, Iwashita, Norris, and Rabie (2002), cited in Wu and Ortega (2013), used a 5-point scale: (1) correct, (2) accurate content repetition with some (ungrammatical or grammatical) changes of form, (3) changes in content or in form that affect meaning, (4) repetition of half of the stimulus or less, and (5) silence or one word repeated or unintelligible repetition. Gaillard and Tremblay (2016) used 7-point scales on five factors: (1) meaning, (2) syntax, (3) morphology, (4) vocabulary, and (5) pronunciation. Spada et al. (2015) used a 2-point scale (correct or incorrect).

## Conclusion

This chapter has dealt with two common elicitation measures purporting to provide information about L2 learners' linguistic knowledge and processing capabilities. They both have a long history in second language research; both are controversial in terms of what they measure; and both have been the subject of examination. Yet, both continue to be used widely. For this reason, it is important for researchers to have a full understanding of strengths and limitations. The intent of this chapter was to elucidate some of the issues surrounding both elicitation techniques.

## Resources for Further Reading

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publishers.

This book is a classic for anyone wanting to understand the underpinnings of judgment data. It is aimed at those conducting research with native speakers, but the methods described are applicable to a wide range of contexts. The book is practical in nature and deals with issues of design, data coding, and statistical analysis.

Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: The University of Chicago Press.

This book presents an overview of the uses of grammaticality judgments throughout the long history of their use in linguistics. Even though it was published 20 years ago, its comprehensive literature review provides insight into the evidential role that judgment data bring to bear on theory construction. The overview includes a discussion about ways in which different factors can influence judgments. For example, a chapter concerning subject-related factors that includes individual differences and experiential backgrounds (e.g., linguistic training and literacy) and a chapter on task-related factors (i.e., factors having to do with procedures and stimuli) are important as they help understand appropriate designs and contextualize interpretation of data.

Spinner, P. & Gass, S. (forthcoming). *Using judgments in second language acquisition research*. New York: Routledge.

This book discusses judgment data from the perspective of second language acquisition. It reviews the literature on judgment data with native speakers as a way of contextualizing their use in L2 research. The focus is the history of and controversies relevant to the use of judgments in linguistics research. With regard to the L2 literature, the book provides a background discussion on historical and current issues such as the knowledge being tapped. It points to the different interpretations of data, depending on whether it is from native speakers or L2 learners. It presents data on current practices and emphasizes best practices in judgment data. The central focus is on traditional judgment data, but other types of judgments (e.g., preference, truth-value, magnitude estimation) are also included.

## Notes

1. I use the term second to refer to language learning beyond the first. Thus, the term “second” or L2 is used as a heuristic for nonprimary language acquisition (i.e., third, fourth, etc.).

2. For an excellent online database of tasks, please see IRIS, a digital repository of data collection instruments for research into second language learning and teaching. This is a free resource of elicitation instruments that have appeared in “peer-reviewed journals, conference proceedings and books” (Marsden, Mackey, & Plonsky 2016, p. 6; [www.iris-database.org](http://www.iris-database.org)).
3. Gass (1979) compared production data with judgment data, coming to different conclusions when analyzing each alone. Through an analysis of both judgment data and written production data, she was able to determine instances of avoidance which reflected the particular construct under investigation (Accessibility Hierarchy).
4. Technically, they are best referred to as acceptability judgments. A grammar is an abstraction and one cannot access that abstraction directly. What one does is ask about the acceptability of a particular sentence and then make inferences about the abstract grammar that underlies that judgment. In this chapter, I refer to them as grammaticality judgments given that this is the most commonly used term in the literature.

## References

- Ayoun, D. (2005). Verb movement phenomena in Spanish: ‘Mixed languages’ and bilingualism. In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), *Proceedings of the 4th Bilingualism Symposium*. Somerville, MA: Cascadia Press.
- Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1), 32–68.
- Batterink, L., & Neville, H. (2013). Implicit and explicit second language training recruit common neural mechanisms for syntactic processing. *Journal of Cognitive Neuroscience*, 25(6), 936–951.
- Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitations as a measure of second-language competence. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 245–261). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bley-Vroman, R., Felix, S., & Ioup, G. (1988). The accessibility of Universal Grammar in adult language learning. *Second Language Research*, 4(1), 1–32.
- Bowles, M. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition*, 33(2), 247–271.
- Chaudron, C. (1983). Research in metalinguistic judgments: A review of theory, methods and results. *Language Learning*, 33(3), 343–377.
- Corder, S. P. (1973). The elicitation of interlanguage. In J. Svartvik (Ed.), *Errata: Papers in error analysis* (pp. 36–48). Lund, Sweden: CWK Gleerup.



- Cowan, R., & Hatasa, Y. (1994). Investigating the validity and reliability of native speaker and second-language learner judgments about sentences. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 287–302). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. London: Sage Publications.
- Ellis, R. (2005). At the interface: Dynamic interactions of explicit and implicit knowledge. *Studies in Second Language Acquisition*, 27(2), 305–352.
- Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27(3), 464–491.
- Falk, Y., & Bardel, C. (2011). Object pronouns in German L3 syntax: Evidence for the L2 status factor. *Second Language Research*, 27(1), 59–82.
- Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of Verbal Learning and Verbal Behavior*, 2(2), 121–135.
- Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, 66(2), 419–447.
- Gass, S. (1979). Language transfer and universal grammatical relations. *Language Learning*, 29(2), 327–344.
- Gass, S., & Alvarez Torres, M. (2005). Attention when? An investigation of the ordering effect of input and interaction. *Studies in Second Language Acquisition*, 27(1), 1–31.
- Gass, S., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gass, S., Mackey, A., Alvarez-Torres, M., & Fernandez-Garcia, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, 49(4), 549–581.
- Gass, S., & Polio, C. (2014). Methodological influences of “interlanguage” (1972): Data then and data now. In Z.-H. Han & E. Tarone (Eds.), *Interlanguage 40 years later*. John Benjamins: Amsterdam, Netherlands.
- Gass, S., Svetics, I., & Lemelin, S. (2003). Differential effects of attention. *Language Learning*, 53(3), 497–545.
- Glew, M. (1998). *The acquisition of reflexive pronouns among adult learners of English*. Unpublished doctoral dissertation, Michigan State University, East Lansing.
- Godfroid, A., Loewen, S., Jung, S., Park, J.-H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge. *Studies in Second Language Acquisition*, 37(2), 269–297.
- Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, 35(3), 423–449.

- Hamayan, E., Saegert, J., & Larudee, P. (1977). Elicited imitation in second language learners. *Language and Speech*, 20(1), 86–97.
- Hopp, H. (2009). The syntax-discourse interface in near-native L2 acquisition: Off-line and on-line performance. *Bilingualism: Language and Cognition*, 12(4), 463–483.
- Hsieh, A., & Lee, M.-K. (2014). The evolution of elicited imitation: Syntactic priming comprehension and production task. *Applied Linguistics*, 35(5), 595–600.
- Hyltenstam, K. (1977). Implicational patterns in interlanguage syntax variation. *Language Learning*, 27(2), 383–411.
- Ionin, T., & Zyzik, E. (2014). Judgment and interpretation tasks in second language research. *Annual Review of Applied Linguistics*, 34, 37–64.
- Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, 64(1), 215–238.
- Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of ESL. *Cognitive Psychology*, 21(1), 60–99.
- Keating, G., VanPatten, B., & Jegerski, J. (2011). Who was walking on the beach? Anaphora resolution in Spanish heritage speakers and adult second language learners. *Studies in Second Language Acquisition*, 33(2), 193–221.
- Kraš, T. (2011). Acquiring the syntactic constraints on auxiliary change under restructuring in L2 Italian: Implications for the Interface Hypothesis. *Linguistic Approaches to Bilingualism*, 1(4), 413–438.
- Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, 36(5), 1247–1282.
- Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning testing and teaching* (pp. 94–112). Bristol: Multilingual Matters.
- Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*, 22(2), 429–445.
- Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York: Routledge.
- Marsden, E., Plonsky, L., Gass, S., Spinner, P., & Crowther, D. (2016). *Looking back and looking forward: A methodological synthesis of judgment tasks in second language research*. Paper presented at EuroSLA, Jyväskylä, Finland.
- Montrul, S., Davidson, J., De La Fuente, I., & Foote, R. (2014). Early language experience facilitates the processing of gender agreement in Spanish heritage speakers. *Bilingualism: Language and Cognition*, 17(1), 118–138.

- Montrul, S., Dias, R., & Santos, H. (2011). Clitics and object expression in the L3 acquisition of Brazilian Portuguese: Structural similarity matters for transfer. *Second Language Research*, 27(1), 21–58.
- Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks: What they measure and how they relate to each other. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 227–245). Mahwah, NJ: Erlbaum.
- Murphy, V. (1997). The effect of modality on a grammaticality judgment task. *Second Language Research*, 13(1), 34–65.
- Myers, J. (2009). Syntactic judgment experiments. *Language and Linguistics Compass*, 3(1), 406–423.
- Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers in Bilingualism*, 2, 1–37.
- Ortega, L., Iwashita, N., Norris, J., & Rabie, S. (2002, October). *An investigation of elicited imitation tasks in crosslinguistic SLA research*. Paper presented at the Second Language Research Forum, Toronto.
- Papp, S. (2000). Stable and developmental optionality in native and non-native Hungarian grammars. *Second Language Research*, 16(2), 173–200.
- Parodi, T., & Tsimpili, I. (2005). Real and apparent optionality in second language grammars: Finiteness and pronouns in null operator structures. *Second Language Research*, 21(3), 250–286.
- Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Amsterdam: John Benjamins.
- Pienemann, M. (2005). Discussing PT. In M. Pienemann (Ed.), *Cross-linguistic aspects of processability theory* (pp. 61–83). Amsterdam: John Benjamins.
- Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, 63(3), 595–626.
- Schachter, J., Tyson, A., & Diffley, F. (1976). Learner intuitions of grammaticality. *Language Learning*, 26(1), 67–76.
- Schimke, S. (2011). Variable verb placement in second-language German and French: Evidence from production and elicited imitation of finite and nonfinite negated sentences. *Applied Psycholinguistics*, 32(4), 635–685.
- Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistics methodology*. Chicago: The University of Chicago Press.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, 10(3), 209–231.
- Slobin, D., & Welsh, C. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson & D. Slobin (Eds.), *Studies of child language development* (pp. 485–497). New York: Holt, Rinehart & Winston.
- Sorace, A. (2010). Using magnitude estimation in developmental linguistic research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 57–72). Amsterdam: John Benjamins.

- Spada, N., Shiu, J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning, 65*, 723–751.
- Spinner, P. (2013). Language production and perception: A Processability Theory study. *Language Learning, 63*(4), 704–739.
- Spinner, P., & Gass, S. (forthcoming). *Using judgment tasks in second language acquisition research*. New York: Routledge.
- Trahey, M., & White, L. (1993). Positive evidence and preemption in the second language classroom. *Studies in Second Language Acquisition, 15*(2), 181–204.
- Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1–30.
- Vafae, P., Suzuki, Y., & Kachisnke, I. (2016). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition, 39*(1), 59–95.
- Vainikka, A., & Young-Scholten, M. (1994). Direct access to X<sup>3</sup>-theory: Evidence from Korean and Turkish adults learning German. In T. Hoekstra & B. Schwarz (Eds.), *Language acquisition studies in generative grammar* (pp. 265–316). Amsterdam: John Benjamins.
- Vainikka, A., & Young-Scholten, M. (1996a). Gradual development of L2 phrase structure. *Second Language Research, 12*(1), 7–39.
- Vainikka, A., & Young-Scholten, M. (1996b). The earliest stages in adult L2 syntax: Additional evidence from Romance speakers. *Second Language Research, 12*(2), 140–176.
- Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325–344.
- Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics, 12*(1), 54–73.
- West, D. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition, 4*(3), 203–222.
- Wu, S.-L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals, 465*(4), 680–704.
- Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497–528.
- Yuan, B., & Dugarova, E. (2012). Wh-topicalization at the syntax-discourse interface in English speakers' L2 Chinese grammars. *Studies in Second Language Acquisition, 34*(4), 533–560.



# 16

## Introspective Verbal Reports: Think-Alouds and Stimulated Recall

Melissa A. Bowles

### Introduction

Traditionally, spoken and written learner language production were the bread and butter of data in the field of second language acquisition (SLA), with some researchers (Selinker, 1972) arguing that they should be the sole source of data for theory-building and investigation and others (Corder, 1973) disagreeing on the grounds that production provides a small piece of the puzzle. In the subsequent decades, as SLA has matured as a discipline, data elicitation techniques have become increasingly varied and sophisticated (see Mackey & Marsden, 2016, for an overview).

Introspective verbal reports from learners are one elicitation method commonly used to gather information about cognitive processes that cannot be gleaned from production data alone. Simply put, verbal reports are verbalizations learners make either while completing a task (concurrent reports, or think-alouds) or some time thereafter (retrospective reports). In applied linguistics research, one particular type of retrospective report known as stimulated recall has been used extensively, in at least 125 language-related studies published between 2000 and 2015 (Gass & Mackey, 2016). In stimulated recall, participants are provided with some sort of stimulus (most often an audio or video recording of themselves completing the task) to help them recall what they were thinking at the time of task completion.

---

M. A. Bowles (✉)

University of Illinois, Urbana, IL, USA

e-mail: [bowlesm@illinois.edu](mailto:bowlesm@illinois.edu)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_16](https://doi.org/10.1057/978-1-137-59900-1_16)

Although both think-alouds and stimulated recall aim to gain insight into learners' cognitive processes, they can be distinguished in several important ways that have implications for their use in research. Apart from the time of reporting, think-alouds and stimulated recall can differ in their modality. Both think-alouds and stimulated recall are most commonly oral (i.e., participants speak their thoughts out loud), but it is possible for responses in a stimulated recall to be written rather than spoken, as in Krauskopf (1963). Additionally, the two types of reporting differ in the amount and type of support that is provided. Since think-alouds occur simultaneous to task completion, the support is the task itself; however, since stimulated recalls occur after task completion, the stimulus that is provided by way of support or prompting from the researcher is an artifact of the task such as a recording.

Despite their widespread use by researchers of different theoretical orientations, including conversational analytic (CA) (Walters, 2007), sociocultural (Poehner, 2007; Smagorinsky, 2001; Swain, 2006), and cognitivist perspectives (Fujii & Mackey, 2009; Hama & Leow, 2010; Qi & Lapkin, 2001; Sachs & Polio, 2007), there has been controversy surrounding the validity of verbal reports. This chapter provides an overview of the history and use of verbal reports in language research, synthesizing research addressing their validity, outlining their advantages and disadvantages, and concluding that they can be valid if implemented appropriately. For further detail and step-by-step data collection and analysis tips that are beyond the scope of this chapter, readers are referred to Bowles (2010) for think-alouds and to Gass and Mackey (2016) for stimulated recall.

## Historical Development

Starting in the early twentieth century, verbal reports were used in cognitive psychology, most extensively to investigate problem-solving strategies on non-verbal tasks and puzzles. By the 1980s, their use was so widespread that “both concurrent and retrospective verbal reports [had become] generally recognized as major sources of data on subjects' cognitive processes in specific tasks” (Ericsson & Simon, 1993, p. xi). That is, their use was no longer confined to problem-solving but had expanded to a host of task types in a wide array of fields, ranging from accounting, care planning, counseling, drug and alcohol addiction treatment, ergonomics, marketing, nursing, and software engineering to medicine, where verbal reports are routinely used in the treatment of autism and developmental disorders, speech pathology, neurology, and cardiology. In each of these fields, verbal reports have become pivotal in providing insight into the cognitive processes of clients and patients.

The specific stimulated recall technique was first used by Bloom (1953), an educational psychologist investigating the thought processes and learning outcomes of college students in lecture classes. His rationale for the use of a stimulus was that “the subject may be enabled to relive an original situation with vividness and accuracy if he is presented with a large number of the cues or stimuli which occurred during the original situation” (p. 161). He therefore argued that the contents of stimulated recalls should be more reliable and accurate than standard retrospective reports, which lack such a stimulus and would be more affected by memory decay. The technique was further refined by Siegel, Siegel, Capretta, Jones, and Berkowitz (1963), who first implemented video-recorded stimuli, and stimulated recall has since become widely used not only in education, where it has been used to evaluate teacher effectiveness (Peterson & Clark, 1978) and teachers’ decision-making processes (Calderhead, 1981), but also in technology education (Fox-Turnbull, 2009) and physical education (Tjeerdsma, 1997) as well as in other disciplines such as conflict resolution (Kressel, Henderson, Reich, & Cohen, 2012), management (Burgoyne & Hodgson, 1983), athletic coaching (Gilbert, Trudel, & Haughian, 1999), counseling (Kagan, Krathwohl, & Miller, 1963), and medicine, where it has been a technique to investigate doctors’ clinical reasoning (Elstein, Shulman, & Sprafka, 1978) and outcomes of nursing training (Daly, 2001).

## Origin and Use in Applied Linguistics

Although they originated in other disciplines, verbal reports have been used increasingly in language research since the 1980s. First language reading and writing research has relied heavily on think-alouds, focusing on strategies used by different groups of students, such as more and less successful readers and writers. In some cases, think-alouds have been used as one component of instructional programs designed to make students aware of successful strategies to improve their reading and writing ability.

In L2 research, both think-alouds and stimulated recalls have also been used to study reading and writing. Additionally, they have been used to investigate a wide array of other phenomena, including L1 and L2 strategy use, lexical organization, the use of translation in L2 learning, interlanguage pragmatics, and vocabulary acquisition through incidental reading (see Bowles, 2010, for references). They have also been used extensively as part of the validation process for large-scale assessments of L2 learners’ reading, speaking, and writing abilities (e.g., Wei & Llosa, 2015, described in Sample Study 16.1). Since think-alouds are collected while a task is being completed, they



are better suited for certain research topics, such as the effects of different levels of awareness on L2 learning and the relationship between explicit and implicit L2 knowledge. On the other hand, since stimulated recall occurs after task completion, it is well suited for oral interaction studies or other sorts of research involving speaking tasks.

### Sample Study 16.1

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, 12(3), 283–304.

#### Research Background

Many high-stakes language tests, such as the TOEFL, have components that are scored by human raters. A central concern in language testing has to do with the consistency and fairness of such ratings, and this study sought to determine whether rater background had an effect on the ratings they assigned to test-takers.

#### Research Method

- *Type of research*: Large-scale language assessment
- *Setting and participants*: Six experienced, trained TOEFL raters (three American and three Indian) assigned to rate 60 TOEFL iBT speech samples of ten Indian test-takers.
- *Instruments/techniques*: Raters instructed to think aloud (non-metalinguistically) while they rated, following the same scoring rubric normally used. Think alouds complemented score data.
- *Data analysis*: Multifaceted Rasch techniques used for score data, with qualitative analyses of the think-aloud data.

#### Research Question

- Do Indian and American TOEFL raters differ in their scoring of Indian test-takers' TOEFL iBT samples?

#### Key Results

Although the results of the quantitative analysis showed that American and Indian raters did not differ in the consistency or severity of their scores, the think-aloud data provided more nuanced information about the rating process, showing that there were in fact some differences based on rater background. Specifically, Indian raters more easily identified and understood features of Indian English in the test-takers' responses, but Indian and American raters did not differ in their use of scoring criteria or in their attitudes toward Indian English.

#### Comments

This study is a good example of how think-alouds can be used to triangulate findings from other data sources, providing additional context for the results from quantitative analyses. Increasingly, think-alouds are being used in large-scale language testing as a component of the validation process.



## Core Issues

Despite the widespread use of verbal reports, there has been controversy surrounding their validity, which hinges on (1) whether verbalizing alters the very thought processes under investigation (reactivity) and (2) whether verbalizations are an accurate reflection of thoughts (veridicality). More than three decades ago cognitive psychologists first raised questions about whether verbal reports actually capture thought processes accurately, or whether they alter those processes (i.e., whether they are “reactive”) (Nisbett & Wilson, 1977; Payne, Braunstein, & Carroll, 1978). Similar concern has since been voiced in SLA. For instance, Jourdenais (2001) specifically cautioned that “the think aloud data collection method itself acts as an additional task which must be considered carefully when examining learner performance” (p. 373). Concern has also been expressed about the accuracy of participants’ memory when they provide retrospective reports (Ericsson & Simon, 1993) or stimulated recalls (Gass & Mackey, 2016; Smagorinsky, 1994). Because of differences between think-alouds and stimulated recall, the validity concerns are manifested in slightly different ways depending on the report methodology in question and each is discussed in turn below.

### Reactivity with Think-Alouds

In a now-classic model of verbalization, Ericsson and Simon (1993) distinguished verbal reports according to time of reporting (concurrent vs. retrospective) and level of detail. The model predicts that concurrent reports will be more complete and accurate (veridical) than retrospective reports, since participants who think aloud during a task are not subject to memory decay, unlike those who are asked to report their thoughts some time after completing a task. That is, for Ericsson and Simon, the chief validity concern for think-alouds is reactivity rather than veridicality, and indeed the vast majority of research across all fields that has examined the validity of think-alouds has been concerned with reactivity.

The model furthermore predicts that verbalizations that are generated as a normal part of the solution process (what Ericsson and Simon refer to as Type 1 verbalizations, or non-metalinguistic verbalizations in the SLA literature) will be generally nonreactive; that is, they should reflect the nature of cognitive processes fairly accurately, although slowing processing slightly.

Conversely, the model predicts that verbalizations that include additional reasoning or justifications that would not normally figure into the normal solution process (what Ericsson and Simon refer to as Types 2 and 3 verbalizations, or metalinguistic verbalizations) are more likely to be reactive (to affect task performance).

Systematic investigation into the validity of verbal reports has been undertaken since the 1950s in psychology, most commonly by comparing the task performance of a group that thinks aloud to a group that completes the same task silently. Statistically significant differences in task performance, when all other variables are held constant, are taken as an indication that the act of verbalizing had an effect on participants' thought processes (i.e., it was reactive). In a meta-analysis of 94 studies with a total of approximately 3500 participants, Fox, Ericsson, and Best (2011) confirmed the predictions of Ericsson and Simon's (1993) model, finding that although think-aloud participants took longer to complete tasks compared to their silent counterparts, those who provided non-metalinguistic verbalizations otherwise performed statistically similarly to silent controls and "the think-aloud effect size [was] indistinguishable from zero ( $r = -0.03$ )" (p. 1). They found an effect size significantly different from zero ( $r = 0.23$ ) when participants were required to describe or explain their thoughts (provide metalinguistic verbalizations).

Although the results of this meta-analysis bode well for the validity of think-alouds in language research, Fox et al. (2011) did not specifically set out to investigate the reactivity of think-alouds used in conjunction with *verbal* tasks. In fact, because they were drawing from the psychology literature, most of the studies included in their meta-analysis involved problem-solving tasks, making the relevance of the findings for linguistics research uncertain.

Bowles (2010) set out to isolate the effects of task type, investigating the issue of reactivity when think-alouds were used with verbal tasks. In her meta-analysis, which included 14 unique sample studies, results were not as decisive. However, she was able to conclude, similar to Fox et al. (2011), that thinking aloud increased time on task across the board, although effect sizes varied substantially, ranging from small ( $d = 0.16$ ) to very large ( $d = 1.16$ ), with the most pronounced increase in time for participants who were asked to think aloud while performing reading tasks. Regarding the impact of thinking aloud on task performance, "in 86 per cent of the effect size calculations, the 95 per cent confidence interval overlap[ped] zero, indicating that the  $d$  value [was] not significantly different from zero" (Bowles, 2010, p. 138). It was not possible to produce grand weighted mean effect sizes because the distributions violated the assumption of homogeneity, but this finding, taken

in conjunction with that of Fox et al. (2011), suggests that with both verbal and non-verbal tasks, thinking aloud does not generally impact task performance or, by extension, alter thought processes.

## Veridicality with Think-Alouds

Ericsson and Simon's (1993) model predicts that think-alouds will be a more accurate representation of thoughts than retrospective reports because there is no time lapse between task completion and verbalization and therefore no possibility for participants to forget their thoughts. Perhaps for this reason, the veridicality of think-alouds has been rather sparsely investigated, with the few psychology studies (Kuusela & Paul, 2000; Robinson, 2001) corroborating Ericsson and Simon, finding that think-alouds tend to be more accurate and to include more information than retrospective reports. Winikoff (1967) examined the veridicality of think-alouds by additionally tracking participants' eye movements during verbalization, finding that the two methodologies corroborated each other. In the SLA literature, two studies to date have examined the issue, finding somewhat mixed results. Similar to Winikoff, Godfroid and Spino (2015) triangulated think-aloud reports by having participants also use their index finger to track words as they read a text. They found that although there was overlap in what the two methods revealed, finger tracking was more sensitive at capturing processes going on at a low level of awareness than think-alouds were. Barkaoui (2011) found that more than a third of the essay raters in his study, who thought aloud as they rated, reported in a subsequent interview that they were unable to verbalize all of the thoughts that went through their minds as they evaluated the essays.

Rather than taking these studies' findings as a challenge to the validity of think-alouds on the grounds that they do not accurately represent thought processes, I would argue that the results in fact indicate that think-alouds are accurate, although they may well not be exhaustive, including every fleeting thought or sensation on every level of awareness that a participant experiences while thinking aloud. Along these lines, Leow, Grey, Marijuan, and Moorman's (2014) review of the ways in which think-alouds, eye tracking, and reaction time measurements can be most profitably used to investigate cognitive processes demonstrated that eye tracking and reaction times appear to be more sensitive at capturing cognition at lower levels of awareness, whereas think-alouds appear to be better suited for capturing cognition at higher levels of awareness.

## Reactivity and Veridicality with Stimulated Recall

Turning now to stimulated recall, “systematic explorations [of validity issues] are seldom found” (Gass & Mackey, 2016, p. 30). Since it is a type of retrospective report, the main threat to the validity of stimulated recall is its potential to be non-veridical, or for participants not to accurately recall what they were thinking at the time of task completion. However, the effects of such memory decay are mitigated by several aspects of the methodology which distinguish it from other retrospective reports, if it is carefully designed and carried out. First and foremost, memory decay is anticipated to be less of a problem with stimulated recall reports than with other retrospective reports because the former have a stimulus whereas the latter do not. Furthermore, if the time lapse between task completion and stimulated recall is short, ideally less than 48 hours, as Gass and Mackey (2016) advise it should be, and if the stimulus provided at the time of reporting is strong, such as a video recording of the participant completing the task, the impact of memory should be minimal. Finally, the instructions provided to the participant should be carefully written and uniformly presented to ensure that participants state what they were thinking *at the time they completed the task* rather than what they think *at the time of reporting*. Careful adherence to these guidelines and to the advice on designing a study using stimulated recall in Chap. 3 of Gass and Mackey (2016) should serve to minimize the risk of non-veridicality with stimulated recall.

Reactivity is not typically a major validity concern for retrospective reports, since verbalization happens after the task has already been completed. However, in the case of stimulated recall, when posttests take place after a recall session, verbalization does have the potential to be reactive, especially since it provides participants with additional exposure to language input that they would not otherwise have had. Two SLA studies have specifically investigated this question, drawing mixed conclusions. In Egi (2007), L2 learners of Japanese received recasts while engaging in task-based interaction. An immediate recall group provided their impressions of the recasts during the task, whereas a stimulated recall group completed the task, followed immediately by a stimulated recall session in which they verbalized their thoughts about the recasts. An immediate posttest showed no significant differences in performance based on report type, but stimulated recall participants significantly outperformed immediate recall participants on the delayed posttest. Thus, Egi’s findings showed that stimulated recall was reactive, having a facilitative impact on L2 learners’ performance. In a follow-up study, Egi (2008) attempted to identify the source of the reactivity by randomly assigning L2

learners to participate in task-based interaction in one of four groups—a stimulated recall group, a stimulus only group (which watched the video silently), an experimental control (which completed the communicative activity but did not verbalize or watch the video), and a test control (which did not complete the communicative activity, verbalize, or watch the video but took all pre- and posttests.) Results showed no significant differences between the stimulated recall, stimulus, and experimental control groups in terms of post-test performance, suggesting non-reactivity for stimulated recall in this case.

As this brief review shows, there is no one-size-fits-all answer to the question of the validity of verbal reports. Validity concerns must always be paramount in researchers' minds, both as they design and plan a study using verbal reports and as they analyze and interpret the verbal report data. Readers who wish to use verbal reports in their research are referred to detailed *how-to* guidance in Bowles (2010) for think-alouds and in Gass and Mackey (2016) for stimulated recall. If appropriate safeguards are taken, verbal reports can provide a rich dataset that would be out of reach without such introspective methods.

## Overview of Research Types

Verbal reports are inherently flexible in the sense that they can be used to provide insight into learners' cognitive processes and strategies on virtually any topic. In part because of this flexibility, they are routinely used not only by researchers who take a cognitive approach to SLA, with which they are perhaps most associated, but also by researchers from sociocultural and conversation analytic (CA) perspectives as well. Depending on the theoretical framework adopted, the starting assumptions about verbal reports and their analysis and interpretation vary, however. Whereas cognitivist SLA researchers take the stance that verbal reports are a window into learners' minds and a reflection of their individual processing, researchers in sociocultural and CA camps tend to view verbal reports less as a reflection of individual thought processes and more as a means of mediating the internalization of new knowledge (Vygotsky, 1987) or, in the case of collaborative dialogue or languaging, a way of socially co-constructing knowledge (Swain, 2006).

Perhaps unsurprisingly given the diverse perspectives of researchers using verbal reports as a data elicitation tool, it is common to find both purely quantitative and purely qualitative studies using think-alouds and/or stimulated recall, as well as studies that combine both quantitative and qualitative methods. In fact, Gass and Mackey (2016) espouse such an approach, writing

that “combining analytical techniques is desirable wherever possible. The shortcomings of one type of analysis may be addressed through the strengths of another... For example, ... qualitative data can be used to shed light on the findings of any quantitative analysis” (pp. 153–154).

Along the same lines, verbal reports can be used with a wide variety of different research designs, ranging from experimental designs in laboratory settings and quasi-experimental designs involving research on intact classes of learners to teacher-initiated action research in classrooms and single or multiple case study designs. Depending on the specific goals and research questions being investigated, sometimes verbal reports are the sole source of data, as in Mackey, Gass, and McDonough (2000) (see Sample Study 16.2), and sometimes the verbal report data serves to complement other data sources, as in Godfroid and Schmidtke (2013), which used a combination of think-alouds, eye tracking, and posttests to investigate incidental vocabulary learning. Most studies that include verbal reports as a data elicitation measure use either think-alouds or stimulated recall, not both, and given that it is easier to minimize validity threats if only one type of verbal report is used, this is generally advisable. However, some applied linguistics studies have incorporated both think-alouds and stimulated recall into their designs (e.g., Nassaji, 2003; Upton & Lee-Thompson, 2001).

### Sample Study 16.2

Mackey, A., Gass, S.M., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, 22, 471–497.

#### Research Background

This study was written just a few years after the publication of the revised version of the Interaction Hypothesis (Long, 1996), when many of its tenets were being empirically tested for the first time. If, as Long claimed, interaction (and resulting feedback) could focus learners' attention on aspects of their production that were not target-like, a necessary precondition is for L2 learners to accurately perceive such feedback in the course of interacting with a highly proficient interlocutor. The study set out to investigate whether this was happening.

#### Research Method

- *Type of research*: Conversational interaction research
- *Setting and participants*: Ten ESL learners and seven Italian as a foreign language (IFL) learners who completed a spot-the-differences task with an English-speaking or Italian-speaking interlocutor.
- *Instruments/techniques*: Stimulated recall was used to gain insight into learners' perceptions of feedback. Interactions were videotaped and interlocutors “were instructed to provide interactional feedback wherever it seemed appropriate and in whatever form seemed appropriate during the interaction” (p. 479). Immediately after the task, a second researcher played the

video of the interaction, pausing it after all of the feedback episodes and allowing participants also to stop it at any time to comment. The goal was to get them to verbalize what they were thinking at the time of the interaction. Both interaction and recall sessions were subsequently transcribed, coded, and analyzed.

- *Data analysis:* Descriptive statistics (frequencies and percentages).

#### Research Questions

- Do L2 learners accurately perceive feedback that occurs during interaction? (i.e., Do they perceive the corrective intent of the feedback and correctly identify the target of the feedback?)

#### Key Results

The study's main finding was that different kinds of feedback were not equally likely to be perceived accurately. Feedback on lexis and phonology (generally in the form of negotiation) tended to be perceived accurately, since learners' stimulated recall comments indicated that they understood the corrective intent and target of a majority of such feedback episodes. However, feedback on morpho-syntax (generally in the form of recasts) was rarely perceived accurately (as being about morphosyntax). Rather, learners' stimulated recall comments indicated that they often thought this feedback was about the content of their statements or semantics.

#### Comments

This was a seminal study, important both because of its revealing findings about how learners (mis)perceive feedback and because of its methodological innovation in the field, as it was the first SLA study to use stimulated recalls. Since then, the method has been used in over 125 published SLA studies (Gass & Mackey, 2016).

## Challenges and Controversial Issues

As the previous sections have shown, verbal reports are a versatile data elicitation tool that can be used either by themselves or in combination with other data sources to answer a seemingly infinite number of research questions in quantitative, qualitative, or mixed-method designs. Along with these benefits come challenges and controversial issues as well. These are the focus of this section.

### Challenges

Perhaps the most significant challenge to using verbal reports to elicit introspective data is that it is a labor-intensive methodology from start to finish, from the time the study is being designed to the time the verbal report data are analyzed, and at all points in between. Although complete sample protocols are presented in Bowles (2010) for think-alouds and in Gass and Mackey



(2016) for stimulated recall, this section will briefly outline key aspects of the data collection and analysis process that require careful planning and attention to ensure that potential threats to the validity of verbal report data are minimized or, ideally, eliminated.

Once the researcher has determined which verbal report method to use given the research questions and design of the study, it is imperative that instructions for participants be carefully written and piloted to ensure that they are clear and elicit the desired type of information. For instance, if a researcher wants to elicit non-metalinguistic think-alouds, the instructions need to make it clear to participants that they should verbalize whatever thoughts naturally go through their mind and not provide additional explanations or justifications for their thoughts or actions while they complete the task. Instructions should also specify what language participants are expected to speak (L1, L2, or a combination) when they verbalize. For stimulated recall only, the researcher must also identify an appropriately strong stimulus to use that will remind participants of the event without being distracting or excessive. The researcher must also plan the logistics of recording the stimulus. Video and audio recordings of task performance are by far the most common stimuli, but depending on the task type, there may be other suitable possibilities (e.g., transcripts of computer-mediated chat (CMC) sessions in Smith, 2012). Piloting instruments is pivotal, as it enables the researcher to estimate the time needed for the session and to make any adjustments prior to beginning full data collection so that all of the effort expended in designing the study and collecting the data results in usable data (e.g., recordings do not have to be discarded because of background noise or poor sound quality).

Researchers should also receive instructions and training prior to collecting verbal report data. Ideally, Gass and Mackey (2016) recommend that the researcher oversee a verbal report session in addition to receiving specific instructions. The training protocol should focus not only on what may seem like basic logistical issues (e.g., when to turn recording devices on and off) but also how to avoid common pitfalls. For think-alouds, researchers must be trained on how to get participants to speak if they fall silent, and the protocol should include scripts to exemplify how this should be done. For stimulated recalls, protocols must include detail about how to get participants to talk about what they were thinking at the time of the task rather than what they are thinking at the time of report, how to select specific excerpts of the stimulus to inquire about during the recall session, and how to inquire about participants' thoughts without asking leading questions. Again, specific scripts should be provided so that researchers have sample language to use when collecting the data.



Once instructions have been written and piloted and both researchers and participants know what they should do, data collection can begin. Data collection with verbal reports is time-consuming in and of itself, since it must typically be done one-on-one. When the verbal reports have been collected in accordance with the established procedures, researchers then proceed to the data analysis phase, which is again labor-intensive. Participants' entire verbal report, or in some cases, selected portions of the report that have been identified in advance, must be transcribed. Although the level of detail of the transcription varies according to the research question, even a simple orthographic transcript can take many times longer to produce than the recording itself is long. The transcriptions should then be checked for accuracy and all researchers involved in coding should be trained to use the established coding scheme. In many cases, the coding scheme may need to be modified as a result of rater socialization, particularly for high-inference coding categories. Then all of the data are coded (typically by one rater) and a randomly selected subset is chosen to be coded by a second rater and used to calculate inter-rater reliability, which should be included in any published reports. With coding complete, quantitative and/or qualitative data analysis can take place and inferences can be drawn from the verbal reports.

## Controversial Issues

The main source of controversy surrounding the use of verbal reports has to do with the reactivity and veridicality concerns addressed previously. These validity considerations should always be at the forefront of researchers' minds as they plan and carry out any study involving verbal reports. Apart from these concerns, two additional factors have been brought to light by recent research and merit discussion.

First, care must be taken to determine the language that participants should use to verbalize their thoughts. In psychology studies, language of reporting is rarely an issue and is in fact seldom mentioned at all because participants in most studies are monolingual. In language research, of course the situation is much more complicated. Verbal reports are more likely to be incomplete if participants cannot express the full range of their thoughts because of language proficiency limitations, so veridicality is a concern particularly when verbal reports are not completed in the L1 or in a language in which the participant is highly proficient. Regardless of whether the researcher chooses to have participants verbalize in L1, L2, or a combination of the two, this is a choice that should not be taken lightly, but rather carefully decided,

standardized across participants, and reported in the final write-up of the study (Bowles, 2010).

Second, even when participants receive carefully piloted instructions and have the opportunity to practice or warm up before beginning the experimental verbalization, some participants verbalize more than others and some are anecdotally more adept at doing so whereas others seem to struggle. In stimulated recall, the researcher periodically refers to particular portions of the stimulus and asks the participant to state what s/he was thinking at that time. This pushes the participant to respond, and researchers must be willing to accept responses that indicate a lack of recall (e.g., “I don’t remember thinking anything in particular”) and know how to respond so that they do not push the participant to fabricate memories. In a typical think-aloud design, the researcher periodically reminds the participant to continue thinking aloud if s/he is silent for more than a few seconds at a time. Normally this jars participants into verbalizing if they have unintentionally fallen silent, but again, participants may in some instances report not being able to verbalize any thoughts at a given time. If this occurs for only a small portion of the verbalization period, such brief lapses are unlikely to affect the quality of the data. However, if a participant is silent for a large portion of the data collection, the representativeness of the verbal report could be called into question and would likely need to be excluded from analysis.

This issue of propensity to verbalize can be subsumed under the umbrella category of individual differences that may have an impact on verbal reporting. Individual differences have been suggested to have more of an effect on think-alouds, since they require participants to perform the dual task of verbalizing while completing another cognitive task, whereas stimulated recall has no such dual task, since it provides a stimulus and asks participants to recall what they were thinking at various moments. In the vast majority of studies on verbal reports (in both psychology and SLA), no mention is made of individual difference variables that might have played a role in verbalization. An implicit assumption seems to be that, at least in designs that require more than a few participants, any individual differences that affect verbal reporting will “come out in the wash” and not have a significant impact on the study’s findings. However, the results of Goo (2010), which showed that thinking aloud was reactive in his study for L2 learners with high working memory but not for those with low working memory, call this practice into question and suggest that researchers should therefore consider collecting individual difference data, including working memory scores, so that they can determine what impacts individual differences have on reactivity.

## Limitations and Future Directions

As the preceding sections show, verbal report is a flexible data elicitation tool that can be used to gain insight into cognitive processes and help to answer a wide range of research questions in applied linguistics that require introspective data. As with any methodology, there are limitations with verbal reports that must be acknowledged. Cohen (1996, 1998) provides a thorough treatment of the expectations and limitations of verbal reports in the context of SLA research. The most central limitation has to do with the potential reactivity and non-veridicality of verbal report data, but if proper protocols are followed, as outlined in this chapter, these validity threats can be minimized. As an additional precaution, it is recommended that all studies with a think-aloud group have a silent control group that is otherwise matched for all relevant variables. This way, the silent and think-aloud groups' performances can be compared to determine whether there was reactivity for that group on the specific task(s) in question. Along the same lines, although it is not mentioned in Gass and Mackey (2016), in research where stimulated recall precedes a posttest, it is advisable to have a control group that participates in all phases except the stimulated recall session, again as a check on reactivity. The second main limitation is that some aspects of cognition are likely to be inaccessible by means of verbal reports; processing at lower levels of awareness may not be captured by verbal report but may require a more sensitive or fine-grained method (Leow et al., 2014). As such, at the outset of a project, researchers should carefully consider whether verbal reports will be suitable to answer their specific research questions or whether they would be better served by another data elicitation instrument.

Looking forward, verbal reports have a bright future in applied linguistics research. Increasingly, researchers have used them to triangulate data from psycholinguistic data collection measures, such as eye tracking or reaction times (Godfroid & Schmidtke, 2013). Introspective data can only help researchers to gain a more nuanced view of language acquisition and use as the field becomes increasingly sophisticated and integrates new, more sensitive measures into its methodological repertoire. As one small example, data from neurolinguistic measures, such as event-related potentials (ERPs) and functional magnetic resonance imaging (fMRI), have often been difficult to interpret in L2 research, and verbal reports provide one option that could serve to clarify and expand on their results.

Without a doubt, as verbal reports expand in their popularity as a data elicitation tool, further research is also needed regarding their reactivity and

validity in language research. Among the topics that have been under-researched to date are (1) how participants' individual difference variables (e.g., working memory and L2 proficiency) affect verbalization and (2) the extent to which thinking aloud is reactive on L2 tasks other than reading, which has been the focus of the lion's share of current reactivity studies.

## Resources for Further Reading

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

This book-length treatment offers an overview of how think-alouds have been used in language research and presents a quantitative meta-analysis of findings from 14 unique sample studies involving verbal tasks and think-alouds (published up through 2009). The book then offers in-depth guidance regarding practical issues of data collection and analysis and concludes by providing implications for the use of think-alouds in language research.

Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.) New York: Routledge.

This updated edition of the now-classic work on stimulated recalls in language research showcases advantages and disadvantages of the elicitation technique and provides examples and step-by-step guidance relevant for both novice and expert researchers alike.

## References

- Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, 28(1), 51–75.
- Bloom, B. (1953). Thought-processes in lectures and discussions. *Journal of General Education*, 7(3), 160–169.
- Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.
- Burgoyne, J., & Hodgson, V. (1983). Natural learning and managerial action: A phenomenological study in the field setting. *Journal of Management Studies*, 20(3), 387–399.
- Calderhead, J. (1981). A psychological approach to research on teachers' classroom decision making. *British Educational Research Journal*, 7(1), 51–57.

- Cohen, A. (1996). Verbal reports as a source of insights into second language learner strategies. *Applied Language Learning*, 7(1 & 2), 5–24.
- Cohen, A. (1998). *Strategies in learning and using a second language*. London: Longman.
- Corder, S. P. (1973). The elicitation of interlanguage. In J. Svartvik (Ed.), *Errata. Papers in error analysis* (pp. 36–48). Lund: CKW Geerup.
- Daly, W. (2001). The development of an alternative method in the assessment of critical thinking as an outcome of nursing education. *Journal of Advanced Nursing*, 36(1), 120–130.
- Egi, T. (2007). Recasts, learners' interpretations, and L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 249–267). Oxford: Oxford University Press.
- Egi, T. (2008). Investigating stimulated recall as a cognitive measure: Reactivity and verbal reports in SLA research methodology. *Language Awareness*, 17(3), 212–228.
- Elstein, A., Shulman, L., & Sprafka, S. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.
- Ericsson, K., & Simon, H. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: MIT Press.
- Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137(2), 316–344.
- Fox-Turnbull, W. (2009, August). Stimulated recall using autophotography: A method for investigating technology education. In A. Bekker, I. Mottier, & M. J. de Vries (Eds.), *Strengthening the position of technology education in the curriculum*. Proceedings PATT-22 Conference (pp. 204–217). Delft, The Netherlands: International Technology and Engineering Educators Association.
- Fujii, A., & Mackey, A. (2009). Interactional feedback in learner-learner interactions in a task-based EFL classroom. *International Journal of Applied Linguistics*, 47(3–4), 267–301.
- Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). New York: Routledge.
- Gilbert, W., Trudel, P., & Haughian, L. (1999). Interactive decision making factors considered by coaches of youth ice hockey during games. *Journal of Teaching in Physical Education*, 18(3), 290–311.
- Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports and vocabulary learning scores. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 183–205). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
- Godfroid, A., & Spino, L. (2015). Under the radar: Triangulating think-alouds and finger tracking to detect the unnoticed. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 73–90). New York: Routledge.

- Goo, J. (2010). Working memory and reactivity. *Language Learning*, 60(4), 712–752.
- Hama, M., & Leow, R. P. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition*, 32(3), 465–491.
- Jourdenais, R. (2001). Cognition, instruction, and protocol analysis. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 354–375). Cambridge: Cambridge University Press.
- Kagan, N., Krathwohl, D., & Miller, R. (1963). Stimulated recall in therapy using video tape: a case study. *Journal of Counseling Psychology*, 10, 237–243.
- Krauskopf, C. J. (1963). Use of written responses in the stimulated recall method. *Journal of Educational Psychology*, 54(3), 172–176.
- Kressel, K., Henderson, T., Reich, W., & Cohen, C. (2012). Multidimensional analysis of conflict mediator style. *Conflict Resolution Quarterly*, 30(2), 135–171.
- Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology*, 113(3), 387–404.
- Leow, R. P., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research*, 30(2), 111–127.
- Mackey, A., Gass, S. M., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, 22(4), 471–497.
- Mackey, A., & Marsden, E. (Eds.). (2016). *Instruments for research into second language*. New York: Taylor and Francis.
- Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly*, 37(4), 645–670.
- Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Payne, J. W., Braunstein, M. L., & Carroll, J. S. (1978). Exploring predecisional behavior: An alternative approach to decision research. *Organizational Behavior and Human Performance*, 22(1), 17–44.
- Peterson, P., & Clark, C. (1978). Teachers' reports of their cognitive processes during teaching. *American Educational Research Journal*, 15(4), 555–565.
- Poehner, M. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *Modern Language Journal*, 91(3), 323–340.
- Qi, D., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing*, 10(4), 277–303.
- Robinson, K. M. (2001). The validity of verbal reports in children's subtraction. *Journal of Educational Psychology*, 93(1), 211–222.
- Sachs, R., & Polio, C. (2007). Learners' uses of two types of written feedback on an L2 writing revision task. *Studies in Second Language Acquisition*, 29(1), 67–100.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, 10, 209–231.
- Siegel, L., Siegel, L., Capretta, P., Jones, R., & Berkowitz, H. (1963). Students' thoughts during class: A criterion for educational research. *Journal of Educational Psychology*, 54(1), 45–51.

- Smagorinsky, P. (Ed.). (1994). *Speaking about writing: Reflections on research methodology*. Thousand Oaks, CA: SAGE.
- Smagorinsky, P. (2001). Rethinking protocol analysis from a cultural perspective. *Annual Review of Applied Linguistics*, 21, 233–245.
- Smith, B. (2012). Eye tracking as a measure of noticing: A study of explicit recasts in SCMC. *Language Learning and Technology*, 16(3), 53–81.
- Swain, M. (2006). Linguaging, agency and collaboration in advanced language proficiency. In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 95–108). London: Continuum.
- Tjeerdsma, B. (1997). A comparison of teacher and student perspectives of tasks and feedback. *Journal of Teaching in Physical Education*, 16(4), 388–400.
- Upton, T. A., & Lee-Thompson, L. (2001). The role of the first language in second language reading. *Studies in Second Language Acquisition*, 23(4), 469–495.
- Vygotsky, L. (1987). *The collected works of L. S. Vygotsky* (Vol. I). New York: Plenum.
- Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing*, 24(2), 155–183.
- Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, 12(3), 283–304.
- Winikoff, A. (1967). *Eye movements as an aid to protocol analysis of problem solving behavior*. Pittsburgh, PA.





# 17

## Corpus Research Methods for Language Teaching and Learning

Magali Paquot

### Introduction

A corpus can be broadly defined as a collection of naturally occurring spoken or written data in electronic format, selected according to external criteria to represent a language, a language variety or a specific domain of language use (e.g., conversation or academic writing). Analyses of corpora using the methods and tools of corpus linguistics have contributed significantly to much more detailed and accurate descriptions of a wide variety of languages. They have also revolutionized language theory and description by placing special emphasis on (1) the frequency with which words and patterns occur in a language, (2) the ways in which words combine in collocations and other phraseological units, (3) the close connections between grammar and lexis, and (4) the ways in which situational factors (e.g., written versus oral communication, news writing versus fiction writing) act as explanatory variables for linguistic variability (Hunston, 2002).

The main objective of this chapter is to review how corpus-based descriptions of language use can have an effect on the teaching syllabus and the design of pedagogical materials by informing decisions about what to teach and when to teach it. Corpora can also be exploited directly by teachers and learners, but so far such uses have largely been confined to advanced proficiency levels and higher education. The reader is referred to Cobb and Boulton (2015) for an overview of direct uses of corpora in the L2 classroom.

---

M. Paquot (✉)

Universite catholique de Louvain, Louvain-la-Neuve, Belgium

e-mail: [magali.paquot@uclouvain.be](mailto:magali.paquot@uclouvain.be)



## Current Issues

### The Role of Frequency

Corpus-derived frequencies are often used to inform two main components of reference and instructional materials design: selection and sequencing. As put by Cobb and Boulton (2015, p. 479), “the basic idea is that frequency of form and meaning is the most reliable predictor of what can be most usefully taught at different points in the learning process.” One of the areas where frequency, among other corpus-based information, first brought about an unprecedented revolution is pedagogical lexicography. It would be inconceivable for lexicographers today to produce a learner’s dictionary in a well-resourced language without making use of corpus-derived frequencies at the very least as a guide to the selection of the lexical entries (i.e., which words to define and illustrate), the ordering of word meanings, and the identification of a defining vocabulary, that is, a restricted list of high-frequency words which can be used to provide simple definitions of words in the dictionary. When corpus-derived frequencies are used in textbook material development, they typically serve to sequence the introduction of new vocabulary words. For example, the authors of the *Vocabulary in Use* series published by Cambridge University Press used word lists derived from the *Cambridge International Corpus* to inform the selection of the most useful vocabulary that students need at each level.

With the increased availability of large reference and specialized corpora, work on frequency lists has been extremely productive over the last years,<sup>1</sup> and several proposals have sought to replace West’s (1953) *General Service List* (GSL), that is, an influential but dated list of 2000 lexical items deemed to be especially useful for the purposes of language pedagogy that was compiled before the advent of computerized corpora. One such proposal is the *New General Service List* (Brezina & Gablasova, 2015), which, unlike the GSL, makes use of lemmas instead of word families (see Paquot, 2010 or Gardner & Davies, 2014 for a synthesis of criticisms levelled at the use of word families to determine word frequencies). Brezina and Gablasova (2015) selected four general corpora to represent a variety of corpus sizes and approaches to sampling and representativeness and compared lists of the top 3000 lemmas for each corpus based on the average reduced frequency (ARF), which is a measure that serves to weight the absolute frequency of a word in terms of its distribution in the corpus.<sup>2</sup> The results showed that there exists a stable core vocabulary of 2122 lemmas (70.7%) among the four corpora.

Wordlists based on specialized corpora have also flourished over the last years, especially those targeting vocabulary needs in higher education. The *Academic Keyword List* (AKL, Paquot, 2010) contains 930 potential academic lemmas, that is, words that are reasonably frequent in a wide range of academic texts but relatively uncommon in other kinds of texts as identified by the three corpus-based techniques of keyword analysis, range, and dispersion. The AKL differs widely from Coxhead's (2000) *Academic Word List* as it is a list of lemmas, not word families, and was developed for productive purposes. As such, it includes high-frequency words (e.g., *aim, argue, because, compare, explain, namely, result*) which have been shown to play an essential structuring role in academic prose but remain a source of difficulty for EFL learners. Gardner and Davies (2014) adopted a similar approach to Paquot (2010) to produce the *New Academic Vocabulary List*, a list of 3000 lemmas derived from the academic subcorpus of the *Contemporary Corpus of American English* (COCA). Specialized wordlists for disciplines such as medicine, agriculture, engineering, and business have also been compiled on the basis of domain-specific corpora over the last ten years (see Lessard-Clouston, 2012/2013 for a selected overview).

The most recent wordlists can certainly be described as superior to previous efforts if only because they are based on ever larger corpora and compiled with the help of more sophisticated corpus-based techniques and measures. However, the large majority still focus on single words despite convergent findings in corpus linguistics, psycholinguistics, and cognitive linguistics that word combinations, be they framed in terms of phraseological units, formulaic sequences, or constructions, play crucial roles in language acquisition, processing, fluency, and idiomaticity (e.g., Ellis & Wulff, 2015; Sinclair, 1991). Three exceptions are

1. The *Academic Formulas List*: a list of 200-core, 200 written-specific, and 200 spoken-specific formulaic sequences which occur significantly more often in academic than in non-academic discourse (Simpson-Vlach & Ellis, 2010);
2. The *PHRASal Expressions List*: a list of 505 most frequent non-transparent multiword expressions such as *of course, at least* and *I mean* in the *British National Corpus* accompanied with frequency information for spoken general English, written general English, and written academic English (Martinez & Schmitt, 2012);
3. The *Academic Collocation List* (ACL): a list of 2468 most frequent and cross-disciplinary lexical collocations compiled from the written component of the *Pearson International Corpus of Academic English* (PICAE; Ackermann & Chen, 2013).

What these three lists have in common is that they all adopted a mixed-method approach that consists of corpus-derived frequency information and expert judgement to select and/or prioritize the most pedagogically useful word combinations.

It is still early to evaluate the impact of the newest corpus-derived lists of words and word combinations in the field of language learning and teaching. However, it is likely that they will be of more use in materials design<sup>3</sup> than in the classroom proper because, with an almost exclusive focus on frequency, they still require extensive pedagogic mediation to meet teachers' and learners' needs. One recent corpus-derived list that promises to be more useful for direct use in the classroom is the *Phrasal Verb Pedagogical List*, a selected list of the 150 most frequent phrasal verbs and their key meaning senses in the COCA (Garnier & Schmitt, 2015; see Sample Study 17.1). Similarly, the AKL words have been described in the *Louvain English for Academic Purposes* dictionary (LEAD) and students majoring in English at the University of Louvain (Belgium) use the web-based dictionary-cum-writing aid in the EAP classroom and at home. The LEAD contains a rich description of academic words, with a particular focus on their phraseology (collocations and recurrent sequences) (cf. Granger & Paquot, 2015; Paquot, 2012). The lexical entries provide information derived from an analysis of a large corpus of academic texts (i.e., the academic component of the *British National Corpus*), as well as a range of discipline-specific corpora and English as a Foreign Language (EFL) learner corpora representing a wide range of first language (L1) populations.

### Sample Study 17.1

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19(6), 645–666.

#### Background

It is notoriously difficult for practitioners to select the items to be taught among the thousands of phrasal verbs in English. Available lists of phrasal verbs (PVs) are of limited value because they do not account for polysemy, which is a serious pedagogical shortcoming as research indicates that phrasal verbs have on average 5.6 meaning senses.

#### Research Gap

The main objective of Garnier and Schmitt's (2015) study was to develop a pedagogical list of PVs that would be helpful to teachers and learners by providing their most frequent meaning senses together with example sentences.

#### Research Method

- Selection of a previously published corpus-based list of the 150 most frequently used PVs in American and British English

- Identification of their most frequent meaning senses in the COCA.
- Two criteria for inclusion in the PHrasal VERb Pedagogical List (PHaVE List): (1) the meaning senses for each item should account for at least 75% of all occurrences of this phrasal verb in their corpus search, and (2) all the meaning senses should account for at least 10% of the same occurrences

### Key Results

Garnier and Schmitt showed that a large majority of the most frequent PVs in English are polysemous, and that, on average, around two meaning senses account for at least 75% of all the occurrences of a single PV in the *Corpus of Contemporary American English*. The PHaVE list thus focuses on the key meaning senses of the 150 most frequent PVs and gives their percentage of occurrence, along with definitions and examples written to be accessible for second language learners.

### Comment

Garnier and Schmitt used corpus data to inform the selection of key meaning senses to be included in the PHaVE list but created example sentence themselves. As discussed above, this will be regarded by some as a limitation. However, among other things, Garnier and Schmitt took care of modelling sentences from various sources found on the Internet as well as from the COCA itself, with the aim to produce as natural and authentic sentences as possible. As a result, I believe the PHaVE is to be commended as the output of a truly corpus-based pedagogic enterprise.

## (Lexis-)grammar in Context

Analyses of raw and grammatically annotated corpora have led to a radical change in the way grammar is conceived of. First, corpus-based research has consistently shown that grammatical patterns differ systematically across registers, that is, language varieties determined by their purposes and situations of use (e.g., newspaper writing vs. academic writing vs. fiction writing). To illustrate, Conrad (2000) shows that linking adverbials—words and expressions such as *however*, *therefore*, and *in other words* that explicitly connect two units of discourse—are used much less frequently in newspaper writing than in academic prose. Second, corpus-based analyses have been instrumental in the development of the concept of spoken or conversational grammar. Among the many contributions these studies made was “the incorporation into the purview of grammar of what, in previous centuries, had typically been regarded as marginal word-classes or as aberrations, items such as discourse markers [*well*, *sort of*, *just*] and interjections [*oh*, *ouch*]” (Carter & McCarthy, 2015, p. 5). Third, corpus-based analyses have demonstrated the patterned nature of language and close connections between grammar and lexis. Lexicogrammatical co-selection phenomena are ubiquitous in language (e.g., the verb *deem* is mostly used with the passive voice; cf. Biber, Johansson,

Leech, Conrad, & Finegan, 1999), which implies that certain grammatical constructions ought to be described in relation to lexical items and vice versa. These developments have had impact on reference tools, notably grammars. For example, the *Longman Grammar of Spoken and Written English* (LGSWE, Biber et al., 1999), the first comprehensive reference grammar to adopt a corpus-based approach, describes the use of grammatical features in academic prose, news, fiction, and conversation, with a special focus on lexicogrammatical patterning.

Textbooks for language learning and teaching lag behind. Quite a few studies have compared the use of language features in general reference corpora with the pedagogical descriptions of the same items in pedagogical material, and most notably course books (cf. Römer, 2005). Meunier and Reppen (2015) have recently shown that major publishers have tried to answer a repeated call for more corpus-informed materials. They carried out a case study on the treatment of the passive in corpus versus non-corpus-informed grammar textbooks published between 2002 and 2012 and showed that unlike non-corpus-informed textbooks, textbooks which are presented as corpus-informed do a better job at describing what we know from corpus research about the use and the lexical associations of the passive. More generally, however, there is still room for improvement, especially in the treatment of collocations and lexicogrammatical patterns. Another area that deserves special attention is the role of learner corpus data.

## The Role of Learner Corpus Data

Learner corpora are systematic collections of continuous, and contextualized language spoken or written by L2 learners stored in electronic format. Since the inception of learner corpus research (LCR) at the turn of the 1990s, the field has always been driven by its desire to have an impact on language teaching and learning. The rationale is that while language descriptions based on native/expert corpora should typically be used to define the teaching agenda and ensure that the language taught reflects typical language use, the learner language descriptions should be used to modify this agenda to take into account the difficulties, needs, and unique trajectories of learners (e.g., Meunier, 2002).

One popular method to analyse learner corpus data is Computer-aided Error Analysis (CEA; Dagneaux, Denness, & Granger, 1998). This method focuses on the identification, correction, and annotation of errors according to a standardized system of error tags with the help of a software tool: an “error editor.” CEA has significant advantages over traditional error analysis.

First, it relies on an error taxonomy that makes the error tagging procedure both more systematic and coherent. Error taxonomies are usually based on structural linguistic categories and include tags for errors that relate to form (i.e., spelling and morphology), grammar, lexis, lexico-grammar, punctuation, word order, and so forth. Second, it enables users to select an error category, for example, that of morphology errors, extract a comprehensive list of all the errors of this type in the learner corpus, sort them in various ways, and analyse them in context.

Despite the controversial status of errors in second language acquisition (SLA) and applied linguistics in general, results of computer-aided error analyses have been among the first types of learner corpus data to find their way in pedagogical materials, especially in monolingual learners' dictionaries. In the second edition of the *Macmillan English Dictionary for Advanced Learners* (MEDAL2, 2007), for example, "Get it right" boxes contain authentic erroneous examples from the *International Corpus of Learner English* (ICLE; Granger, Dagneaux, Meunier, & Paquot, 2009), clear explanations of the source of the problems, and practical advice on how the errors can be corrected and avoided. With a view to ensuring maximum representativeness of the error notes, only errors that were both frequent and widespread across the 16 mother tongue backgrounds represented in the ICLE were covered (Gilquin, Granger, & Paquot, 2007). The focus on errors is also in evidence in a range of recent pedagogical grammars mostly published by Cambridge. In *Grammar and Beyond 4* (2013), for example, each unit features an "Avoid Common Mistakes" box informed by careful analysis of the error-annotated subset of the *Cambridge Learner Corpus* (CLC) as illustrated in Fig. 17.1. The CLC

## Avoid Common Mistakes

---

- 1. Do not use *who* with inanimate nouns.**

*that*  
A study *who* showed the benefits of being an only child was published last year.
- 2. Do not omit the relative pronoun in subject relative clauses.**

*who*  
Children *^* have older siblings tend to be somewhat dependent.
- 3. Remember that the subject and the verb must agree in relative clauses.**

*have*  
Children who *has* siblings often become secure and confident adults.
- 4. Use the same *as*, not the same *than*.**

*as*  
Middle children often have the same level of creativity *than* last-born children.

Fig. 17.1 *Grammar and Beyond 4*, "Avoid Common Mistakes" box (p. 75)

contains over 60 million words mostly of written exam scripts produced by English language learners taking Cambridge English exams all over the world, a large proportion of which is error-annotated.

The most large-scale research project based on the CLC, that is, the English Profile Programme, however, adopts the opposite approach: instead of focusing on errors, it aims primarily at investigating what grammar and vocabulary learners can typically use correctly and appropriately at each level of the *Common European Framework of Reference for Languages* (CEFR, Council of Europe, 2001). Results of the English Vocabulary Profile and the English Grammar Profile are available online in the form of a listing of known vocabulary (e.g., the verb “to agree” is attested at A2 level in its meaning of “to have the same opinion as someone” and at B1 in its meaning of “to decide something with someone”) and Can Do statements targeting core grammatical features such as tenses, articles, modal and auxiliary verbs, and word order by CEFR level (e.g., a B1 learner can use the past continuous with an increasing range of adverbs in the normal mid position) (<http://englishprofile.org/>).

While the outcomes of the English Programme Profile should prove most useful for curriculum and syllabus design, material design, and assessment, they are also limited in one important way: the exclusive focus on first appearance of specific vocabulary items or grammatical features in learner language. Thus, the conjunction “because” and the adverb “however” are described as A1 and A2 level words respectively; the approach adopted does not make it possible to highlight that these two connectors are then used with extremely high frequency by intermediate and advanced learners at the expense of other means of expressing cause and contrast. To uncover such patterns of overuse and underuse, the method of Contrastive Interlanguage Analysis (CIA; Granger, 1996) has been extensively used in learner corpus research. CIA involves two types of comparison:

1. A comparison of one or more interlanguage varieties with one or more reference varieties which aims to shed light on a range of linguistic features that characterize learner writing and speech, “not only errors but also instances of under- and overrepresentation of words, phrases and structures” (Granger, 2002, p. 12);
2. A comparison of different interlanguage varieties of the same language, i.e. the English of French learners, Spanish learners, Dutch learners, and so forth so as to “differentiate between features which are shared by several learner populations and are therefore more likely to be developmental and those which are peculiar to one national group and therefore possibly L1-dependent” (Granger, 2002, p. 13).



**BE CAREFUL!** When learners express degrees of possibility or certainty, they often limit themselves to modal auxiliaries, and neglect the many other expressions that can be used for the same purpose. Verbs, adverbs, adjectives, and nouns can all express various degrees of certainty, and these are discussed in the following sections.

Fig. 17.2 “Be careful note” on the overuse of modal auxiliaries (MEDAL2, p. 17)

Patterns of overuse and underuse shared by various learner populations should be used to inform generic tools such as pedagogical dictionaries and textbooks. However, this is still extremely rare. One exception is MEDAL2 where learner corpus-based information from ICLE was used not only to compile error notes but also to inform an extended writing section focusing on 12 rhetorical or organizational functions particularly prominent in academic writing (e.g., comparing and contrasting, expressing cause and effect). The analysis of learner corpora and their comparison with native corpora highlighted a number of problems which non-native learners from various mother tongue backgrounds experience when writing academic essays (e.g., lack of register awareness, phraseological infelicities, semantic misuse, and restricted lexical repertoire). These sources of difficulties are addressed explicitly in MEDAL2: Fig. 17.2, for example, shows a “Be careful!” note which warns the reader against the excessive use of modal auxiliaries to express the function of expressing possibility and certainty. It is followed by a series of verbs, adverbs, adjectives, and nouns which, though common in native academic writing, are rarely used by learners.

The fact that most learners’ dictionaries and other pedagogical materials are generic, that is, target all learners whatever their first language, has precluded the inclusion of notes and exercises that focus on difficulties that are specific to one particular L1 or family of L1s to date. With the generalized use of the Internet and online pedagogical tools, however, more customization is expected to become the norm, and the LEAD, which contains both generic and L1-specific error notes that show up depending on the first language selected by the user, will probably not remain the only one of its kind (see above). This is highly desirable since results from learner corpus research have repeatedly demonstrated the influence of the mother tongue on correct and incorrect language use by (even advanced) learners, most particularly perhaps at the phraseological level (e.g., Paquot, 2015, 2017; see Sample Study 17.2).



### Sample Study 17.2

Paquot, M. (2017). L1 frequency in foreign language acquisition: Recurrent word combinations in French and Spanish EFL learner writing. *Second Language Research*, 33(1), 13–32.

#### Background

Empirical evidence suggests that learners are sensitive to L2 frequency (Ellis, 2002). This sensitivity applies not only to individual lexical items but to the frequency of L2 collocations, lexical bundles and constructions as well (e.g., Ellis, O'Donnell, & Römer, 2014). Cross-linguistic influence is another important variable in SLA, but it is usually approached in terms of the potential influence on the L2 of the presence or absence of a linguistic phenomenon in the L1.

#### Research Gaps

Despite the large number of studies focusing on cross-linguistic influence and current interests in frequency effects in SLA, there has been very little research on the potential influence of L1 frequency on L2 acquisition and use. The study thus aimed to answer the following research questions:

1. To what extent is there a relationship between the frequency of a lexical bundle in the EFL learners' written productions and the frequency of its equivalent form in the learners' first language?
2. Can French and Spanish learners' different use of lexical bundles be explained by frequency differences in the two Romance languages?

#### Method

- Extraction of statistically significant three-word lexical bundles in the Spanish and French sub-components of the second version of the *International Corpus of Learner English* (ICLE; Granger et al., 2009)
- Manual selection of lexical bundles with discourse or stance-oriented function
- Use of Jarvis' (2000) unified methodological framework for the study of L1 influence
- Identification of L1 translational equivalent forms
- Frequency comparisons across the two L1s on the basis of web corpora of French and Spanish

#### Results

Strong and positive monotonic correlations were found between the frequency of a lexical bundle in the EFL learners' written productions and the frequency of its equivalent form in the learners' first language. Results also confirm that different patterns of use across the two L1 learner populations (e.g., Spanish learners' frequent use of the lexical bundle "in the case of") may be explained by frequency differences in L1 French and Spanish (Spanish "en el caso de" is much more frequent than French "dans le cas de").

#### Comment

The lexical bundle approach represents "corpus linguistic methodology at its most heuristic, i.e. as a raw discovery procedure" (De Cock, 2004, p. 227). Coupled with Jarvis' (2000) framework, it proved most useful to identify in learner corpora transfer effects that until now have been little documented in the SLA literature.

## Challenges and Controversial Issues

As already highlighted above, corpus evidence does not necessarily equate with pedagogical relevance and the results of corpus studies should always be balanced with other factors such as learners' proficiency levels, learners' needs, and teaching objectives before they are used to inform curriculum design and teaching material development (see also Granger, 2015). Results of corpus studies, however, also crucially depend on the type of corpora used. What corpus should be used to inform English Language Teaching (ELT) around the world? Should a general reference corpus or a specialized corpus be used? Should the corpus be made up exclusively of native language production or is it acceptable if the corpus includes expert or near-native data irrespective of mother tongue background? What is the target language variety or norm? Should a corpus of British English or American English be favoured? In what circumstances should we opt for a corpus representing another English language variety? Is there a role for corpora of English as a lingua franca in ELT? The way we answer these questions will impact heavily on the type of corpus findings obtained. Similar questions also need to be addressed when a learner corpus is compared with another corpus to bring out learners' difficulties. Depending on the objectives of the study, learner corpus researchers also need to decide whether they want to compare EFL learner language with expert and/or student productions (cf. Ädel 2006, pp. 206–207 for a discussion).

Another controversial issue relates to the use of authentic texts from corpora in language teaching. While Garnier and Schmitt (2015), for example, illustrate the different meanings of phrasal verbs with made-up examples, other researchers believe that corpus-derived (sometimes simplified) examples should be used because “intuition is often unreliable when it comes to patterns of language use” (Meunier & Reppen, 2015, p. 501). The debate is far from resolved and more research is needed into whether and to what extent learners benefit from exposure to authentic examples in the L2 classroom, especially as a function of variables such as the learners' proficiency levels, the teaching objectives, and the topic of the course.

## Limitations and Future Directions

More insights from corpus linguistics, not just frequency information, need to find their way into curriculum design and material development. Future textbooks and other pedagogical tools should provide learners with a more integrated perspective on lexis and grammar; they should also place more emphasis on word combinations such as collocations and phrases and focus

on variability in language use. I have also argued elsewhere for a corpus-based context-sensitive treatment of phraseology in (electronic) learners' dictionaries whereby collocations would not appear in undifferentiated lists but be organized by domain, mode, or register (Paquot, 2015).

For such a "revolution" to happen, however, there is a need for more pedagogical development and translation to make the results of corpus-based studies really usable in the classroom. Recent corpus-based resources such as the new GSL developed by Brezina and Gablasova (2015) are raw material and not as pedagogically useful as they could be (i.e., unlike West's (1953) GSL, the new GSL comes with no rich description such as meaning senses or examples). More generally, the implications of current corpus-based research for pedagogy are usually not developed in any great detail. As Meunier argues in a discussion of learner corpus research and L2 language learning,

[t]here is an urgent need to go beyond the usual last paragraph of articles (or last slide of conference presentations) stating that 'foreign language instruction could profit from this kind of investigation' and efforts should be made towards providing teachers with ready-to-use teaching materials. (2011, p. 465)

More systematic and large-scale comparisons of textbooks and other pedagogical resources with corpus data, of the type initiated by Römer (2005), should also be conducted, targeting not only materials produced by major publishing houses but also by local publishers, with a view to better document the discrepancies between textbook language and "real" language.

Ultimately, however, the future role of corpus-based findings in L2 syllabus and material design will probably rest on the field's ability to answer a repeated call for more pedagogically relevant corpora (cf. Meunier, 2011). Until very recently, learner corpus research has focused on English for academic and specific purposes in higher education. As a result, a large majority of learner corpora consist of exam scripts or argumentative essays; more learner corpora such as the Spanish Learner Language Oral Corpora (SPLLOC, <http://www.splloc.soton.ac.uk/index.html>) representing less advanced proficiency levels as well as other domains, registers, and modes are definitely needed. Importantly, for similar developments as the ones witnessed in ELT to occur in the teaching and learning of other foreign languages, there is an urgent need for more, better and bigger corpora of languages other than English. In a methodological synthesis of learner corpus research, Paquot and Plonsky (2017) found that 88% of the 378 studies in their sample examined English as the target language.

## Resources for Further Reading

Biber, D., & Reppen, R. (2015). *The Cambridge handbook of corpus linguistics*. Cambridge: Cambridge University Press.

The *Cambridge handbook of corpus linguistics* offers a comprehensive survey of corpus-based linguistic research on English, including chapters on collocations, phraseology, grammatical variation, discourse, and the description of English varieties (e.g., dialects, World Englishes, learner language) and a critical discussion of the state of the art in the field. Part IV presents corpus research into a variety of areas that are particularly relevant to language learning and teaching such as corpus-derived vocabulary lists and classroom applications of corpus analysis.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24.

In this article, Granger discusses the major criticisms Contrastive Interlanguage Analysis (CIA), a comparative framework for analysing learner corpus data, has faced since its introduction in 1996 and presents a revised model, CIA<sup>2</sup>, which, she argues, better acknowledges the important role played by variation in LCR and is generally more in line with the current state of foreign language theory and practice.

Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

The *Cambridge handbook of learner corpus research* provides a unique account of learner corpus collection, annotation, methodology, theory, analysis, and applications. Part III is devoted to the links between learner corpus research and SLA. Part IV will be of special interest to all readers who want to know more about the applications of learner corpora to teaching, pedagogical material development, and testing. Part V explores the diverse and growing natural language processing tasks that rely on learner corpora.

Leech, G. (2011). Frequency, corpora and language learning. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 7–31). Amsterdam & Philadelphia: Benjamins.

In this chapter, Leech considers how the “corpus revolution” made frequency information available in a totally unprecedented way from the 1960s onward. Then, he discusses the equation “more frequent = more important to learn,” what questions of frequency we really need to ask, and how far they can be answered in the present state of corpus linguistics.

Timmis, I. (2015). *Corpus linguistics for ELT research and practice*. New York: Routledge.

This volume is a user-friendly overview of corpus research in selected areas of language study relevant to ELT: lexis, grammar, English for specific purposes, and spoken language. It provides a practical guide to undertaking ELT-related corpus research and bringing corpus material to the classroom.

## Notes

1. Work on frequency lists has also spread to other languages, as seen in the “Routledge Frequency Dictionaries” series which now offers corpus-based core vocabulary (alphabetical, frequency, and thematically organized) lists for learners of 14 languages including Arabic, Dutch, French, German, Japanese, Mandarin Chinese, Portuguese, Russian, and Spanish.
2. Thus, if a word occurs with relatively high frequency in the corpus but is only found in a limited number of corpus parts of the same length, the ARF will be low.
3. For example, the Academic Keyword List has been used to inform the *Oxford Learner’s Dictionary of Academic English* and is presented as a major asset of the five-level academic English course *Skillful* published by Macmillan Education. Similarly, the Academic Collocation List appears in an appendix of the *Longman Collocations Dictionary and Thesaurus*.

## References

- Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: Benjamins.
- Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL)—A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4), 235–247.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.
- Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), 1–22.

- Carter, R., & McCarthy, M. (2015). Spoken grammar: Where are we and where are we going? *Applied Linguistics (Advance Access)*, 38(1), 1–20.
- Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of corpus linguistics* (pp. 478–497). Cambridge: Cambridge University Press.
- Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, 34(3), 548–560.
- Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, 34(2), 213–238.
- Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System: An International Journal of Educational Technology and Applied Linguistics*, 26(2), 163–174.
- De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures*, 2, 225–246.
- Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, 24(2), 143–188.
- Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, 4(4), 405–431.
- Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 75–93). New York and London: Routledge.
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305–327.
- Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19(6), 645–666.
- Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319–335.
- Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37–51). Lund: Lund University Press.
- Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, S. Petch-Tyson, & J. Hulstijn (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (Vol. 6, pp. 3–33). Amsterdam and Philadelphia: John Benjamins.
- Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 485–501). Cambridge: Cambridge University Press.
- Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *The International corpus of learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

- Granger, S., & Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1), 118–141.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Lessard-Clouston, M. (2012). Word lists for vocabulary learning and teaching. *The CATESOL Journal*, 24(1), 287–304.
- Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics*, 33(3), 299–320.
- Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung, S. Petch-Tyson, & J. Hulstijn (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 119–142). Amsterdam and Philadelphia: John Benjamins.
- Meunier, F. (2011). Corpus linguistics and second/foreign language learning: Exploring multiple paths. *Revista Brasileira de Linguística Aplicada*, 11(2), 459–477.
- Meunier, F., & Reppen, R. (2015). Corpus versus non-corpus-informed pedagogical materials: Grammar in focus. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of corpus linguistics* (pp. 498–514). Cambridge: Cambridge University Press.
- Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. London and New York: Continuum.
- Paquot, M. (2012). The LEAD dictionary-cum-writing aid: An integrated dictionary and corpus tool. In S. Granger & M. Paquot (Eds.), *Electronic lexicography* (pp. 163–185). Oxford: Oxford University Press.
- Paquot, M. (2015). Lexicography and phraseology. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of corpus linguistics* (pp. 460–477). Cambridge: Cambridge University Press.
- Paquot, M. (2017). L1 frequency in foreign language acquisition: Recurrent word combinations in French and Spanish EFL learner writing. *Second Language Research*, 33(1), 13–32.
- Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, 3(1), 61–94.
- Römer, U. (2005). *Progressives, patterns, pedagogy. A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam: Benjamins.
- Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *ELT Journal*, 31(4), 487–512.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- West, M. (1953). *A general service list of English words*. London: Longman.





# 18

## Digital Discourses Research and Methods

Christoph A. Hafner

### Introduction

Digital discourses, by which I mean digitally mediated texts and interactions and their associated ideologies, are having a profound effect on the kinds of real-world communication practices that applied linguists study. Firstly, digital tools feature more and more prominently in such communication practices: indeed, when it comes to written communication, the use of digital tools is now seen as the norm. As Warschauer, Zheng, and Park (2013, p. 825) put it, “there is little serious writing that is not done digitally.” As a result, applied linguists are increasingly likely to encounter the use of digital tools when studying issues of communication in traditional domains of applied linguistic inquiry: institutions, professions and workplaces are becoming increasingly reliant on digital discourses. In some cases, the technology clearly plays a central role in establishing effective communication: take the case of virtual workplaces, where multinational teams use remote collaboration tools to achieve their goals without recourse to face-to-face meetings (Lockwood & Forey, 2016). In other cases, the technology may seem to play a somewhat more peripheral role, as where lawyers perform negotiations by marking up versions of an agreement with the reviewing tools of a word processor and sharing their drafts over email (Townley & Jones, 2016).

---

C. A. Hafner (✉)

Department of English, City University of Hong Kong, Kowloon, Hong Kong  
e-mail: [elhafner@cityu.edu.hk](mailto:elhafner@cityu.edu.hk)



Secondly, as digital tools have become more and more integrated into people's lives, there has been a reordering of the social contexts that form the focus of applied linguistic research. Health care provides an instructive example, but there are many others. With digital media, some of the roles and functions that have traditionally been performed by health care professionals are gradually being devolved to networks of interested amateurs. One example of this is the "quantified self" movement, whose members make use of technological devices to constantly record health-related data: heart rate, posture, movement, mood, food intake, exercise, sleep, all can be digitally observed, logged and the data aggregated over time (Jones, 2015). Members share and discuss this information and its implications for their health in dedicated online social networking sites, which may or may not be moderated by medically trained professionals. In this way, a new and powerful digital discourse has emerged, with important consequences for traditional health care institutions. As yet, there is still little applied linguistic study of such innovative digital practices, even though they clearly impact directly on the applied linguistic research agenda. The applied linguistic field of inquiry must expand to take such user-generated initiatives into account.

Finally, digital discourses offer applied linguists a plentiful new source of easily accessible data. Many digital communication tools record, archive and store interactions by default. As a result of this "persistence," researchers can obtain records of communications that might not otherwise be available. In a recent case study of collaborative project-based learning in Hong Kong, David Li, Lindsay Miller and I used a range of ethnographic techniques to learn about the language practices of undergraduate students as they went about completing an English language project (Hafner, Li, & Miller, 2015). One set of data that we collected comprised the more than 6000 digitally mediated messages (emails, Facebook posts, WhatsApp posts) that the students sent one another. These messages provided us with a detailed record of authentic, out-of-class interactions between university students in the plurilingual environment of Hong Kong. Students were able to send us logs of interactions that had occurred over a number of months. It would have been much harder to capture the same kind of data using traditional observation methods. Nevertheless, the use of such digital data sets raises a number of challenging questions. How do we collect data in an ethical way? How do we process the data that we find? How do we analyse it? As we discovered in our analysis of students' digital interactions, existing discourse analytical tools may not be up to the task.

## Current/Core Issues

When it comes to studying digital discourses as part of an applied linguistics research project, it is important to understand that these digital discourses do not exist in a contextual vacuum. Instead, they are part of a larger, real-world problem that the applied linguist is interested in. The examples that I have provided above—virtual workplace teams collaborating remotely, lawyers negotiating agreements between clients, self-quantifiers tracking and sharing health information, university students communicating about an English project—all are situated in real-world activities of which they make up just a part. Sometimes the research questions may lead to a focus on the digital discourses alone. At other times, the questions may lead to an examination of how such digital discourses link up with other “offline” texts and interactions as part of a larger language ecology. At all times, the analysis of digital discourses can be supported by accounts of the people in institutions, professions and workplaces, who are participating in the digital practices in focus. What I am getting at here is this: the starting point for applied linguistic study of digital discourses is the same as any other applied linguistics research project. It begins with the identification of a real-world problem, in which language plays a central role. Digital discourses provide one possible source for an answer.

It is also important to realize that digital discourses are not just digital versions of analogue text types, transported onto the internet (though these exist, too). In this chapter, we are primarily concerned with the features of digital discourses that stand out and deserve special attention. Studies of language and digital media show that the affordances of digital tools, that is, what we perceive digital tools to enable us to do, provide new resources for creating texts, interacting, building new kinds of identities and relationships. As Rodney Jones and I have pointed out, digital media provide us with the tools to establish new ways of doing, meaning, relating, being and even thinking (Jones & Hafner, 2012). In studying digital discourses, it is therefore important to begin with a solid understanding of how the affordances and constraints of digital tools, coupled with social factors of use, impact on fundamental concepts for applied linguistic study, like text, interaction and context. Both the affordances and constraints of the tools themselves and their social “cultures-of-use” (Thorne, 2003), the way that they have come to be used by social groups of users, are important here. For this reason, Herring (2007) uses the term “socio-technical mode” “to refer not just to CMC systems, but also to the social and cultural practices that have arisen around their use” (p. 3).

## Context

Early research into digital discourses assumed a fairly clear division between physical/virtual, online/offline activity (e.g., Turkle, 1995). We now understand, however, that these spaces overlap in various ways to create a layered context for interaction, which can, of course, shape the form of communication. The context can consist of (1) what/who is in the physical environment; (2) what/who is in the virtual environment or on the communication platform; (3) what/who is on the screen (see Jones, 2002). Each of these contextual layers can have an influence on the nature of the online interaction, as described below:

1. The physical environment: Individuals engaged in online communication may use different language and communication strategies when others are present, especially if the other is someone's parent or their boss. This aspect of context has become more complicated, thanks to the ubiquity of mobile devices, which can so easily be accessed and consulted, even when in conversation with physically co-present others.
2. The virtual environment: Virtual environments like online social networks bring about a kind of "context collapse" (Marwick & boyd, 2011) bringing together contacts from different walks of life: school, work, friends, family and so on. This can affect the form of communication, as the intended audience for communications becomes blurred.
3. The screen environment: Individuals may shift their attention between multiple different communication platforms on screen, and this may affect the way that messages are composed. They can also begin a conversation in one platform (e.g., an instant messaging app) and continue it in another (e.g., an online social network), a possibility which complicates data collection and analysis.

To this we can add that the individual's experience of the communication platform may vary on different devices: compare the look and feel of the Facebook website with its mobile app. Additionally, with algorithms that customize advertising or the order of search results, different individuals often have a different experience of the same platform. Such differences in experience can also raise issues for data collection, data processing and data analysis. In summary, the context of digital discourses is multilayered and highly complex. Researchers need to develop appropriate strategies that take these contextual features into account.

## Text and Interaction

Digital tools provide affordances for meaning making that ultimately allow us to make new kinds of texts and have new kinds of interactions. Elsewhere, I have pointed out that digital composition offers the writer resources of hyper-text, multimodality and interactivity (Hafner, 2013). Briefly, a digital text can easily incorporate references (hyperlinks) to other texts and point readers to them, potentially altering the reading experience. Digital texts can also easily combine multiple modes, such as speech, writing, graphics, sound, still and moving images. Finally, digital texts allow for interactions between writers and readers, as in the comments function of a blog or on a Wikipedia talk page. As a result of all of these technical affordances, we see new practices emerging in digital discourses: collaboration on a massive scale; remixing of digital content, involving the appropriation of digital texts, which are repurposed and recontextualized (Iedema, 2003) for different audiences, often leading to the development of hybrid genres.

Herring (2007) has developed a faceted classification scheme that provides a comprehensive list of factors that can have an impact on language in digital discourse. The scheme is divided into medium factors and situation factors. Medium factors include synchronicity (does the communication occur simultaneously?), message transmission (one-way vs. two-way), persistence of transcript, size of message buffer (i.e., characters allowed in a message), channels of communication (e.g., video, audio, text), anonymous messaging, private messaging, filtering, quoting, message format (i.e., the order in which messages appear). Situation factors include participation structure, participant characteristics, purpose, topic or theme, tone, activity, norms, code. Often, digital discourses involve novel groupings of participants, where the normal roles are reversed, as when teenagers, experts at online role-playing games, mentor professionals much older than themselves.

We often take the characteristics of digital discourses for granted. But the increased multimodal nature of the communication and potential for interaction can fundamentally change the quality of text and interaction. Let's consider the example of Twitch.tv, an online real-time video streaming service, which is often used by gamers to demonstrate gameplay to a massive online audience. Twitch.tv allows users to set up channels where they can host live gaming sessions. Simplifying a little, each channel is divided into two main areas: near the middle of the screen is a video interface. This displays a live screen recording of the game that the host is playing as well as an inset showing

the host, commenting on the gameplay. On the right of the screen is a text chat area, where the thousands of viewers participating can enter comments. Some of these comments will be read out loud by the host, who will respond to them while playing the game (but will not respond via text). Viewers can also interact with the host by making a donation, which appears as a text message on the streaming video screen and the host usually responds to these messages with some kind of “thank you” and by reading out the donor’s username.

In Twitch.tv, multimodal semiotic resources are combined: the gameplay screencast, the video of the host, the chat area. Interactions cross from one semiotic mode to another, as where a user asks a question in chat and the host responds through the video interface. As a result, analysing one mode on its own (e.g., the chat log) will not sufficiently capture the interaction. In any case, in conversation analytic terms, the chat log is likely to display the kind of “disrupted” turn adjacency that Herring has observed elsewhere (Herring, 1999) partly because of the massive numbers of people involved. In order to make sense of the interaction, it is necessary to construct some kind of multimodal transcription system that can account for what is going on in each of the socio-technical modes at any given point in time and how these “goings-on” relate to each other (e.g., Recktenwald, 2017). Traditional tools of conversation and discourse analysis must be adapted for this platform: the affordances of digital tools to facilitate the use of multiple semiotic modes in interactions happening simultaneously between massive numbers of participants create a novel form of text and interaction.

Another platform that draws on the multimodal semiotic resources of digital tools is YouTube. Benson’s (2015) study of informal language learning, described in Sample Study 18.1, is interesting because of the way that an analytical system is developed to categorize the discourse of YouTube comments and search for patterns of learning in the interactions.

### Sample Study 18.1

Benson, P. (2015). Commenting to learn: Evidence of language and intercultural learning in comments on YouTube videos. *Language Learning & Technology*, 19(3), 88–105.

#### Research Background

In the field of language and literacy education, it has been observed that the internet provides a new context for autonomous language learning. The many globalized online spaces that have sprung up—fan fiction sites, virtual worlds and gaming sites, video-sharing sites—provide language learners with new opportunities to interact and develop their language skills. In spite of these observations, little empirical evaluation of these new learning contexts has been carried out.

### Research Problems/Gaps

This exploratory and methodologically oriented study aimed to develop the necessary analytical tools to identify evidence of language and intercultural learning in the online video-sharing site, YouTube.

### Research Method

- Comments were collected from YouTube videos that involved translanguaging, that is, videos that combined Chinese with English.
- A snowball sampling procedure was used: first, videos that were well-known in the Hong Kong context of the study were identified; then, related videos suggested by YouTube were also examined.
- Ten different genres of videos were identified, and within each genre, the videos with the most comments were collected, initially yielding 8850 comments.
- 2840 comments that specifically related to issues of language or culture were retained for analysis, while the rest of the comments were discarded.
- A quantitative analysis was conducted by manually coding comments using Sinclair and Coulthard's (1975) exchange structure framework (Initiation, Response, Evaluation), Stenström's (2014) taxonomy of interactional acts (e.g., opine, evaluate, challenge), three different kinds of stance marking in the discourse (cognitive activity, status of knowledge and sources of knowledge).
- A qualitative microanalysis of a single exchange was performed.

### Key Results

While only exploratory in nature, the study showed how discourse analytical tools could be employed to provide evidence of language and intercultural learning in YouTube comments on translanguaging videos. The quantitative analysis demonstrated a rich interactional context in comments focusing on language and culture. This could in theory create an environment that is conducive to learning. The associated microanalysis provided stronger evidence of learning, showing how one commenter progressed from a state of confusion about Chinese song lyrics to a state of understanding. The analysis illustrated the potential process of learning involved.

### Comments

This computer-mediated discourse analysis combined quantitative coding with qualitative analysis. The analysis had to take into account multiple modes. In the context of the YouTube page, the video itself was considered an initiation move, which comments responded to in various ways. According to the analysis, 45% of the comments were part of an exchange structure involving the video and at least two comments. An analysis of the comments alone would not sufficiently have accounted for the interaction.

## Challenges and Controversial Issues

We have seen that the affordances of digital tools can fundamentally alter the nature of context, text and interaction. Not only this, but new digital tools are constantly emerging, each one changing the rules of the game again, ever so slightly. Thus, one of the most pressing challenges is to develop appropriate

approaches to the study of digital discourses, with appropriate methods of data collection, processing and analysis. A possible controversial issue is that of research ethics: in a world where conceptions of privacy are rapidly changing, how can researchers ensure that they adopt appropriate standards of ethics? I will begin this section by providing the briefest of sketches of some prominent approaches to the study of digital discourses (constraints of space permit no more). I will then outline some of the issues related to data collection, processing and analysis. Finally, I will provide a brief discussion of the question of ethics.

## Approaches to the Study of Digital Discourses

Here I will outline four prominent approaches: digital ethnography; computer-mediated discourse analysis; corpus analysis; network analysis.

1. Digital ethnography: This approach applies ethnographic techniques like participant observation to the study of online communities and their online communication practices. It usually involves collecting and analysing a large amount of different kinds of data: texts and interactions, interviews with community members, stated principles or regulations of the community and so on. The strength of this approach is the in-depth understandings that it can generate by focusing in detail on communication practices in particular cases. See Varis (2016) for a detailed account.
2. Computer-mediated discourse analysis (CMDA): Herring (2004, p. 339) defines CMDA as “any analysis of online behaviour that is grounded in empirical, textual observations.” The analysis of text can be supported by surveys, interviews or ethnographic observations. The primary analysis is a language-focused content analysis, which can be either qualitative or quantitative. If it is a quantitative analysis, it involves systematically coding and counting language features, similar to Benson’s study in Sample Study 18.1. This approach aims to investigate features of online discourse and social practices.
3. Corpus analysis: A quantitative approach which makes use of corpus tools in order to generate linguistic profiles of texts. The analyst must first compile a suitably large corpus of digital texts, for example, posts to an online forum, before the linguistic features can be quantitatively analysed.
4. Network analysis: Network analysis provides a quantitative, mathematical way of analysing a network. The main units of analysis are nodes and arcs which show how different “units” (nodes) are related to each other in the network. Visualizations can be produced based on these which also allow



the analyst to take into account the frequency of connections in the network. This approach can be used “to observe the structure of a community by examining communication among its members” (Paolillo, 2016, p. 39). Network analysis usually relies on massive sets of data, including meta-data (e.g., the location, time, method of communication of online posts or texts).

There are increasing calls to mix methods, in order to benefit from the power of quantitative methods to spot patterns and the ability of qualitative methods to generate rich insights. Page (2016) provides an interesting example of a study that combined quantitative network analysis with qualitative interactional sociolinguistics in order to investigate online trolling behaviour in a spoof Facebook page dedicated to the late Margaret Thatcher. Analysis of meta-data helped to identify an atypical user (the “troll”), and microanalysis identified a range of positioning strategies used. Related to the call to mix quantitative and qualitative methods in this way, applied linguists also need to develop ways to collect large amounts of data, including the kind of meta-data described above. Some studies have employed the application program interface (API: the routines, protocols and tools that developers use to build programmes based on other platforms) of web services like YouTube in order to automate data collection to some extent and convert the “chaos” into “data” (Androutsopoulos & Tereick, 2016, p. 365).

## Data Collection, Processing and Analysis

Another challenge concerns the processing and analysis of digital data. Some digital discourses may be relatively straightforward to process, as long as the analysis is limited to linguistic, not multimodal, features. For example, text chat or tweets can generally be cut and pasted without much modification. Even so, the analyst will have to make decisions about how to deal with features that are more or less native to digital discourse, like hyperlinks, hashtags, emoticons and embedded images or videos. Such features are akin to paralinguistic gestures in speech and can add significant meanings. In processing data for our study of students’ CMC communications on a collaborative English project, we annotated links to files, images and videos that students were sharing, in order to indicate the dominant language and whether these linked texts were drafts created by students or material found on the internet (Hafner et al., 2015). Where the analyst is interested in the multimodal aspect of the communication, a robust system of multimodal transcription is needed. Baldry and Thibault (2006) provide a useful exploration of this topic, and the



work of O'Halloran and colleagues (O'Halloran, Tan, Smith, & Podlasov, 2013) on the development of multimodal corpora is also instructive. The example of Twitch.tv that we considered above shows how streaming video and text chat interact in real time, requiring a transcription system that not only accounts for multimodal features (e.g., of the video) but also accounts for time.

While analytical tools like discourse analysis, genre analysis and conversation analysis can be applied to digital discourses, this is not as straightforward as one might think. Let's take the example of conversation analysis of online interaction, for example, in text chat, forums and social media (a more detailed discussion can be found in Giles, Stommel, Paulus, Lester, & Reed, 2015). Some of the assumptions of conversation analysis tend to unravel in digital discourses. First, in digital interactions, it can be difficult to clearly identify a conversation, its beginning and its end. With the persistence of digital discourses, online interactions can be picked up and continued after lengthy time gaps: do such delayed contributions continue the conversation or start a new one? Second, and related to that, digital interactions do not follow the relatively coherent, linear patterns of speech. Many turns are unanswered and, even in interactions between two participants, it is possible to have multiple simultaneous conversation threads. Third, conversations can cross digital platforms, beginning in one space and continuing in another. Fourth, in public spaces anyway, turns may be addressed not only at individuals but also at the group, for example, a forum community. Fifth, the system itself, governed by algorithms, can become a participant in the conversation. For example, in social media groups on platforms like Facebook or WhatsApp some posts are system messages that record admin actions like creating the group or adding a member. More research is needed to develop theories that can address such analytical challenges.

## Research Ethics

The study of digital discourses raises new ethical challenges for applied linguistic research (see Sterling & De Costa, Chap. 8). These challenges stem from particular socio-technical features of digital media: changing conceptions of privacy, which mean that individuals may discuss issues traditionally considered private, in the context of public or semi-public communities on the internet; persistence, the tendency for digital communications,

including such private discussions, to be permanently recorded; and searchability, the ability of users, including academic readers, to use search engines to locate these permanent records. A key question is to what extent applied linguistic researchers can treat information on the publicly accessible internet, for example, in internet forums or social media, as public information and therefore useable for research purposes without further consent of participants. Generally, such information would not have been compiled with knowledge of participants that it would be used as data in academic research. This question may be answered differently for different kinds of participants: compare the corporate twitter feed of a news organization like the BBC with a health forum, where members of the public share and discuss personal health problems (see Angouri & Sanderson's (2016) study, described in Sample Study 18.2).

From an ethical point of view, a key concern is whether the human subjects research model should be applied to research of digital discourses and therefore, on a very practical level, whether informed consent should be obtained (Bassett & O'Riordan, 2002). Adopting different metaphors to understand the internet is likely to affect the answer to this question. As Hudson and Bruckman (2004, p. 128) point out, "For some researchers, the Internet is like a public square, and for others, a private living room, a town hall meeting, or a newspaper letters column." However, their study also shows that, even though participants in chatrooms are posting in publicly accessible forums, they nevertheless view the chatroom as a kind of private space, where research is often unwelcome. They ultimately argue that, in some studies, researchers may need to rely on IRBs to issue a formal waiver of consent. This requires a consideration of issues of: (1) consent (whether it can practically be obtained); (2) harm (what the potential for harm is and the sensitivity of the data); (3) data protection and retention (how long and who will have access); (4) anonymization (whether identifiers are removed and when); and (5) credit (whether participants desire credit for their work). On the issue of harm, it can be especially difficult to identify minors in online data sets, making it hard to exclude them. Also, Varis (2016) points out the issue of searchability, which has an impact on the reporting of digital ethnographies that focus on sensitive online discussions. In order to protect anonymity, reporting may need to be limited so that readers are not able to search utterances and link people to the study.

As mentioned above, one study that had to deal with these kinds of ethical issues is Angouri and Sanderson's (2016) work on interactions in a health forum.

### Sample Study 18.2

Angouri, J., & Sanderson, T. (2016). 'You'll find lots of help here': Unpacking the function of an online Rheumatoid Arthritis (RA) forum. *Language & Communication, 46*(1), 1–13.

#### Research Background

Online health forums, where lay internet users interact with one another about a particular medical condition which they suffer from, are becoming increasingly common. Such forums are thought to empower the patients who use them. Patients gain access to others in similar situations, who provide emotional support and information. Such forums may also have implications for the role of health care professionals and possibly even threaten professional dominance. However, little research has been conducted into the way that users of such forums discursively construct a sense of online community and the roles and functions they perform. This article used a combination of thematic analysis and discourse analysis to address the issue.

#### Research Problems/Gaps

This article aimed to describe and explain how members of an online Rheumatoid Arthritis (RA) forum “project roles on the group and construct their community in their postings” (p. 3).

#### Research Method

- A RA forum with 5065 registered members (25% estimated to be active) was identified.
- Gathering data raised ethical issues: participants who had contributed posts to the forum were not aware, at the time that they posted, that their contributions would be used as data for a research project in this way.
- Researchers therefore gained support from the “arthritis partner” (p. 4), ensured that the project followed forum terms of use and created a post to the forum to explain the project and provide an opportunity for forum members to ask for their username to be excluded from the research. Where users self-disclosed that they were under 18, their posts were excluded from the research.
- Researchers collected 12 popular forum threads for analysis and focused their analysis on postings related to “new diagnosis,” interactions between new members to the forum and more experienced members.
- Thematic analysis of the forum posts was conducted, focusing on the semantic content and coding for emerging patterns or themes.
- A discourse analysis of the postings was performed, focusing on the linguistic forms of the messages.
- Researchers combined a macroanalysis of broader themes in the data with a microanalysis of interactions.

#### Key Results

In the forum observed, a diagnosis of RA was a prerequisite for community membership. Members provided each other with a range of different kinds of support. They performed task-orientated functions such as providing information and emotional support, for example, sharing narratives of their own experience. They also performed rapport-orientated functions, identifying through shared

experiences and encouraging new members to use the forum. Angouri and Sanderson point out that for many chronic conditions, the health system is designed to intervene only when these conditions reach crisis point. They conclude with an observation about the potential importance of such online forums: "online communities provide exactly what is often reported as 'missing' from the medicalised model of treatment, a holistic understanding of the condition and 'expertise' in the mundane daily activities (Rodriguez, 2013) where 'normal' is redefined by the patient" (p. 7).

#### Comments

This study incorporates elements of digital ethnography in its design. Angouri and Sanderson describe their analysis as "ethnographically informed" (p. 4), meaning that the analysis was not only based on the posts themselves but also informed by talks with the arthritis charity, moderators of the forum and users, where they could be involved.

## Limitations and Future Directions

This chapter has highlighted a number of challenges related to the use of digital discourses in applied linguistics research. Resolving these challenges will require concerted research efforts. If I were to highlight one area of particular interest for future research in applied linguistics, it would be the changing nature of relationships in digital media, including the way that digital discourses have broadened audiences and blurred distinctions between specialists and non-specialists. An important feature of some digital discourses is that they are part of a participatory culture (Jenkins, 2006), in which individual members of the public are producers of information as well as consumers. Members of institutions and professions now have powerful tools of communication, like social media for example, that they can use to reach out to people in order to further their goals. However, such specialists share the resulting spaces with well-informed, well-organized, networks of amateurs (Lessig, 2004), whose cultural productions and contributions to knowledge have been very impressive. As individuals now go about their lives, informing themselves about their political choices, health issues and legal options, they will often be getting such information not from traditional sources but instead from user-generated sources. Such a shift obviously has a tremendous impact on the applied linguistic research agenda. Going forward, it is going to become increasingly important to understand such user-generated spaces, what role specialists and non-specialists have to play in them, and how they contribute to institutional and professional communities that have been the traditional focus of applied linguistic inquiry.

## Resources for Further Reading

Georgakopoulou, A., & Spilioti, T. (Eds.). (2016). *The Routledge handbook of language and digital communication*. Abingdon, Oxon: Routledge.

Reputable scholars from a range of applied linguistics disciplines address issues, including issues of methods and analysis, in research on language and digital media.

Jones, R. H., Chik, A., & Hafner, C. A. (Eds.). (2015). *Discourse and digital practices: Doing discourse analysis in the digital age*. Abingdon, Oxon: Routledge.

Well-known scholars in discourse analysis and language and literacy education discuss the question of how to “do” discourse analysis in digitally mediated environments.

Page, R. E., Barton, D., Unger, J. W., & Zappavigna, M. (2014). *Researching language and social media: A student guide*. Abingdon, Oxon: Routledge.

A practical guide on how to design and carry out research projects on language in social media.

## References

- Androutsopoulos, J., & Tereick, J. (2016). YouTube: Language and discourse practices in participatory culture. In A. Georgakopoulou & T. Spilioti (Eds.), *The Routledge handbook of language and digital communication* (pp. 355–370). Abingdon, Oxon: Routledge.
- Angouri, J., & Sanderson, T. (2016). ‘You’ll find lots of help here’: Unpacking the function of an online Rheumatoid Arthritis (RA) forum. *Language & Communication*, 46(1), 1–13.
- Baldry, A., & Thibault, P. J. (2006). *Multimodal transcription and text analysis*. London: Equinox.
- Bassett, E. H., & O’Riordan, K. (2002). Ethics of Internet research: Contesting the human subjects research model. *Ethics and Information Technology*, 4(3), 233.
- Benson, P. (2015). Commenting to learn: Evidence of language and intercultural learning in comments on YouTube videos. *Language Learning & Technology*, 19(3), 88–105.

- Giles, D., Stommel, W., Paulus, T., Lester, J., & Reed, D. (2015). Microanalysis of online data: The methodological development of "digital CA". *Discourse, Context & Media*, 7, 45–51.
- Hafner, C. A. (2013). Digital composition in a second or foreign language. *TESOL Quarterly*, 47(4), 830–834.
- Hafner, C. A., Li, D. C. S., & Miller, L. (2015). Language choice among peers in project-based learning: A Hong Kong case study of English language learners' plurilingual practices in out-of-class computer-mediated communication. *Canadian Modern Language Review*, 71(4), 441–470.
- Herring, S. C. (1999). Interactional coherence in CMC. *Journal of Computer-Mediated Communication*, 4(4).
- Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behaviour. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for virtual communities in the service of learning* (pp. 338–376). Cambridge and New York: Cambridge University Press.
- Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4(1), 1–37.
- Hudson, J. M., & Bruckman, A. (2004). "Go away": Participant objections to being studied and the ethics of chatroom research. *The Information Society*, 20(2), 127–139.
- Iedema, R. (2003). Multimodality, resemiotization: Extending the analysis of discourse as multi-semiotic practice. *Visual Communication*, 2(1), 29–57.
- Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York: New York University Press.
- Jones, R. H. (2002). The problem of context in computer mediated communication. In P. LeVine & R. Scollon (Eds.), *Discourse and technology: Multimodal discourse analysis* (pp. 20–33). Washington, DC: Georgetown University Press.
- Jones, R. H. (2015). Discourse, cybernetics, and the entextualization of the self. In R. H. Jones, A. Chik, & C. A. Hafner (Eds.), *Discourse and digital practices: Doing discourse analysis in the digital age* (pp. 28–47). Abingdon, Oxon: Routledge.
- Jones, R. H., & Hafner, C. A. (2012). *Understanding digital literacies: A practical introduction*. London: Routledge.
- Lessig, L. (2004). *Free culture: How big media uses technology and the law to lock down culture and control creativity*. New York: Penguin Press.
- Lockwood, J., & Forey, G. (2016). Discursive control and power in virtual meetings. *Discourse & Communication*, 10(4), 323–340.
- Marwick, A. E., & boyd, d. m. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114–133.
- O'Halloran, K. L., Tan, S., Smith, B. A., & Podlasov, A. (2013). Multimodal digital semiotics: The interaction of language with other resources. *Text & Talk*, 33(4–5), 665–690.

- Page, R. (2016). Moving between the big and small: Identity and interaction in digital contexts. In A. Georgakopoulou & T. Spilioti (Eds.), *The Routledge handbook of language and digital communication* (pp. 403–407). Abingdon, Oxon: Routledge.
- Paolillo, J. C. (2016). Network analysis. In A. Georgakopoulou & T. Spilioti (Eds.), *The Routledge handbook of language and digital communication* (pp. 36–54). Abingdon, Oxon: Routledge.
- Recktenwald, D. (2017). Toward a transcription and analysis of live streaming on Twitch. *Journal of Pragmatics*, 115, 68–81.
- Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. London: Oxford University Press.
- Stenström, A. B. (2014). *An introduction to spoken interaction*. London: Routledge.
- Thorne, S. L. (2003). Artifacts and cultures-of-use in intercultural communication. *Language Learning & Technology*, 7(2), 38–67.
- Townley, A., & Jones, A. (2016). The role of emails and covering letters in negotiating a legal contract: A case study from Turkey. *English for Specific Purposes*, 44(1), 68–81.
- Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. New York: Simon & Schuster.
- Varis, P. (2016). Digital ethnography. In A. Georgakopoulou & T. Spilioti (Eds.), *The Routledge handbook of language and digital communication* (pp. 55–68). Abingdon, Oxon: Routledge.
- Warschauer, M., Zheng, B., & Park, Y. (2013). New ways of connecting reading and writing. *TESOL Quarterly*, 47(4), 825–830.

# Part III

## Data Analysis

There are ten chapters in this part of the Handbook. In each chapter, the authors present a conceptual overview of one or more fundamental approaches to data analysis in applied linguistics. Authors also discuss current and core issues, key procedures, challenges and controversial issues, and limitations and future directions. The authors also provide resources for further reading.

- In Chap. 19 (Correlation and Simple Linear Regression in Applied Linguistics), Reza Norouzian and Luke Plonsky provide an applied description of two key methods to evaluate the association between two research variables. The authors discuss the various behaviors of the correlation statistic, Pearson  $r$ , under different scenarios. The authors also present a conceptually overarching view that links the regression methods to all other methods that applied linguists often use to detect important patterns in their data.
- In Chap. 20 (Exploratory Factor Analysis), Aek Phakiti explains how exploratory factor analysis (EFA) is used to explore common factors that account for participants' responses to research instruments, such as Likert-type scale questionnaires and tests. The author also provides an overview of important aspects, considerations, and practical guidelines for conducting EFA.
- In Chap. 21 (Confirmatory Factor Analysis and Structural Equation Modeling), Aek Phakiti examines the core principles of confirmatory factor analysis (CFA) and structural equation modeling (SEM) that can be used in applied linguistics research. The author provides essential guidelines for not only how to read CFA and SEM reports but also how to perform CFA using EQS Program.



- In Chap. 22 (Analyzing Group Differences), Luke Wander Amoroso identifies appropriate uses of analysis of variance (ANOVA), explains how to conduct a statistically sound ANOVA, and also discusses null hypothesis testing and the importance of reporting and interpreting effect sizes in the context of means-based analyses. The author also describes the process of checking statistical assumptions, interpreting statistical output, and calculating effect sizes.
- In Chap. 23 (Statistics for Categorical, Non-parametric, and Distribution-Free Data), Jesse Egbert and Geoffrey T. LaFlair describe statistical methods for handling such data. The authors introduce methods for analyzing categorical data, including the use of basic descriptive statistics such as measures of central tendency, measures of dispersion, frequency counts, and normed rates of occurrence. They then introduce when, why, and how to use alternatives to traditional parametric statistical tests, including non-parametric analogs, permutation tests, and bootstrapping.
- In Chap. 24 (Reliability Analysis of Instruments and Data Coding), Kirby Grabowski and Saerhim Oh begin with a general discussion of different types of reliability (both internal and external to an instrument itself) including the different indices and models used to estimate reliability and their respective interpretations. Throughout the chapter, guidelines for addressing common limitations with respect to reliability analysis and reporting in AL research are outlined, including suggestions for how to address these issues in operational contexts.
- In Chap. 25 (Analyzing Spoken and Written Discourse: A Role for Natural Language Processing Tools), Scott Crossley and Kristopher Kyle provide a general overview of research methods used in the analysis of both spoken and written discourse. The authors also provide a specific overview of how natural language processing (NLP) tools that measure lexical, syntactic, rhetorical, and cohesion features of text can be used to examine spoken and written discourse.
- In Chap. 26 (Narrative Analysis), Phil Benson introduces narrative analysis as a relatively recent addition to the toolkit of applied linguistics. Its basic premise is that the telling of stories can elucidate the meanings attached to participants' experiences. This chapter is mainly concerned with uses of storytelling and narrative writing in data analysis and presentation of research findings.
- In Chap. 27 (Interaction Analysis), Elizabeth Miller explores interaction analysis, a methodology used to examine the sequential development of turn-taking in interaction in order to understand how interlocutors create and/or transform social order in interaction. The author discusses its utility

for examining how language learning is practiced, afforded, and constrained in specific situations and how individuals demonstrate learning through their changing participation in particular discursive practices, as co-constructed activity.

- Finally, in Chap. 28 (Multimodal Analysis), Jesse Pirini, Tui Matelau, and Sigrid Norris present the five most prominent approaches to multimodal data analysis: Multimodal (Inter)action Analysis, Mediated Discourse Analysis, Systemic Functional Multimodal Discourse Analysis, Social Semiotics, and Multimodal Conversation Analysis.



# 19

## Correlation and Simple Linear Regression in Applied Linguistics

Reza Norouzian and Luke Plonsky

### Introduction and Conceptual Overview

Correlations and regression analysis form a critical foundation for many of the statistics found most commonly in applied linguistics research. Nearly all quantitative research in the field, in fact, relies in some way or another on these procedures or some extension thereof (i.e., the general linear model (see Norouzian & Plonsky, 2018; Plonsky, 2013)). Some of the many other statistical procedures based on correlation and regression include *t*-test, analysis of variance (ANOVA), exploratory and confirmatory factor analysis, structural equation modeling, and multi-level modeling (Cohen, 1968), which are covered elsewhere in this handbook (e.g., chapters by Amoroso and Phakiti, respectively) as well as in other, recent volumes (e.g., Norris, Ross, & Schoonen, 2015; Plonsky, 2015a; Roever & Phakiti, 2017).

Why are correlations and regression analyses, often referred to as “analyses of association,” used so frequently? Put simply, they are both useful and applicable in a wide variety of research situations. Before addressing some of the more theoretical and statistical issues associated with this type of analysis, we will outline two of the most important and common scenarios in which analyses of association are found.

---

R. Norouzian  
Texas A&M University, College Station, Texas, USA  
e-mail: [rnorouzian@gmail.com](mailto:rnorouzian@gmail.com)

L. Plonsky (✉)  
Northern Arizona University, Flagstaff, AZ, USA  
e-mail: [lukeplonsky@gmail.com](mailto:lukeplonsky@gmail.com)

The first and perhaps most common use of analyses of associations is found when scholars in this field, as in many other educational and social sciences, explore certain types of expectations or hypotheses. More specifically, a correlation can be used when researchers are interested in examining and quantifying the relationship between two continuously measured variables. These kinds of relationships between variables are often quite intuitive and do not necessarily require statistical evidence to be understood: height and weight, L2 vocabulary knowledge and reading comprehension, class attendance and course grades, motivation and achievement. To provide a concrete example, in their now-classic study, Johnson and Newport (1989) used a correlation, which is expressed with an  $r$  value that can range from  $-1$  to  $+1$ , to examine the relationship between age of arrival and grammar knowledge. The result of  $r = -0.77$  was interpreted as fairly strong evidence to indicate that those who arrived in the target-language environment at younger ages achieved higher grammatical proficiency (see Sample Study Box 19.1). Likewise, in the realm of corpus linguistics, Egbert and Plonsky (2015) were interested in examining the relationship between linguistic and stylistic features of conference abstracts and the ratings they received from reviewers. Among a number of other variables examined, they found that abstract length, which they measured in total number of words, correlated fairly well with abstract ratings ( $r = 0.38$ ). In other words, longer abstracts tended to receive higher scores.

A second type of scenario when analyses of association are used is to examine the reliability and/or validity of a quantitative measure such as a language test. Although they are not always viewed as such, nearly all indices used to express interrater reliability, internal consistency (e.g., Cronbach's alpha) and concurrent validity are based on correlations (see Plonsky & Derrick, 2016, for an overview and guide to interpreting reliability estimates; see Grabowski, Chap. 24). We can also use correlational analyses to establish the concurrent validity of two or more instruments. For example, although many institutions use the Test of English as a Foreign Language (TOEFL) to inform admissions decisions regarding students' academic English ability, some universities may prefer to develop their own in-house instruments. In order to test the concurrent validity of the in-house test, the university might give a sample of students in the target population both the in-house test and a TOEFL and then run a correlation between the two. A high correlation in this case would be taken as evidence of the (concurrent) validity of the in-house assessment.

In the section that follows, our discussion delves deeper into the purpose and applications of the correlation. Specifically, we provide a conceptual view of the notion of correlation between two research variables. Using small datasets, we discuss the various behaviors of the correlation statistic, Pearson's  $r$ , under different scenarios. Then, we turn our attention to a neighboring but practically different concept within the analyses of association: *regression*. By end of the chapter,

we provide an overarching view that links the regression to virtually all other methods that applied linguists use to find important pattern in quantitative data.

## Pearson's Product-Moment Correlation Coefficient ( $r$ )

The most familiar form of analysis of association, based on the relationship between two continuously measured variables, is certainly the *linear correlation* ( $r$ ) popularized by Karl Pearson (Pearson, 1896). The mathematics of Pearson's  $r$  (including why it is called *product-moment*) is beyond the scope of this chapter (but see Rosenthal & Rosnow, 2008). Conceptually, however, we need to know what it means for the scores forming two continuous variables to be linearly correlated with each other.

We explore this question and the very nature of what it means for two variables to be correlated by first using a small dataset used for illustration purposes. Imagine we have obtained data on two variables,  $X$  and  $Y$ , from six participants as shown in Table 19.1. We can imagine  $X$  and  $Y$  to represent any two measured variables such as vocabulary knowledge and reading comprehension, years of study and proficiency, or working memory and performance.

As shown in Table 19.1, the six students have each scored differently on the variables  $X$  and  $Y$ . However, upon closer scrutiny, it becomes apparent that despite these differences in scores, variables  $X$  and  $Y$  share two important characteristics. First, both  $X$  and  $Y$  have exactly the same shape as indicated by the two *shape statistics*: *skewness* and *kurtosis*. Second, scores on variables  $X$  and  $Y$  are ordered in the same manner. To better illustrate this second characteristic in Table 19.1, the scores on variables  $X$  and  $Y$  also appear in ranks (e.g., the smallest score = 1, second smallest score = 2, and so on) next to the original

**Table 19.1** Two variables showing a *perfectly positive* Pearson's  $r$

Student no.	$X$	$X_{(\text{rank})}$	$Y$	$Y_{(\text{rank})}$
1.	1.00	1	3.00	1
2.	2.00	2	4.00	2
3.	3.00	3	5.00	3
4.	7.00	6	9.00	6
5.	5.00	5	7.00	5
6.	4.00	4	6.00	4
Mean	3.67		5.67	
SD	2.16		2.16	
Skewness	<b>0.46</b>		<b>0.46</b>	
Kurtosis	<b>-0.30</b>		<b>-0.30</b>	
Pearson's $r$		(+ 1.00)		

*Note:* Quantities boldfaced are known as shape statistics

scores for each variable. These rank-transformed data are labeled  $X_{(\text{rank})}$  and  $Y_{(\text{rank})}$  in Table 19.1. Now, to determine its *upper limit*, Pearson's  $r$  conceptually asks two main questions of the two variables of interest to the researcher:

1. Do the scores forming the two variables have the same shape (i.e., “same shape statistics”)?
2. Do the two variables order their scores in the same order (i.e., “same ranks”)?

In fact, a perfectly positive Pearson's  $r$  is bound by  $+1$ . For example, without needing to compute a Pearson's  $r$  using a software package, because variables  $X$  and  $Y$  in Table 19.1 have exactly the same shape (see “shape statistics”), and because their scores appear in exactly the order (see “ranks”), we can see that they have a Pearson's  $r$  of exactly  $+1$ . Such a relationship would be characterized not only as positive but also as “perfect.” It is also entirely possible that the distribution of scores for  $X$  and  $Y$  could have the same shape but with the order of scores in the opposite (or nearly opposite) direction. In such a situation, Pearson's  $r$  would approach or reach its lower limit of  $-1$ . When Pearson's  $r$  reaches its lowest limit, we say the linear correlation between  $X$  and  $Y$  is “perfectly negative.” One scenario in applied linguistics when negative correlations have been observed is in studies of the relationship between age of onset and accentedness: the lower the age of onset, the more nativelike (i.e., higher) a learner's speech is often rated. Table 19.2 presents an example of perfectly negative Pearson's  $r$  for six different students. (Note: This and other mocked-up datasets in our chapter are based on very small samples, which allow for easier and more economical presentation. Quantitative applied linguistics is notoriously underpowered (see Plonsky, 2015b), and we do not intend to imply that such  $N$ s are adequate.)

In reality, it is extremely rare to observe scenarios such as these where there is a perfect correlation between two variables. Non-perfect correlations (i.e., any  $r$  between  $-1$  and  $+1$ ) can be due to the differences in (a) the two variables'

**Table 19.2** Two variables showing a *perfectly negative* Pearson's  $r$

Student no.	$X$	$X_{(\text{rank})}$	$Y$	$Y_{(\text{rank})}$
1.	1.00	<b>1</b>	9.00	<b>6</b>
2.	2.00	<b>2</b>	8.00	<b>5</b>
3.	3.00	<b>3</b>	7.00	<b>4</b>
4.	7.00	<b>6</b>	3.00	<b>1</b>
5.	5.00	<b>5</b>	5.00	<b>2</b>
6.	4.00	<b>4</b>	6.00	<b>3</b>
Mean	3.67		6.33	
SD	2.16		2.16	
Skewness	<b>0.46</b>		<b>0.46</b>	
Kurtosis	<b>-0.30</b>		<b>-0.30</b>	
Pearson's $r$		(-1.00)		

**Table 19.3** Imperfect  $r$  due to differences in *ordering*

Student no.	$X$	$X_{(\text{rank})}$	$Y$	$Y_{(\text{rank})}$
1.	1.00	<b>1</b>	5.00	<b>3</b>
2.	2.00	<b>2</b>	6.00	<b>4</b>
3.	3.00	<b>3</b>	3.00	<b>1</b>
4.	7.00	<b>6</b>	4.00	<b>2</b>
5.	5.00	<b>5</b>	9.00	<b>6</b>
6.	4.00	<b>4</b>	7.00	<b>5</b>
Mean	3.67		5.67	
SD	2.16		2.16	
Skewness	0.46		0.46	
Kurtosis	-0.30		-0.30	
Pearson's $r$		(+0.10)		

**Table 19.4** Imperfect  $r$  due to differences in *scores' shapes*

Student no.	$X$	$X_{(\text{rank})}$	$Y$	$Y_{(\text{rank})}$
1.	1.00	1	1.00	1
2.	2.00	2	3.00	2
3.	3.00	3	5.00	3
4.	7.00	6	8.00	6
5.	5.00	5	7.00	5
6.	4.00	4	6.00	4
Mean	3.67		5.00	
SD	2.16		2.61	
Skewness	<b>0.46</b>		<b>-0.61</b>	
Kurtosis	<b>-0.30</b>		<b>-0.65</b>	
Pearson's $r$		(+0.95)		

**Table 19.5** Imperfect  $r$  due to differences in *scores' shapes and ordering*

Student no.	$X$	$X_{(\text{rank})}$	$Y$	$Y_{(\text{rank})}$
1.	1.00	<b>1</b>	9.00	<b>6</b>
2.	2.00	<b>2</b>	2.00	<b>1</b>
3.	3.00	<b>3</b>	6.00	<b>3</b>
4.	7.00	<b>6</b>	7.00	<b>4</b>
5.	5.00	<b>5</b>	8.00	<b>5</b>
6.	4.00	<b>4</b>	5.00	<b>2</b>
Mean	3.67		6.17	
SD	2.16		2.48	
Skewness	<b>0.46</b>		<b>-0.87</b>	
Kurtosis	<b>-0.30</b>		<b>0.74</b>	
Pearson's $r$		(+0.16)		

shape of scores, (b) their ordering of scores, and/or (c) a combination of (a) and (b). Tables 19.3, 19.4 and 19.5 present an example of each of these situations. Boldfaced quantities in these tables correspond to situations (a), (b), and (c), respectively. These datasets also exemplify different types of relationships

between variables. Whereas the correlations in Table 19.3 ( $r = 0.10$ ) and Table 19.5 ( $r = 0.18$ ) might be characterized as positive and relatively small or relatively, the correlation in Table 19.4 ( $r = 0.95$ ) is quite strong. See comments below on interpreting correlation coefficients. To compute a Pearson's  $r$  in each of the above cases, a simple way is to use the Excel command: = Pearson (list  $X$  scores, list  $Y$  scores).

To summarize thus far, a correlation can characterize the relationship between two variables in a variety of ways. Depending on the shape and order of the data, correlations can be either positive or negative and can indicate a variety of relationships between variables ranging from no relationship ( $r = 0$ ) to a perfect relationship ( $r = |1|$ ).

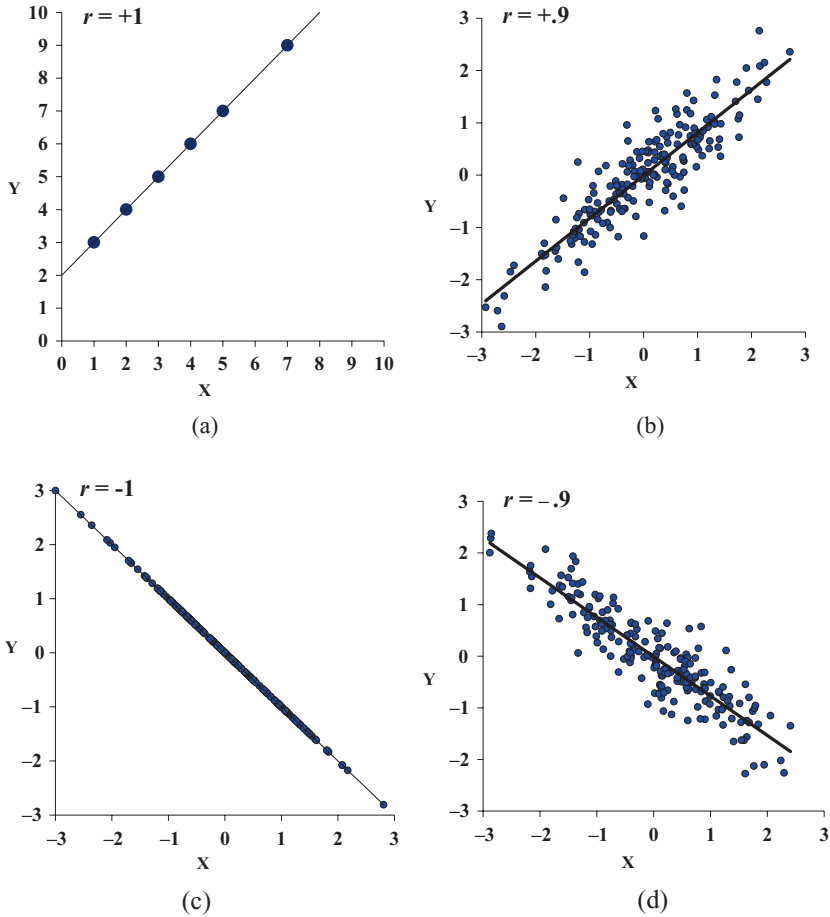
## Graphical Representation of Pearson's $r$ : Two Views

Up to this point, we have explored the nature of Pearson's  $r$  correlation on a largely conceptual and mathematical basis. This statistic and, more precisely, the associations it is used to explain can also be represented visually. In fact, the two views or plots that we discuss in this section, both of which can be used to display correlational data and the relationships they represent, can be more intuitive and informative than the correlation coefficient.

### The Scatterplot

A common way to visualize the correlation between two variables is through a “scatterplot.” A scatterplot is a diagram that has two axes corresponding to the variables being correlated. The  $X$  (horizontal) and  $Y$  (vertical) axes of the scatterplot in Fig. 19.1 (a) correspond to the range of scores on the two variables for our six participants ( $n = 6$ ) in Table 19.1. Given the perfectly positive relationship ( $r = +1$ ) between variables  $X$  and  $Y$ , a straight, upward-sloping line (referred to as a regression line) should run through all six pairs of scores. Figure 19.1 (b), (c), and (d), based on data for 200 students ( $n = 200$ ) simulated in the software package R, better illustrates the tendency of regression lines and their corresponding correlation coefficients to catch or represent many—but not all—of the data points in positively and negatively linear correlations. Scatterplots can be created quite easily using a number of statistical and spreadsheet-based programs such as Excel and SPSS, both of which provide point-and-click menus for this exact purpose.





**Fig. 19.1** Scatterplots of four samples of students' scores

These scatterplots and the regression lines laid over them are relatively unambiguous in representing very strong relationships between  $X$  and  $Y$ . Such magnitudes are rare. We generally observe much more moderate relationships. In fact, according to Plonsky and Oswald (2014), the typical (median) correlation in L2 research is about  $r = 0.37$ . Figure 19.2 provides a scatterplot for a correlation of this strength as well as for  $r = 0.25$  and  $r = 0.60$ , which Plonsky and Oswald found to be the 25th and 75th percentiles for correlation coefficients in L2 research, respectively. The authors tentatively propose these  $r = 0.25$  as representing a somewhat small correlation and 0.60 as somewhat large. These benchmarks cannot be applied universally, and the actual interpretation of a given correlation coefficient will depend on a number of factors.

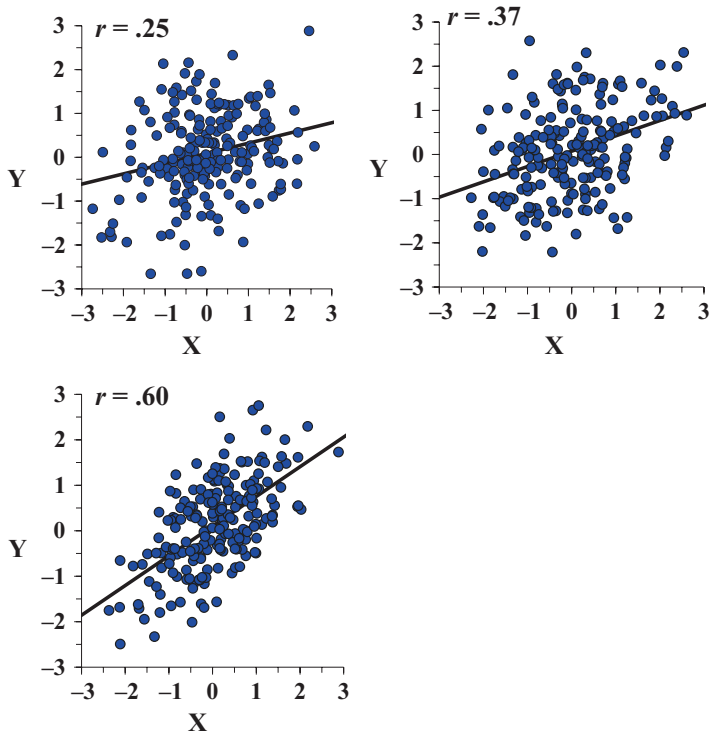


Fig. 19.2 Scatterplots indicating small, medium, and large  $r$  in L2 research

Nevertheless, developing an intuition for understanding the general strength of correlation, whether expressed as a coefficient or in a scatterplot, can be very useful. Toward this end we recommend playing (yes, playing!) the game found at <http://guessthecorrelation.com/>.

## Area View and the Notion of Shared Variance ( $r^2$ )

In the score view of Pearson's  $r$  just described, we represented *original scores* of participants on the two variables of interest using a scatterplot. We can also represent variables according to their respective dispersions in a squared metric (e.g., variance). In the *area view*, Pearson's  $r$  itself is expressed in a squared metric. When calculated, Pearson's  $r$  always standardizes variables (i.e., makes them have a mean of 0 and a variance of 1). Consequently, Pearson's  $r^2$  is equal to the proportion of variance shared between two standardized variables. Figure 19.3 represents two standardized variables' ( $X$  and  $Y$ ) variances with

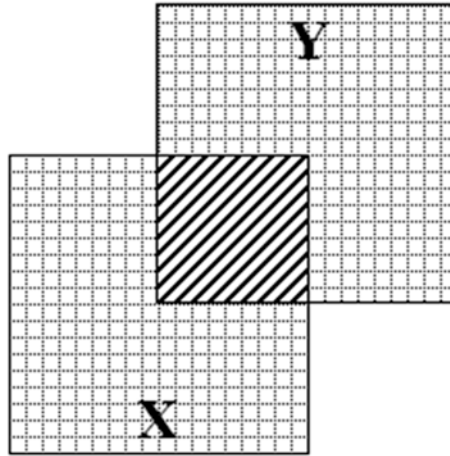


Fig. 19.3 Crosshatched area representing an  $r^2$  of 0.25 (25%)

two squares of the same size (as both have a variance of 1) where  $r = 0.5$  and, therefore,  $r^2$  between the variables is 0.25. This means that 25% of the variance in either variable is shared with the other. The small boxes inside each square in Fig. 19.3 illustrate how the two squares might overlap to show an  $r^2$  of 0.25. Note that in the area view, we are no longer able to determine the sign (– or +) of the relationship between two variables. This is because squaring necessarily removes the sign of Pearson's  $r$ .

## Regression: A Directional Analysis of Association

### Simple Linear Regression

Pearson's  $r$  is ideal if a researcher is simply interested in exploring how  $X$  and  $Y$  vary together (i.e., are associated). However, this discussion implies that Pearson's  $r$  is insensitive to the role of  $X$  and  $Y$  as research variables (i.e., dependent variable vs. independent variable). This is because asking about the correlation of  $X$  with  $Y$  is the same as asking about the correlation of  $Y$  with  $X$  (i.e.,  $r_{X\text{with } Y} = r_{Y\text{with } X}$ ). Figure 19.4 presents a simple view of the non-directional nature of Pearson's correlation.

Now, imagine an applied linguist is interested in knowing the extent to which length of residence ( $X$ ) in a target-language community can *predict* (not simply relate to) oral language ability ( $Y$ ). In this case, we can no longer arbitrarily decide on the roles of independent and dependent variables;

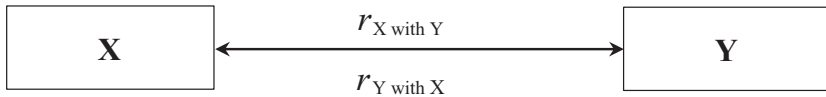


Fig. 19.4 Representation of Pearson's  $r$  as a non-directional measure

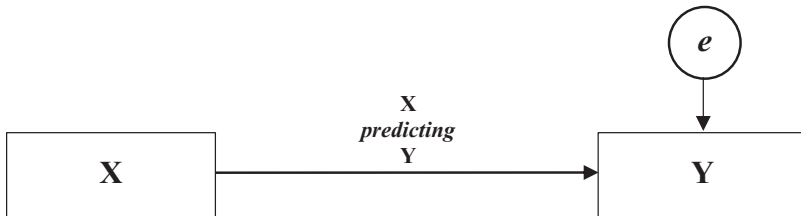


Fig. 19.5 Representation of regression as a directional measure

if we switch  $X$  and  $Y$  in terms of their role, we must accordingly reverse the wording of our research question! That is, the RQ would then be “To what extent can oral language ability predict (*not simply relate to*) length of residence?” In such situations, we must formally declare a model, known as a *linear regression model*, which makes explicit the directionality of our research question (i.e., we assign the role of dependent and independent variables). Such models only predict values of the dependent variable for given values of the independent variable. Thus, our diagrammatic view becomes directional as shown in Fig. 19.5.

Although regression models can take several forms, by far, the most common form in the social and behavioral research is that of the linear regression model. A *simple linear regression model* only involves one independent variable measured in some unit (e.g., years) and one dependent variable measured in some unit (e.g., an oral language ability score). The simple linear regression takes the form:

$$Y_i \leftarrow \hat{Y}_i = a + b(X_i), \quad (19.1)$$

where  $\hat{Y}_i$ , read *Y hat*, denotes the predicted value for the study participant  $i$  on the dependent variable ( $Y_i$ ) using his or her  $X_i$ , the score of the study participant  $i$  on the independent variable ( $X$ ) in our study.  $a$  and  $b$  are some numbers (coefficients) provided by any computer routine to optimize the prediction. The left pointing arrow ( $\leftarrow$ ), read *the prediction of*, indicates that the output of Eq. (19.1),  $\hat{Y}_i$ , is the prediction of the actual score of the study

participant  $i$  on the dependent variable ( $Y_i$ ) obtained in our study. Simply put, this equation is our linear forecasting device, which we refer to as a “linear model.” We can ask our model to linearly predict the value of the  $Y_i$  (dependent variable) for a given value of  $X$  (independent variable). Our linear forecasting model defines a straight line in the same manner that we saw in the case of Pearson’s  $r$ . Thus, the mathematics of the simple linear regression and of Pearson’s  $r$  are the same except that simple linear regression is directional and sensitive to the role of variables (dependent vs. independent) in the research question being posed. This is the most basic difference between correlation and (simple) regression when only two variables are involved in the analysis.

Let us now return to our example of length of residence (henceforth LR) in a target-language community predicting oral language ability (henceforth OLA). For simplicity’s sake, let’s assume LR is measured in years and OLA is measured using a rating scale with a maximum score of 25 points. Our hypothetically collected dataset (small, again, for illustration purposes) is shown in Table 19.6.

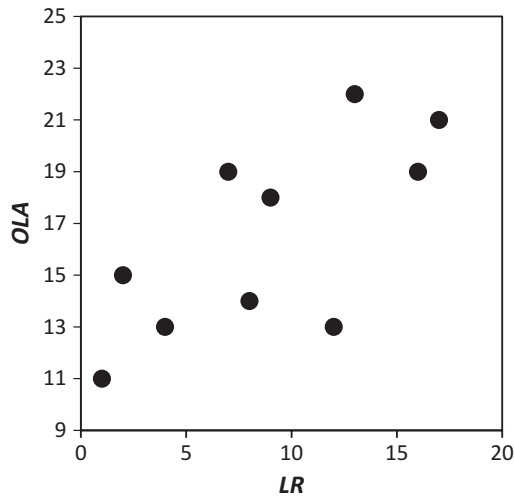
As noted earlier, the framework for regression analysis when only two variables (one predictor and one dependent) are involved is similar to that of Pearson’s  $r$ . We can also represent our *actual* scores collected on LR<sub>(years)</sub> and OLA on the  $X$ -axis and the  $Y$ -axis, respectively, to set up a scatterplot, as in Fig. 19.6. For purposes of clarity, the scatterplots for regression examples appear in boxes in this chapter.

**Table 19.6** Data for predicting OLA from LR ( $N = 10$ )

Participant no.	LR <sub>(Years)</sub>	OLA <sub>(Score)</sub>
1.	1	11
2.	8	14
3.	2	15
4.	9	18
5.	7	19
6.	4	13
7.	12	13
8.	13	22
9.	16	19
10.	17	21
Mean	8.90	16.50
SOS	280.90	128.50
Variance	31.21	14.27
Skewness	0.03	0.05
Kurtosis	-1.21	-1.48

*Note:* Parentheses next to variables’ names contain their unit of measurement

SOS = Sum of squares



**Fig. 19.6** Scatterplot for predicting OLA from  $LR_{(\text{years})}$

At this point, any statistical package can provide us with the information required for setting up the best straight line that can, on average, be closest to the actual study scores shown in the scatterplot. The output of the software package can also be used to fill in the general Eq. (19.1) presented earlier, as will become clear shortly.

Given its wide familiarity in the applied linguistics community, we use IBM® SPSS® (version 23) to run this simple linear regression example. After entering  $LR_{(\text{years})}$  and OLA scores into the Data Editor, we can locate *Regression* from the *Analyze* menu and then click on *Linear* (see Fig. 19.7).

This will open a dialogue box (Fig. 19.8) that enables us to drag and drop the dependent (OLA) and the independent ( $LR_{\text{year}}$ ) variables in the appropriate spaces.

Next, we need to click on *Statistics* (see Fig. 19.9) and then choose *Confidence Intervals* and press *Continue* as shown below. Finally, we can click *OK* to see the output.

## Making Sense of the Output

There are two key tables that we usually look for to answer the question “Do I have anything noteworthy in my results?” and make sense of the Statistical Package for the Social Sciences (SPSS) output. As will be shown, only if the answer to this question based on these two tables is “YES” can one proceed to prediction. The first table is called *Model Summary* as shown in the SPSS exhibit below.

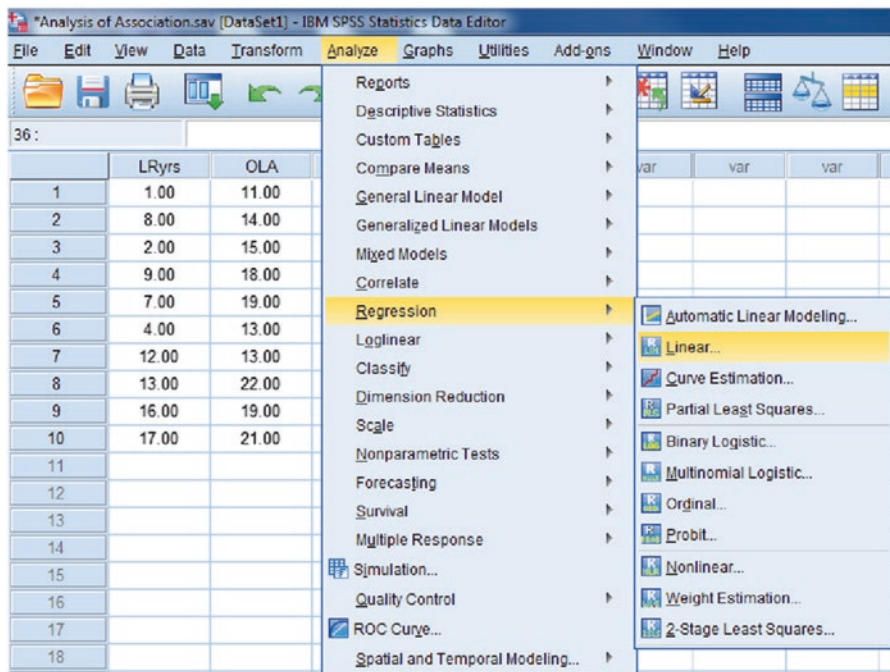


Fig. 19.7 Menu for selecting simple regression analysis in SPSS

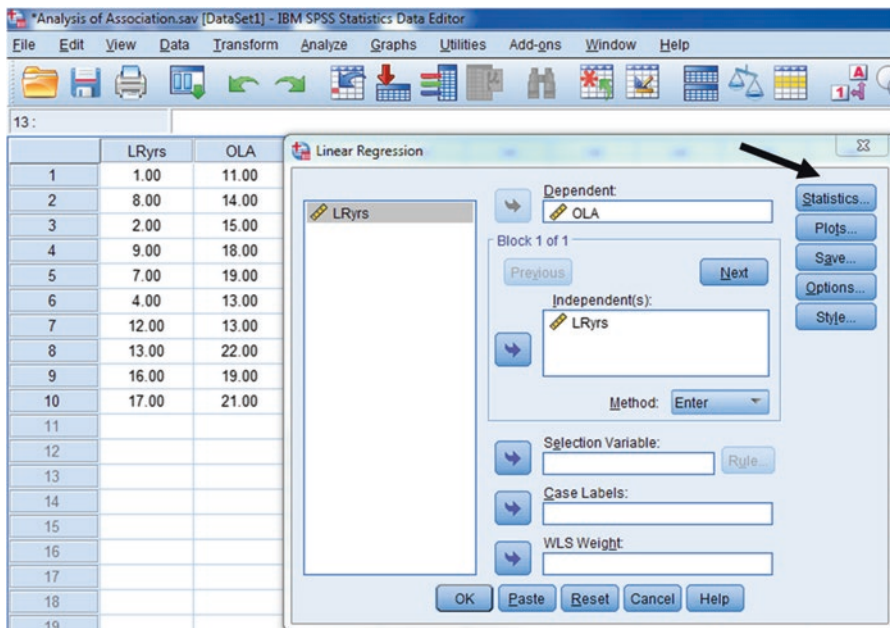


Fig. 19.8 Selections for running regression analysis in SPSS

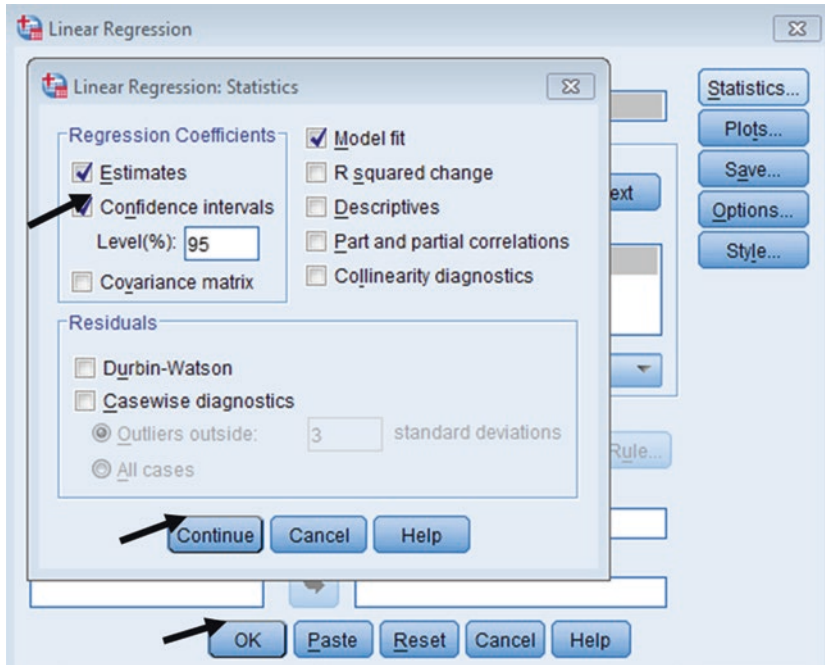


Fig. 19.9 Statistics for running regression in SPSS

Table 19.7 SPSS output of model summary from simple regression analysis

Model summary <sup>a</sup>				
Model	<i>R</i>	<i>R</i> square	Adjusted <i>R</i> square	Std. error of the estimate
1	0.708 <sup>b</sup>	0.501	0.439	2.831

<sup>a</sup>Dependent variable: OLA

<sup>b</sup>Predictors: (constant), LRys

Table 19.7, labeled *Model Summary*, shows an “*R*” of 0.708. *R* is a fit statistic. More specifically, *R* gives us information about how much our straight line fits our study data in the scatterplot. Thus, *R* ranges from 0 (complete lack of fit) to 1 (perfect fit). In reality, *R* can be measured after the specifications of our best-fitting line become available. We will soon see the specifications for our fitting line in one of the tables in the output reported by SPSS called *coefficients*.

“*R* Square” (henceforth  $R^2$ ) is the squared form of *R*. Similar to the  $r^2$  statistic discussed above,  $R^2$  gives us the area view of the amount of the variance in our OLA (dependent variable) scores that is predicted by (or overlapping with) the LR (predictor). In the next section, we will see how this  $R^2$  is obtained and then tested for statistical significance under the regression framework.



Next to  $R^2$  in the Model Summary table is “Adjusted  $R$  Square” (henceforth adjusted  $R^2$ ). In an ideal world, we want our adjusted  $R^2$  to be equal to our “ $R$  square” and not much smaller. Conceptually, adjusted  $R^2$  takes into account design problems as regards the number of study participants, number of predictors, and size of the  $R$ . Specifically, because in our heuristic example the number of study participants is very small ( $N = 10$ ), it is expected that adjusted  $R^2$  (0.439) would be smaller than  $R^2$  (0.501). Thus, adjusted  $R^2$  can be useful in arriving at a more conservative estimate of the ability of a given variable to predict another. The last piece labeled “Std. Error of the Estimate” (standard error of the estimate) will be discussed in the next section.

## Computation and Significance Testing of $R^2$

The second table of information, which we might look at to answer the question “Do I have anything noteworthy in my results?”, is the ANOVA table (Table 19.8). In the previous section we found that the proportion of variance in OLA (the dependent variable) that  $LR_{(\text{years})}$  can predict amounts to  $R^2$ . In common software packages, this  $R^2$  is computed and then tested against the null hypothesis of  $R^2 = 0$ . Both computation and significance testing of  $R^2$  are usually done via the ANOVA (analysis of variance) procedure as shown in the SPSS output below. Although most applied linguists view ANOVA and regression as distinct, they are quite the same from a statistical perspective and under the general linear model. In fact, as presented in Cohen (1968) we can think of one-way ANOVA as a special case of regression wherein a single categorical variable is used to predict outcomes in the dependent variable. For discussion of the relationship between these two tests in the realm of applied linguistics, see Plonsky and Oswald (2017).

The ANOVA procedure uses sum of squares (henceforth SOS) of the dependent variable (OLA) rather than variance as a dispersion index of the dependent variable to compute  $R^2$ . The reason for this is that ANOVA creates its own especial type of variances only used for significance testing purposes

**Table 19.8** ANOVA output table for simple regression analysis in SPSS

		ANOVA <sup>a</sup>				
Model		Sum of squares	<i>df</i>	Mean square	<i>F</i>	Sig.
1	Regression	64.401	1	64.401	8.038	0.022 <sup>b</sup>
	Residual	64.099	8	8.012		
	Total	128.500	9			

<sup>a</sup>Dependent variable: OLA

<sup>b</sup>Predictors: (constant), LRyrs

(labeled *Mean Square*). Instead, SOS is the true score dispersion index that we could obtain even before ANOVA procedure was conducted. For example, total SOS of 128.500 in this SPSS ANOVA table corresponds exactly to the SOS of our dependent variable (OLA) which we calculated when we first presented our data in the bottom portion of Table 19.6. When we use only one predictor, as in the case of our example study, the focus of ANOVA is on partitioning the total SOS in the dependent variable (OLA) into two parts, and list the size of those two parts in the ANOVA table. First we see the part that the predictor variable (LRyears) can predict out of the total SOS in the dependent variable (listed as *SOS Regression*). Second comes the part that the predictor variable (LRyears) cannot predict (listed as *SOS Residual*, which we can think of as error or variance in the dependent variable that is unaccounted for in the model). In our running example, SOS regression (64.401) is the part that LRyears has predicted out of total SOS (128.500) in the dependent variable (OLA). The SOS residual (64.099), then, is the part of total SOS that LRyears has not been able to predict. If we sum up these two parts, we get the total SOS in OLA (128.500). The partitioning mechanism of ANOVA for OLA (our dependent variable) is summarized in Fig. 19.10.

Given this partitioning, we are able to derive our  $R^2$  of 0.501 (or 50.1%) from the sums of squares listed in the table labeled ANOVA by asking “what proportion of the OLA’s total SOS has been predicted by LR<sub>(years)</sub>.” That is:

$$R^2 = \frac{\text{SOS}_{(\text{Regression})}}{\text{SOS}_{(\text{Total})}} \quad (19.2)$$

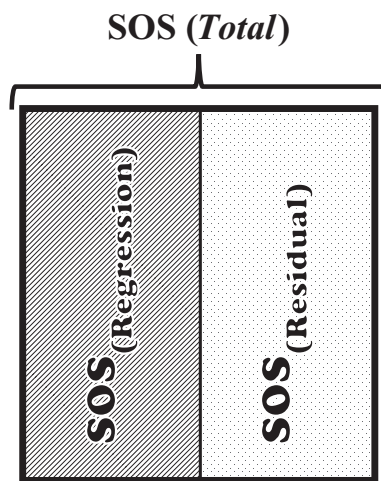


Fig. 19.10 ANOVA partitioning of total sum of squares (SOS) in OLA ( $R^2 = 50.1\%$ )

Thus, our  $R^2$  is:

$$R^2 = \frac{64.401}{128.500} = 0.501 \text{ or } 50.1\%$$

The statistical significance of  $R^2$  may not always be of interest, for a variety of reasons (see, e.g., Norris, 2015; Plonsky, 2015b). Nevertheless, in our example, the  $p$ -value of 0.022 is smaller than 0.05 showing that  $R^2$  of 50.1% is statistically different from 0 given 0.05 as our criterion for detecting statistical significance.

We can then also see that “*Std. Error of the Estimate*” in the *Model Summary* table is simply some function of the amount of the total SOS of OLA that our single predictor, LRyears, could *not* predict (i.e.,  $SOS_{\text{residual}}$ ). The computation of “*Std. Error of the Estimate*” is very simple but it is one that does not concern us here. What is important here, however, is that we want *Std. Error of the Estimate* to be small just like we want our  $SOS_{\text{residual}}$  to be small.

If  $R^2$  (or  $R$ ) is acceptably noteworthy in terms of size (e.g., compared to a theoretical predictor or to previous studies in the same domain), which will likely but not necessarily also correspond to a statistically significant  $p$ -value, we can one move on to the next table in the SPSS output called *Coefficients*. It is at this point that we can then attempt to make predictions for the dependent variable-based values of our independent or predictor variable. However, if we do not have a noteworthy  $R^2$ , several possible conditions may be present which prevent such predictions. For example, we may have chosen the incorrect predictor, our instruments may be unreliable, the pattern between the variables might be non-linear, and/or other design flaws (e.g., insufficient sample size) might be present. In any event, under such circumstances prediction using coefficients in the next section is not possible.

## Regression Coefficients: $a$ , $b$ , and $\beta$

The third table of information, *Coefficients*, enables us to better understand the prediction process. As noted earlier, a computer routine can help us find the coefficients for  $a$  (the intercept) and  $b$  (the slope), which we can then use to fill in our linear forecasting regression model in Eq. (19.1). These coefficients are shown in the SPSS output below (Table 19.9).

Recall that our simple linear regression model had the form:

$$Y_i \leftarrow \hat{Y}_i = a + b(X_i) \quad (1, \text{repeated})$$

**Table 19.9** Output for regression coefficients in SPSS

Model	Coefficients <sup>a</sup>						
	Unstandardized coefficients	Standardized coefficients		95.0% confidence interval for B			
	<i>B</i>	Std. error	Beta	<i>t</i>	Sig.	Lower bound	Upper bound
1 (Constant)	<b>12.239</b>	1.749		6.996	0.000	8.204	16.273
LRyrs	<b>0.479</b>	0.169	0.708	2.835	0.022	0.089	0.868

<sup>a</sup>Dependent variable: OLA

**Table 19.10** Result of prediction of OLA from LRyears for our ten participants

Participant no.	LR <sub><i>i</i></sub>	OLA <sub><i>i</i></sub>	← OĤA <sub><i>i</i></sub>	(OLA <sub><i>i</i></sub> – OĤA <sub><i>i</i></sub> ) = e <sub><i>i</i></sub>
1.	1	11	← 12.717	–1.717
2.	8	14	← 16.069	–2.069
3.	2	15	← 13.196	1.803
4.	9	18	← 16.547	1.452
5.	7	19	← 15.590	3.409
6.	4	13	← 14.153	–1.153
7.	12	13	← 17.984	–4.984
8.	13	22	← 18.463	3.536
9.	16	19	← 19.899	–0.899
10.	17	21	← 20.378	0.621
Mean	8.900	16.500	← 16.500	0.000
SOS	280.900	<b>128.500</b>	← <b>64.401</b>	64.099
Variance	31.211	<b>14.277</b>	← <b>7.155</b>	7.122
SD	5.586	3.778	← 2.675	2.669
Skewness	0.033	0.054	← 0.032	–0.343
Kurtosis	–1.209	–1.483	← –1.208	–0.237

SPSS uses “(Constant)” for *a* (here 12.239), and right below that, SPSS provides *b* (here 0.479). Other quantities in the table of *Coefficients* relate to confidence intervals for *b* or testing whether *a* and *b* (or beta; to be discussed shortly) are statistically different from zero or not. Using *a* and *b*, we can fill in our equation for our example of LR (*X*) predicting OLA (*Y*) as follows:

$$OLA_i \leftarrow O\hat{L}A_i = 12.239 + 0.479 \times (LR_i) \tag{19.3}$$

With these values we can then input values for length of residence in years (LR<sub>*i*</sub>) for a participant in our study and obtain the predicted oral language ability score (OĤA<sub>*i*</sub>) for that participant in the target population. As we had

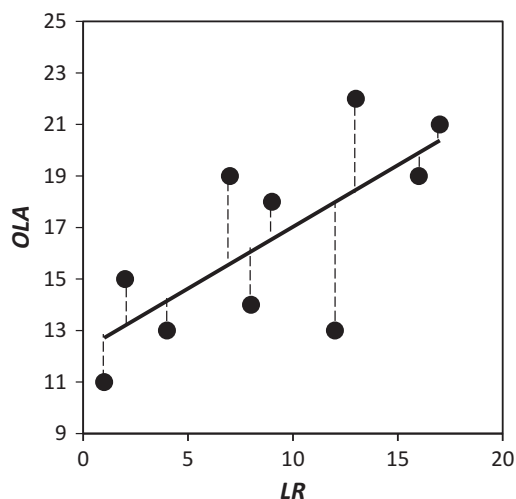


Fig. 19.11 Scatterplot with for  $LR_{(\text{years})}$  predicting OLA with the regression line

Table 19.11 Coefficients table with modified scale for predictor variable

Model	Coefficients <sup>a</sup>						
	Unstandardized coefficients		Standardized coefficients		95.0% confidence interval for $B$		
	$B$	Std. error	Beta	$t$	Sig.	Lower bound	Upper bound
1 (Constant)	12.239	1.749		6.996	0.000	8.204	16.273
LRmths	<b>0.040</b>	0.014	<b>0.708</b>	2.835	0.022	0.007	0.072

<sup>a</sup>Dependent variable: OLA

ten participants in our illustrative example, we will have ten predicted  $\hat{OLA}_i$  scores of our participants as shown in Table 19.10.

In simple terms, our ten predicted scores in the column labeled  $\hat{OLA}_i$  are key to forming a straight line that can best fit our collected data on OLA for ten participants. Figure 19.11 shows how ten predicted scores define a straight line to match the actual research data. As can be seen in the figure, our prediction is not exact. The dashed vertical lines from the data points in the scatterplot to the straight line fitted to the data points indicate the error in the prediction for each participant. These errors for each participant ( $e_i$ ) are also found in Table 19.11 (further below), in the far-right column.

## Additional Notes on $a$ , $b$ , and $\beta$

### Coefficient $a$

In many research situations, including our running example,  $a$ , which we call the *intercept*, is not immediately interpretable. This is because, by default,  $a$  is the predicted score on the dependent variable when the predictor variable is 0. In our example,  $a$  is the predicted score on oral language ability when the length of residence of a person in the target-language community is 0. Basically, this has no meaning to an L2 researcher who has collected data on participants whose length of residence in a target-language community has to be measured in some years. Indeed, participants who have never resided in the target-language community are not even the matter of study, and their predicted scores on OLA are of no empirical use, value, or meaning.

### Coefficient $b$

Unlike  $a$ , the interpretation of the  $b$  coefficient is critical to communicating our predictive results. Generally,  $b$ , which we call *slope*, shows the average amount of linear change in our prediction for the dependent variable ( $\hat{OLA}_i$ ) for one unit of increase in the predictor variable ( $LR_{\text{years}}$ ). This means that interpretations of  $b$  always involve clearly stating the original unit of measurement (e.g., scores, hours, years) for our research variables. A  $b$  with a positive sign indicates an increasing trend in our prediction for the dependent variable for one unit of increase in the predictor variable ( $LR_{\text{years}}$ ). A  $b$  with a negative sign, on the other hand, shows a decreasing trend in our prediction for the dependent variable for one unit of increase in the predictor variable ( $LR_{\text{years}}$ ). Thus in our case, where  $b$  is +0.479, we interpret it as follows:

For one more year of residence ( $X$ ) in the target-language community, our model, on average, predicts +0.479 of a score increase (i.e., improvement) in oral language ability ( $Y$ ).

### Coefficient " $\beta$ "

In the previous section, we noted that the interpretation of  $b$  requires that we always clearly state the original unit of measurement for the dependent (e.g., scores) and the predictor variables (e.g., years). But this implies that  $b$  coefficient is subject to change if we measure our predictors or dependent variable

using a different unit of measurement. For instance, if for our current example, we measured *length of residence* in *months* (i.e., LRmths) rather than in years, our previous  $b$  coefficient would suddenly change from +0.479 to now +0.040 as shown in the SPSS exhibit in Table 19.11.

Simply, our previous  $b$  of +0.479 based on year as the unit of measurement for length of residence is now divided by 12 (i.e.,  $+0.479/12 = 0.040$ ) to adjust for the change of the unit of measurement for length of residence from *year* to now *month*. This dynamic is troubling if one intends to compare a result across a number of studies in the related literature. Here is where  $\beta$  coefficients come to our aid.

Beta ( $\beta$ ) coefficients, listed as “Beta” in the SPSS output above, are like  $b$  coefficients, but they are independent of the unit of measurement (i.e., standardized). For example, our  $\beta$  when we measure length of residence in *years* ( $\beta = +0.708$ ) is the same as when we measured length of residence in *months* ( $\beta = +0.708$ ). This gives  $\beta$  an advantage over  $b$  because it allows for comparison of results across studies, which may have studied the same phenomenon using different units of measurement. In reality,  $\beta$  coefficients use standardized scores on the predictor and the dependent variables (i.e.,  $z$  scores). Standardized scores all are based on a standard deviation (SD) of 1. This allows for the interpretation of  $\beta$  coefficients to be expressed in SD units (again, similar to  $z$  scores or Cohen’s  $d$  values). In terms of sign,  $\beta$  can be negative or positive but has the same interpretation as we saw in the case of  $b$  (see above). Thus, in our case, we can interpret our  $\beta$  coefficient of +0.708 as follows: For 1 SD unit of increase in *length of residence* (no matter measured in years, months, or else), our model, on average, predicts +0.708 SD unit of increase (i.e., here improvement) in the oral language ability (no matter what the unit of measurement).

## Assumptions of Correlation and Regression

As with all statistics, running correlation and regression analyses are only appropriate under predefined assumptions about the data. Nevertheless, many applied linguists employing correlation and regression analyses fail to examine whether such assumptions are met (see Ghanbar & Plonsky, [in press](#)). We will not present here a detailed discussion of the various theoretical assumptions, diagnostic techniques (e.g., outlier detection), or remedial strategies (e.g., data transformation). A complete coverage of these issues is provided elsewhere (Cohen, Cohen, West, & Aiken, 2003, Ch. 4; Draper & Smith, 1998, Ch. 2; Field, 2013, Ch. 8; Graybill & Iyer, 1994, Ch. 3; Howell,

2013, Chs. 9 and 15; Kutner, Nachtsheim, Neter, & Li, 2005, Ch. 3; and Wilcox, 2016, Ch. 8). However, we would remind readers of a few basic assumptions that anyone who reads or conducts research using correlational analyses should be aware of.

First, all statistical approaches, including Pearson's  $r$ , assume that our measurements of study participants' performance are error free. However, this assumption must be checked using one of the appropriate score reliability indices (see Grabowski, Chap. 24; Plonsky & Derrick, 2016; Thompson, 2003) available to us. Put briefly, Pearson's  $r$  based on unreliable scores cannot be validly interpreted. Note that reliability is not guaranteed by simply using a well-known instrument. Any instrument must be checked for the reliability of the scores it produces in each study. Second, recent research shows that the number of participants in a Pearson's  $r$  study needs to be around 30, so that our generalizations to larger pools of participants would be reasonably unbiased (De Winter, Gosling, & Potter, 2016). Third, some outlying scores have been shown to distort our conclusions regarding generalizing correlation size to larger populations of learners (Wilcox, 2016). To allow replicability and verification of correlational results by others, treatment of certain problematic outlying scores using robust correlation methods is required (Wilcox, 2016). In regression, the regression line that we obtain is meant to show the average performance of a great number of participants in the population. Always a confidence interval for the slope ( $b$  or  $\beta$  coefficients) allows evaluating how well our study regression line does at representing that of the population. Prediction is only limited to the many more participants outside our study that fall exactly in the same category in terms of the predictor variable. For instance, in the example developed in this chapter, our predictive results may not be used to predict the mean OLA scores of L2 learners whose length of residence (predictor variable) is much lower or higher than those included in our collected data.

## Conclusion

Approximately four decades ago, Tukey (1977, p. 208) elegantly noted that any research data may be summarized as:

$$\text{Data} = \text{fit} + \text{residual} \quad (19.4)$$

Analyses of association offer direct insights into this view. Understanding what data has to offer requires that we fit a known and reasonable model (e.g.,



linear, quadratic) to it to determine if we can obtain reasonable fit. What, then, cannot be fitted is residual (i.e., error). The fit itself is obtained using a universal principle. We simply weight our theoretically relevant independent variables by applying some weighting coefficients (e.g.,  $a$  and  $b$ , or  $\beta$ ) to them to obtain the fit. That is:

$$\text{fit} = \text{weight} \times \text{independent variable} \quad (19.5)$$

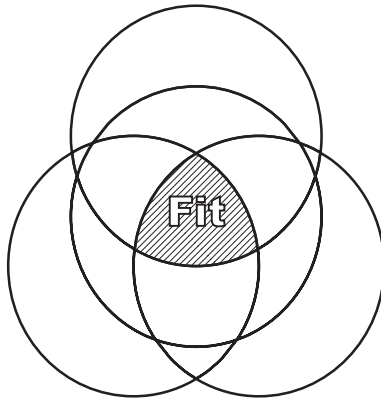
Equation (19.5) is the same as Eq. (19.1), which we used for prediction, and holds across a wide range of analyses of association to obtain the fit. For example, in exploratory factor analysis (EFA), a “factor” or “component” (see Thompson, 2004; see also Phakiti, Chaps. 20 and 21) is obtained by applying weights to the independent variables. The common variance among these variables is simply the factor. Figure 19.12 shows the area view of a factor from four standardized variables.

Under classical test theory, the concept of reliability of test scores (see Thompson, 2003) is conceptualized using:

$$\text{Reliability} = \frac{\text{fit}}{\text{fit} + \text{residual}} \quad (19.6)$$

In ANOVA, we always fit a modified linear regression model (see Plonsky & Oswald, 2017) to our data to obtain a fit (SOSregression). The rest is left unexplained (SOSresidual).

When we encounter a situation in applied linguistics when we believe effects are multiply caused, and causes have multiple effects, our models grow in complexity.



**Fig. 19.12** Factor shown as the commonly shared area among standardized variables

In multivariate approaches (e.g., MANOVA or its mirror descriptive discriminant analysis, or canonical correlation analysis) used to deal with more complex relationships among variables, we always use analogous fitting processes between multiple potential causes and multiple effects in question (see Pituch & Stevens, 2016). The same concepts also apply to more advanced associational approaches, namely, structural equation models (see Kline, 2015; Schoonen, 2015) and multi-level models (see Randenbush & Bryk, 2002).

In truth, fits among a set of L2 variables contain information about what all L2 researchers wish to understand (i.e., the “latent Constructs” that underlie observed variables). In essence, this is the reason why applied linguistics researchers can greatly benefit from understanding analyses of association.

### Sample Study 19.1

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22, 499–533.

#### Background

Proponents of the critical period hypothesis (CPH) hold that human language development becomes constrained beyond puberty. The relationship between age of acquisition (or initial exposure) and ultimate attainment, however, has been disputed in the realms of second language acquisition as well as language policy. DeKeyser sought to shed light on this relationship by extending previous findings in this area.

#### Research Methods

DeKeyser administered a grammaticality judgment tasks (GJT) to 57 Hungarian users of English as a second language. The participants exhibited a wide range of ages at which they had moved to the United States when they had, presumably, begun to be exposed to English. Each participant also took a test designed to measure language learning aptitude. Like other studies in this domain, the main analytical tool for this study was a correlation between the participants' age of arrival and their score on the GJT.

#### Findings

The results of the main analysis revealed a correlation of  $-0.63$ , depicted in the study both numerically and, quite usefully, as a scatterplot. DeKeyser interpreted this result as further evidence of the inverse relationship between age of arrival and ultimate attainment. However, DeKeyser took his analysis one step further. Among other findings, he was also able to show that late learners with high aptitude were able to overcome the limitations imposed by the CPH.

#### Comments

This article presents a clear study with a fairly straightforward design and set of findings to match. It is also worth noting that, whereas some studies interested in age effects divide their participants into arbitrarily formed “early” and “late” groups, DeKeyser chose wisely to preserve the variability in this variable. By doing so, the study was able to examine and shed a great deal of light on the two main variables under investigation: age and morphosyntactic proficiency.

## Resources for Further Reading

Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 131–159). New York, NY: Routledge.

Placed within an edited volume on somewhat more advanced quantitative methods, this chapter provides a hands-on, SPSS-based tutorial the use of multiple regression for L2 research.

Plonsky, L., & Oswald, F. L. (2014). How big is ‘big’? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

This study is primarily concerned with the interpretation of two statistics, one of which is the correlation coefficient described in this chapter. Based on a synthesis of hundreds of studies and dozens of meta-analyses, the authors describe the size of the correlations typically found in the field. Based on their findings, the authors then go on to propose a general set of guidelines for interpreting the strength of correlations in L2 research.

Roever, C., & Phakiti, A. (2017). *Quantitative methods for second language research: A problem-solving approach*. New York, NY: Routledge.

Chapter 5 provides a clear and concise introduction to the conceptual and design-related underpinnings of correlations as applied to second language research. Both parametric and non-parametric applications are described. The authors also provide a very useful guide to the procedures involved in running correlations using SPSS.

## References

- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*, 426–443.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York, NY: Routledge.
- DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.
- De Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, *21*, 273–290.

- Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. New York, NY: Wiley.
- Egbert, J., & Plonsky, L. (2015). Success in the abstract: Exploring linguistic and stylistic predictors of conference abstract ratings. *Corpora*, *10*, 291–313.
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Thousand Oaks, CA: SAGE.
- Graybill, F. A., & Iyer, H. K. (1994). *Regression analysis*. New York, NY: Duxbury Press.
- Howell, D. C. (2013). *Statistical methods for psychology*. Belmont, CA: Cengage Learning.
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99.
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY: Guilford.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. New York, NY: McGraw-Hill.
- Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, *34*, 257–271.
- Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(Supp. 1), 97–126.
- Norris, J. M., Ross, S., & Schoonen, R. (2015). Improving second language quantitative research. *Language Learning*, *65*(Supp. 1), 1–8.
- Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions A*, *373*, 253–318.
- Pituch, K. A., & Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. New York, NY: Routledge.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687.
- Plonsky, L. (Ed.). (2015a). *Advancing quantitative methods in second language research*. New York, NY: Routledge.
- Plonsky, L. (2015b). Statistical power,  $p$  values, descriptive statistics, and effect sizes: A “back-to-basics” approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–45). New York, NY: Routledge.
- Plonsky, L., & Derrick, D. J. (2016). A Meta-Analysis of Reliability Coefficients in Second Language Research. *Modern Language Journal*, *100*, 538–553.
- Plonsky, L., & Ghanbar, H. (in press). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R values. *Modern Language Journal*.
- Plonsky, L., & Oswald, F. L. (2014). How big is ‘big’? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

- Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, 39, 579–592.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: SAGE.
- Roever, C., & Phakiti, A. (2017). *Quantitative methods for second language research: A problem-solving approach*. New York, NY: Routledge.
- Rosnow, R. L., & Rosenthal, R. (2008). *Essentials of behavioral research: Methods and data analysis*. New York, NY: McGraw-Hill.
- Schoonen, R. (2015). Structural equation modeling in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 213–242). New York, NY: Routledge.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
- Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Newbury Park, CA: SAGE.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Wilcox, R. (2016). *Understanding and applying basic statistical methods using R*. New York, NY: Wiley.



# 20

## Exploratory Factor Analysis

Aek Phakiti

### Introduction

Exploratory factor analysis (EFA) is a multivariate statistical technique for identifying the factors that account for the variation in participants' responses to research instruments, such as Likert-type scale questionnaires and tests. This chapter provides an overview of important aspects, considerations and practical guidelines for conducting EFA. An analysis of actual empirical data via IBM SPSS (Statistical Package for Social Sciences) is used to illustrate the EFA methodology.

### Basic Concepts

#### Why Is EFA Used?

There are two key reasons for using EFA in applied linguistics research:

1. To keep the dataset under analysis to a manageable size for subsequent data analysis. Responses to items are interrelated when they are influenced by the same underlying factor. EFA helps researchers form a single score based

---

A. Phakiti (✉)

Sydney School of Education and Social Work, The University of Sydney,  
Sydney, NSW, Australia

e-mail: [aek.phakiti@sydney.edu.au](mailto:aek.phakiti@sydney.edu.au)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_20](https://doi.org/10.1057/978-1-137-59900-1_20)

on various items that represent a construct (e.g., a factor or composite score).

2. To obtain evidence of the *convergent* and *discriminant validity* of the theoretical constructs of interest (see Brown, 2015). Convergent validity refers to the extent to which items that measure the same construct are related (items influenced by the same underlying factor will be strongly related), whereas discriminant validity refers to the extent to which items of the different constructs are related. Outcomes from EFA can be used as part of an instrument development and validation (e.g., Davis, 2016).

Several studies in applied linguistics research have used EFA to identify factor structures (see, e.g., Plonsky & Gonulal, 2015). For example, Asención-Delaney and Collentine (2011) employed EFA to analyse data in a written L2 Spanish corpus. The study identified four distinctive discourse types that had two stylistic variations, narrative and expository. O'Brien (2014) investigated the nature of accentedness, fluency and comprehensibility of native and non-native German speech as assessed by German L2 learners. O'Brien used EFA to identify the links between speech stream characteristics (e.g., phonological, fluency-based and lexical/grammatical) and speech ratings.

## EFA and Confirmatory Factor Analysis (CFA)

EFA and CFA take different approaches to seek underlying factors on observed variables (Brown, 2015). EFA differs from CFA in that no a priori pattern of the relationships among the observed variables factors is assumed. It is, therefore, exploratory in nature, even to the extent that the number of common factors is initially unknown. This number may be determined through a series of EFAs. Some researchers first conduct EFA and then use the results to perform CFA (e.g., Mizumoto & Takeuchi, 2011; Peng & Woodrow, 2010; Phakiti, 2006). However, EFA is not a prerequisite for CFA, even though CFA can be used to validate or confirm EFA results.

## EFA and Principal Component Analysis (PCA)

Unlike CFA, both EFA and PCA are exploratory in nature and can be used for data reduction. However, they make different theoretical assumptions. On the one hand, EFA assumes that a latent variable explains observed variables (EFA is, therefore, an *effect indicators model*; Fabrigar & Wegener, 2012). That is, it is determined by the correlation coefficients (i.e., covariance) among

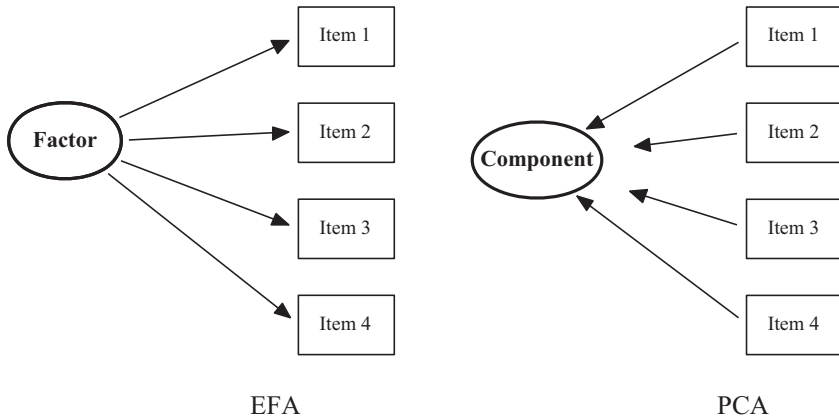


Fig. 20.1 EFA versus PCA

items. On the other hand, PCA utilises the variance of observed variables to derive a component (known as a *causal indicators model*; Fabrigar & Wegener, 2012). While EFA aims at generalising to the target population, PCA only aims at reproducing the sample being used (Osborne, 2014). Figure 20.1 illustrates the differences in these theoretical assumptions.

PCA is often misconceived as a type of EFA, since the results of the analysis from both may yield a similar common factor with comparable factor loadings (Brown, 2015; Fabrigar & Wegener, 2012; Osborne, 2014). Furthermore, a statistical programme, such as SPSS, includes PCA as part of the EFA extraction methods (discussed further below), thereby causing further confusion. PCA is often employed in applied linguistics research (e.g., Cheng, Andrews, & Yu, 2010; Kozaki & Ross, 2011; Murray, Riazi, & Cross, 2012). For example, Kozaki and Ross (2011) used PCA to extract common factors (e.g., aspiration to professional pursuit (APP) and orientation to the social mainstream (OSM)) from a 27-item questionnaire ( $N = 1682$ ). They found a six-PCA solution, employed CFA to confirm the component solution and subsequently used three-level multilevel modelling to examine changes over time (i.e., two years).

### Common Versus Error Factors in EFA

The relationships between latent factors and observed measures are expressed in terms of the regression weights that connect observed indicators to factors. Each observed variable has a linear relationship to one (or more) common factor and one unique factor—a combination of both *systematic* (i.e., non-random)



and *random* errors. For example, if a factor loading on an item is 0.50, 25% of the variance of the item is explained by the underlying latent factor. The total amount of variance in a set of observed items is referred to as *communality variance* or  $h^2$  (the sum of squared factor loadings for each factor; e.g.,  $0.55 = 55\%$ ), and the remaining variance is referred to as *unique variance* (e.g.,  $1 - h^2$ ; e.g., 0.45). Here the unique variance is 45%, which may be due to systematic or specific factors, such as item wording and other random factors. In reality, not all items in EFA will have a strong factor loading to only one factor. Cross-factor loadings (i.e., when a variable has two or more substantive correlation coefficients with other factors) are expected in EFA (e.g., Fabrigar & Wegener, 2012).

## Computer Programmes

Several programmes can perform EFA, and in this chapter SPSS is used to illustrate how EFA can be conducted.

- CEFA (Comprehensive Exploratory Factor Analysis). Visit <http://faculty.psy.ohio-state.edu/browne/software.php>.
- IBM SPSS (Statistical Package for Social Sciences). Visit <https://www.ibm.com/products/spss-statistics>.
- R. Visit <https://www.r-project.org/>.
- SAS (Statistical Analysis Software). Visit <http://www.statisticssolutions.com/statistical-analysis-software-sas/>.

## Critical Requirements and Statistical Assumptions for EFA

The following are the key statistical assumptions and general quantitative requirements for EFA (see Fabrigar & Wegener, 2012; Field, 2013; Osborne, 2014; see also, e.g., Roever & Phakiti, 2018, for a range of quantitative assumptions).

1. *Expected common factors*. This is not a statistical assumption per se. Researchers need to make the assumption that there are underlying common factors that influence the observed items in their instrument, even though they may be unknown at the outset of the study (Fabrigar & Wegener, 2012).

2. *Sample size.* EFA is a large sample size statistical method. There is no consensus on the optimal sample size for EFA. However, some researchers suggest a minimum of five to ten participants per item being used, and a sample size of at least 200 is critical for factor structures to produce correct model solutions and be replicable (Fabrigar & Wegener, 2012; Osborne, 2014).
3. *Number of items.* Researchers need to strive towards a balance between too few items and too many items in a research instrument. On the one hand, a questionnaire that has five items may produce one common factor, even though in principle EFA will produce five factors. It is not worthwhile performing EFA if five individual items measure five different constructs, which may not be related to one another. On the other hand, a questionnaire that has 60 items will result in many common factors, partly due to strong cross-loadings.
4. *Types of data.* Nominal data (e.g., numerical codes assigned to names, such as languages) cannot be used in EFA since they do not have mathematical properties (Fabrigar & Wegener, 2012). If data are binary (e.g., correct/incorrect, coded 1 or 0), they may be used with caution since binary data can introduce a new factor called a *difficulty factor* (Fabrigar & Wegener, 2012, p. 94) that is associated with ability level. A factor extraction method, such as unweighted least squares (ULS), may be used for binary data. Ordinal data (e.g., Likert-type scale) and interval and ratio data that meet the normality assumption are suitable for EFA after screening and cleaning.
5. *Univariate normality.* Ideally data for EFA should be univariately normal because EFA builds on correlational analysis. The key question concerning univariate normality is whether there is sufficient variability in the dataset. Descriptive statistics at an item level should be examined and reported, or made available for readers to examine. The normality of a distribution can be examined using statistics, such as skewness and kurtosis. For binary data, a normal distribution assumption cannot be met, yet it is still possible to perform EFA. When the normality assumption is not met, it is recommended that the *principal axes factor* (PAF) extraction method be adopted in EFA, even though it can also be adopted when the normality assumption is met (illustrated further below).
6. *Linearity.* This statistical assumption is related to the correlations among items. If the items are not statistically correlated, it is unlikely that a common factor will be found. A number of sizable correlation coefficients (e.g., > 0.30) are also needed for EFA to be appropriate. Researchers need

to examine a correlation matrix among items before performing EFA (a correlation matrix can be generated as part of EFA in SPSS).

7. *Outliers based on zero or near-zero  $R^2$  (shared variance)*. It is likely that some items will not be correlated at all with others. For EFA, an item that has a near-zero  $R^2$  with other items is considered an outlier and should not be included in EFA.
8. *Outliers based on perfect or near-perfect  $R^2$* . While there is a need for a number of sizable correlations, researchers do not want to have many perfect or near-perfect correlations in the dataset. With a perfect correlation,  $R^2$  is 1 or is close to 1, which signifies singularity and multicollinearity in the dataset. Such items are also considered outliers, and when found, they should be excluded from subsequent EFA.

## Essential EFA Steps

Although researchers can instruct a programme such as SPSS to perform EFA in a single step (e.g., by specifying extraction and rotation methods at the outset), this chapter emphasises the importance of conducting EFA one step at a time because the early detection of problematic items or choices of extraction methods can prevent difficulties in interpreting common factor structures at a later stage, thereby promoting the validity of EFA. This is a parsimonious approach to EFA. Figure 20.2 provides 12 essential steps in EFA. While this diagram appears sequential, the actual steps are largely iterative.

To illustrate how EFA can be performed using SPSS and following these steps, an unpublished dataset (available upon request for the purpose of practice only) from 275 international students who answered a 37-item, Likert-type scale questionnaire (1 = not at all true of me, 2 = not true of me, 3 = somewhat true of me, 4 = true of me and 5 = very true of me) about their strategy use in academic lectures is used. Note that Item 32 was reverse coded. Figure 20.3 shows a screenshot of part of this dataset.

### Step 1: Entering, Checking and Cleaning Data

This step is common to all quantitative analyses. Unless data are obtained online, researchers first enter data into an SPSS spreadsheet and examine if there are missing data or errors in the data entry. Odd values, for example, have an implication for the variances used in EFA and can result in a different

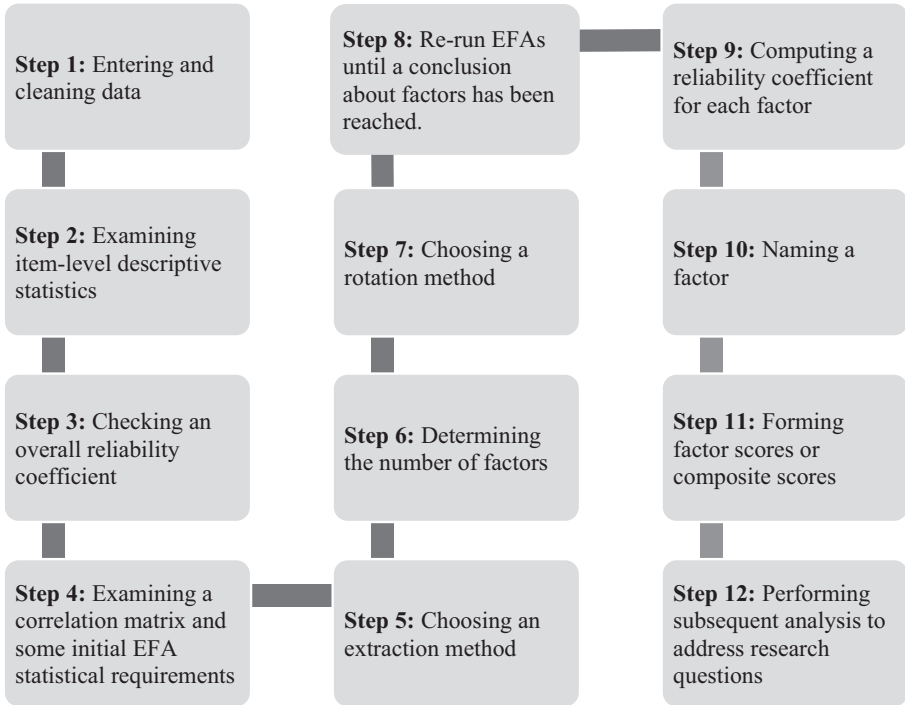
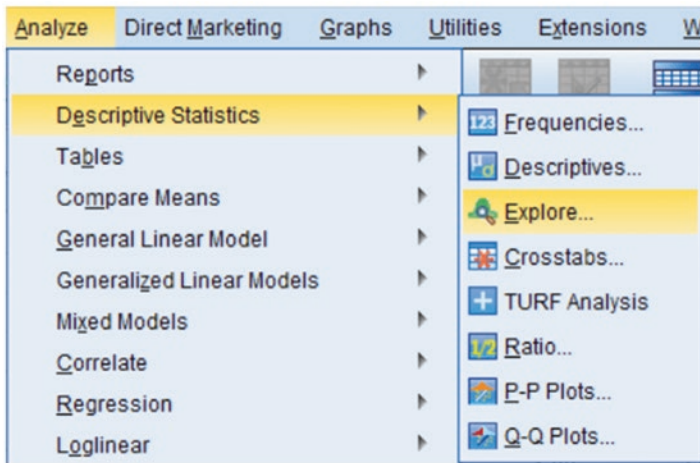


Fig. 20.2 12 essential steps in EFA

	Code	item1	item2	item3	item4	item5
1	1	3.00	3.00	2.00	1.00	2.00
2	2	4.00	3.00	3.00	2.00	4.00
3	3	3.00	3.00	3.00	4.00	4.00
4	4	4.00	3.00	4.00	3.00	3.00
5	5	4.00	3.00	1.00	1.00	1.00
6	6	3.00	3.00	1.00	1.00	1.00
7	7	4.00	3.00	4.00	4.00	4.00
8	8	2.00	2.00	2.00	2.00	2.00
9	9	5.00	4.00	3.00	4.00	4.00

Fig. 20.3 Screenshot of the strategy use in lectures data



**Fig. 20.4** Descriptive statistics options in SPSS

factor structure. Participants with missing values are not used in EFA (see Step 7), which will reduce the sample size. Missing data are common in quantitative research, and there are several methods that have been developed to deal with missing data, such as *mean substitution*, *regression* and *multiple imputation* (see Fig. 20.17). See Brown (2015), Fabrigar and Wegener (2012) and Osborne (2014) for further discussion of how to deal with missing data in EFA (Fig. 20.4).

## Step 2: Examining Item-Level Descriptive Statistics

In research articles, authors may decide not to report item-level descriptive statistics due to the word limits. However, this does not mean that researchers do not need to examine the item-level descriptive statistics, which include the mean, median, mode, standard deviation as well as skewness and kurtosis statistics. Item-level descriptive statistics are essential as they allow researchers and readers to understand the nature of the dataset prior to EFA. Table 20.1 is an example of descriptive statistics for items one to five. The data were normally distributed as the skewness and kurtosis statistics ranged within the limits of  $\pm 1$ .

## Step 3: Checking an Overall Reliability Coefficient

After examining the descriptive statistics, researchers perform reliability analysis. In SPSS, this can be performed in the “Scale” sub-menu under “Analyze.” The Cronbach’s alpha is commonly used for test and Likert-type scale data. If

**Table 20.1** Descriptive statistics of items one to five

		Statistics				
		Item1	Item2	Item3	Item4	Item5
N	Valid	275	275	275	275	275
	Missing	0	0	0	0	0
Mean		3.0036	2.8800	2.5491	2.4800	2.7127
Median		3.0000	3.0000	3.0000	2.0000	3.0000
Mode		3.00	3.00	3.00	2.00	3.00
Std. deviation		0.94559	0.97606	0.89624	0.88507	0.92473
Skewness		-0.242	-0.137	0.219	0.045	-0.040
Std. error of skewness		0.147	0.147	0.147	0.147	0.147
Kurtosis		-0.263	-0.428	-0.110	-0.709	-0.476
Std. error of kurtosis		0.293	0.293	0.293	0.293	0.293
Minimum		1.00	1.00	1.00	1.00	1.00
Maximum		5.00	5.00	5.00	4.00	5.00

**Table 20.2** Cronbach's alpha coefficient of the questionnaire

Reliability statistics	
Cronbach's alpha	N of items
0.865	37

the overall Cronbach's alpha is lower than 0.70, researchers need to decide whether to continue with an EFA approach. In SPSS, it is useful to choose the option "Scale if item deleted" in reliability analysis, so that some items that negatively affect the reliability estimates may be identified and excluded from EFA. Table 20.2 presents the overall Cronbach's alpha of the questionnaire, which was 0.87.

#### Step 4: Examining a Correlation Matrix and Some Initial EFA Statistical Requirements

This step is related to points 2, 8, 9 and 10 in the assumptions discussed above. This step can be performed from the factor analysis menu in SPSS. To start performing EFA, go to "Analyze," "Dimension Reduction" and then "Factor" (see Fig. 20.5).

Figure 20.6 is the main dialog box for EFA. Here, move items 1 to 37 to the Variables pane on the right.

In Fig. 20.6, there are sub-menu options that allow researchers to perform various stages of EFA. The Descriptives option is relevant to this step. Click

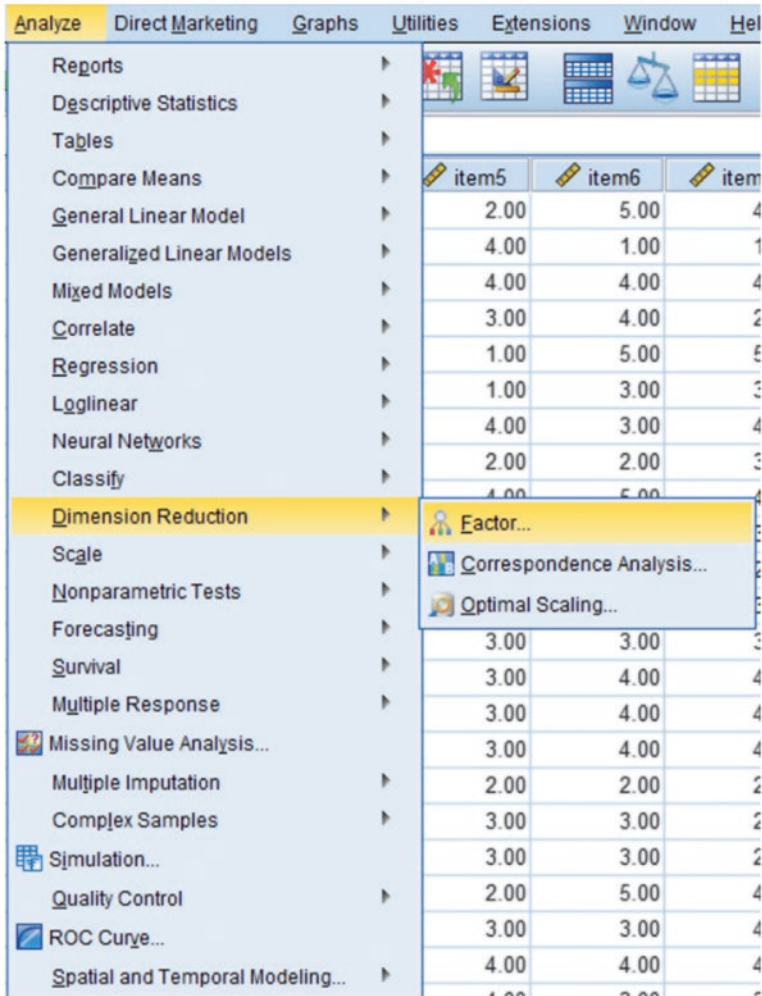


Fig. 20.5 EFA in SPSS

“Descriptives” and a new dialog box will appear (see Fig. 20.7). In this option, tick the following boxes: initial solution, coefficients and Kaiser-Meyer-Olkin (KMO) and Bartlett’s test of sphericity. Then click “Continue” and “OK.” More options can be chosen (see e.g., Field, 2013), but for this chapter, this selection is sufficient.

The descriptive statistics option is not selected because it only reports the means, standard deviations and the number of cases (which is not as comprehensive as in Step 2). In this step, we first check whether the correlation coefficients in the matrix are reasonable and second check whether there are perfect



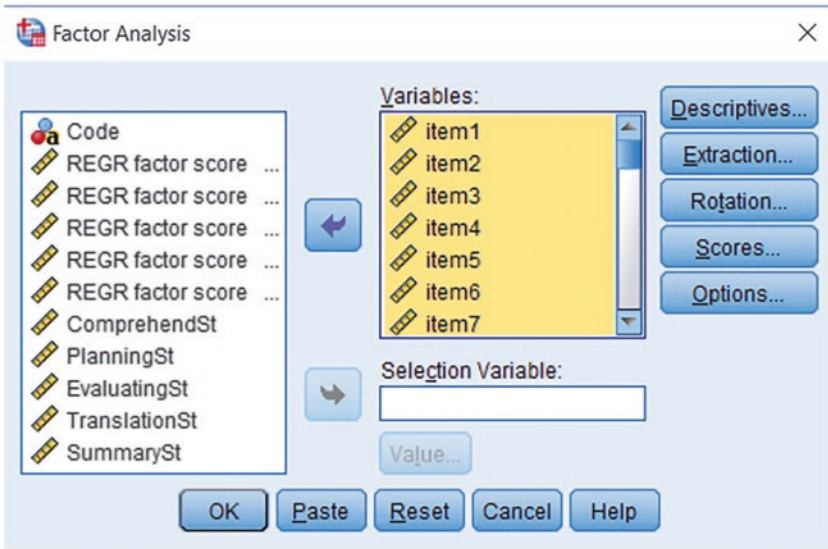


Fig. 20.6 Factor analysis menu

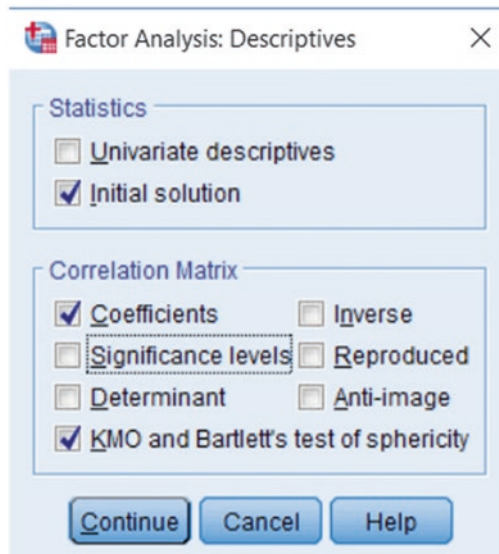


Fig. 20.7 SPSS Descriptives dialog box

correlations or several zero or near-zero correlations (refer to points 9 and 10 in the assumptions section). Due to limitations of space, this matrix is not presented here. A careful inspection suggests that several correlations were larger than 0.30, and there were no perfect correlations or near-zero correlations of a



**Table 20.3** KMO and Bartlett's test based on 37 items

KMO and Bartlett's test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.837
Bartlett's test of sphericity	Approx. chi-square	3790.175
	<i>df</i>	666.000
	Sig.	0.000

particular item to any other items, indicating that factor analysis is plausible. Table 20.3 presents the results of the KMO and Bartlett's test.

The Kaiser-Meyer-Olkin measure is used to indicate data sampling adequacy and this statistic should be larger than 0.60 (Fabrigar & Wegener, 2012; Field, 2013). In this dataset, it was 0.84. The Bartlett's test of sphericity is used to determine whether the dataset is factorable, and this statistic should be significant at 0.05 (Fabrigar & Wegener, 2012; Osborne, 2014), and in this output, it should be read as  $p < 0.001$ .

## Step 5: Choosing an Extraction Method

To choose an extraction method, follow the steps in Fig. 20.5. There is no need to reset the analysis; the extraction option should be continued from the previous step (i.e., leave the Descriptives option in Step 4 unchanged). To specify an extraction method, click "Extraction." Figure 20.8 presents the extraction dialog box. In this dialog box, there are seven extraction methods: principal component (or PCA), unweighted least squares (ULS), generalised least squares (GLS), maximum likelihood (ML), principal axes factor (PAF), alpha extraction (AE) and image factor (IF) (see Fabrigar & Wegener, 2012; Osborne, 2014).

As noted above, if PCA is chosen, the entire analysis will be treated as PCA, rather than EFA. ULS should be chosen when the data are non-normal (especially for binary scores), while GLS relies on variables with strong correlations and the outcome can be sensitive to those with weak correlations. ML is a robust method that aims to produce the population correlation matrix and the data need to be univariately and multivariately normal. ML in EFA also provides a goodness-of-fit statistic of the extracted factors (namely the chi-square test of goodness-of-fit test, which can be used as the basis for hand calculations of other fit indices such as the root-mean-square error of approximation (RMSEA); Fabrigar & Wegener, 2012; see Phakiti, Chap. 21). However, EFA researchers are not normally concerned with fit statistics. PAF functions similarly to ML, but should be used when the multivariate normality condition is not met. AE focuses on the reliability of the extracted factors (i.e., generalising

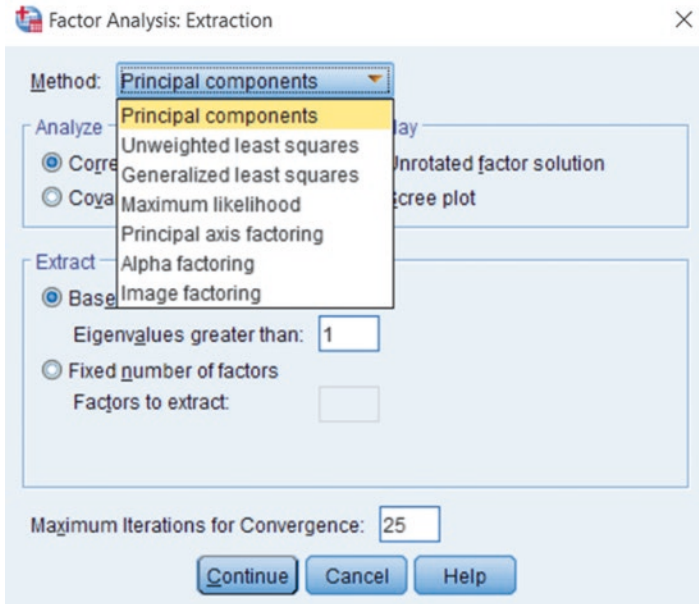


Fig. 20.8 SPSS extraction dialog box

a population of measures instead of a population of individuals (Osborne, 2014)). This extraction technique can cause confusion in interpreting the results, even after factor rotation. IF uses the image (i.e., communality) covariance matrix for factor extraction (similar to that used in PCA). In brief, for EFA, PAF is a reasonably robust method for factor extraction and is commonly chosen (Fabrigar & Wegener, 2012; Field, 2013; Osborne, 2014).

The decision to use a particular extraction method depends on the research purpose and the types of data gathered. Some researchers may begin their EFA by starting with the PCA option as a strategy to identify the number of factors to be extracted, followed by EFA (as illustrated in this chapter). This option is viable given the entire concept of EFA is exploratory and parallel analysis appears to be more conservative in producing eigenvalues for PCA than for EFA (eigenvalues based on random data for EFA are normally smaller than those for PCA). In several comparative analyses, EFA and PCA have been found to produce similar outcomes (as in the illustrated case here).

In Fig. 20.9, tick the following boxes: “covariance matrix,” “unrotated factor solution,” “scree plot” and “based on eigenvalue larger than 1.” The maximum iterations for convergence is by default 25. Then click “Continue” and “OK.” Note that this Extraction menu will be revisited several times during the extraction process. Table 20.4 presents the communalities, both initial and

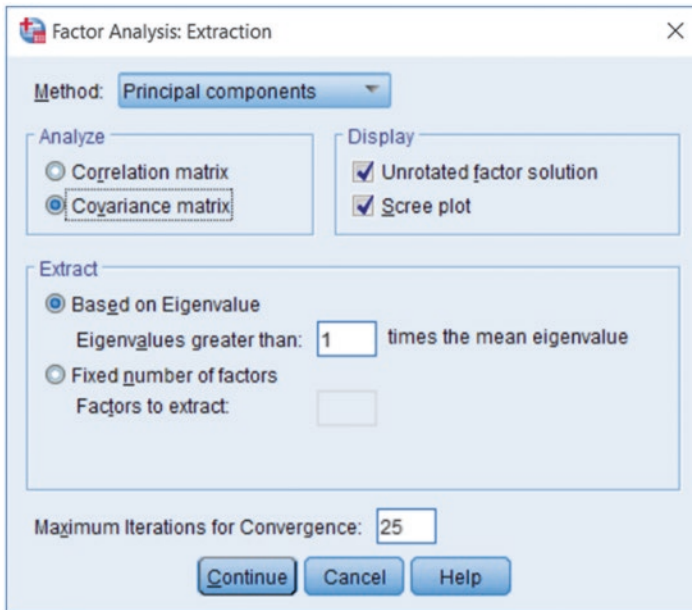


Fig. 20.9 SPSS extraction dialog box

extracted. In PCA, the initial (rescaled) communalities are normally 1 for all items (Osborne, 2014). Generally the extracted value of each item should be at least 0.3; items lower than 0.3 may be excluded from further extraction. Table 20.4 suggests that all extracted values were greater than 0.3.

## Step 6: Determining the Number of Factors to Retain

There are several statistical methods that help researchers decide on the number of factors to be extracted, for example, the *Kaiser criterion*, the *scree plot* and *parallel analysis*. Apart from these statistical criteria, researchers should use the theories underlying their study (if any) to inform the number of factors to be kept (Fabrigar & Wegener, 2012). Table 20.5 presents an output on factor extractions. Only factors with eigenvalues greater than 1 are included in this table (using the Kaiser criterion).

The Kaiser criterion for an eigenvalue greater than 1 is by default set in SPSS. An eigenvalue is the sum of the squared factor loadings, and factors with low loadings will have an eigenvalue less than 1. An eigenvalue less than 1 suggests that the variance explained by the factor is less than the variance of the item (Brown, 2015). According to Table 20.5, there are eight factors with eigenvalues greater than 1. While often used as a single determinant of the

**Table 20.4** Communalities (initial and extracted)

	Communalities			
	Raw		Rescaled	
	Initial	Extraction	Initial	Extraction
Item1	0.894	0.635	1.000	0.710
Item2	0.953	0.648	1.000	0.680
Item3	0.803	0.529	1.000	0.659
Item4	0.783	0.572	1.000	0.730
Item5	0.855	0.567	1.000	0.663
Item6	0.914	0.625	1.000	0.684
Item7	1.069	0.854	1.000	0.799
Item8	1.276	1.020	1.000	0.799
Item9	0.964	0.536	1.000	0.555
Item10	1.071	0.771	1.000	0.720
Item11	1.171	0.908	1.000	0.775
Item12	0.772	0.383	1.000	0.496
Item13	0.728	0.417	1.000	0.573
Item14	0.623	0.369	1.000	0.591
Item15	0.867	0.588	1.000	0.678
Item16	1.040	0.698	1.000	0.671
Item17	0.824	0.529	1.000	0.642
Item18	0.891	0.549	1.000	0.616
Item19	0.999	0.609	1.000	0.610
Item20	1.227	0.925	1.000	0.754
Item21	1.064	0.587	1.000	0.552
Item22	1.650	1.502	1.000	0.910
Item23	0.739	0.292	1.000	0.396
Item24	0.739	0.368	1.000	0.498
Item25	0.695	0.371	1.000	0.533
Item26	0.962	0.561	1.000	0.584
Item27	0.813	0.381	1.000	0.469
Item28	0.807	0.354	1.000	0.439
Item29	0.605	0.272	1.000	0.449
Item30	0.813	0.498	1.000	0.613
Item31	0.976	0.601	1.000	0.616
Item32	0.986	0.719	1.000	0.729
Item33	0.746	0.372	1.000	0.498
Item34	0.955	0.537	1.000	0.562
Item35	0.825	0.385	1.000	0.467
Item36	0.861	0.529	1.000	0.614
Item37	1.063	0.708	1.000	0.666

Extraction method: principal component analysis

factors to be extracted, this statistical index of eigenvalues greater than 1 is limited in terms of its accuracy (see, e.g., Ledesma & Valero-Mora, 2007) and should therefore be used in consultation with other methods, such as the scree plot and parallel analysis (Fabrigar & Wegener, 2012; Field, 2013; Plonsky & Gonulal, 2015). Figure 20.10 presents the scree plot (based on PCA), which also suggests eight common factors.

Table 20.5 Total variance explained

Component	Total variance explained					
	Initial eigenvalues <sup>a</sup>			Extraction sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	6.494	19.086	19.086	6.494	19.086	19.086
2	3.636	10.688	29.774	3.636	10.688	29.774
3	2.621	7.704	37.477	2.621	7.704	37.477
4	1.949	5.728	43.205	1.949	5.728	43.205
5	1.569	4.612	47.817	1.569	4.612	47.817
6	1.343	3.947	51.764	1.343	3.947	51.764
7	1.123	3.299	55.064	1.123	3.299	55.064
8	1.102	3.239	58.303	1.102	3.239	58.303

Extraction method: principal component analysis

<sup>a</sup>When analysing a covariance matrix, the initial eigenvalues are the same across the raw and rescaled solution

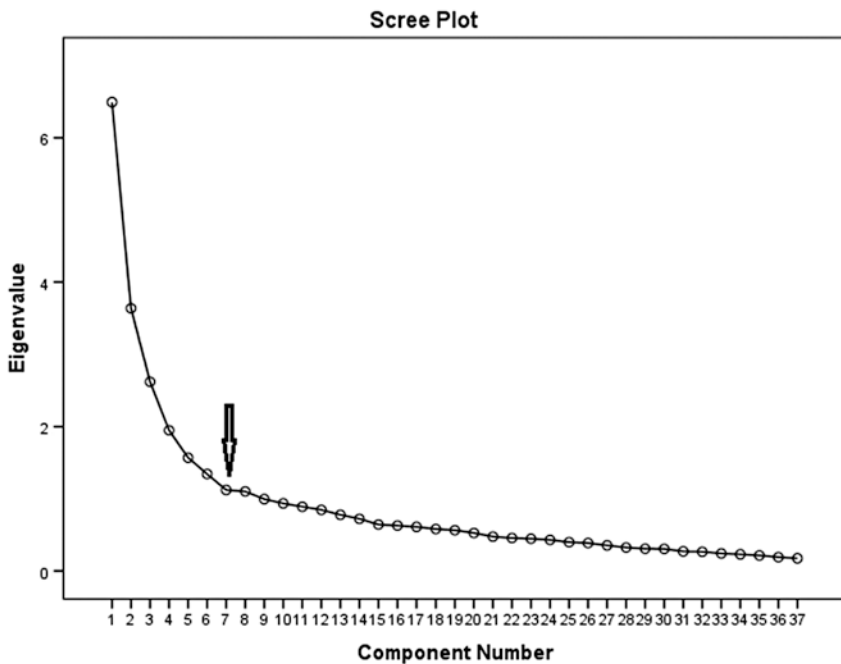


Fig. 20.10 Scree plot (PCA)

The scree plot allows researchers to determine the point at which the curve of the eigenvalues changes or becomes flat (known as the elbow; as indicated by the arrow). Some scree plots do not always have a clear elbow. Like the Kaiser criterion, the factors to retain from the scree plot may still be

inaccurate and the appropriate factors may be one or two above the elbow (Osborne, 2014). Based on the scree plot, six factors may be retained, given that the curve to the right of factor 7 is quite flat.

*Parallel analysis* (Horn, 1965) has gained a recognition for its usefulness in determining the number of common factors to be extracted in EFA (see Fabrigar & Wegener, 2012, for several options for parallel analysis). Parallel analysis is a Monte Carlo simulation method that calculates the number of expected eigenvalues from random data (Fabrigar & Wegener, 2012; Ledesma & Valero-Mora, 2007). SPSS does not conduct parallel analysis during the extraction stage. However, O'Connor (2000) has written an SPSS syntax that can be used in SPSS (this is obtainable from <https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html>; choose parallel.sps, or [http://global.oup.com/us/companion.websites/9780199734177/supplementary/factors/factors\\_a/](http://global.oup.com/us/companion.websites/9780199734177/supplementary/factors/factors_a/); choose “SPSS\_Parallel\_Analysis\_Syntax.sps”). Note that this internet-based parallel analysis (<http://analytics.gonzaga.edu/parallelengine/>) provides similar results to those illustrated below.

It is quite simple to instruct SPSS to use this syntax. First, in SPSS, click “File,” “New” and then “Syntax” (see Fig. 20.11). Then copy and paste all the text from the “SPSS\_Parallel\_Analysis\_Syntax.sps” here. Alternatively, open the saved syntax to edit to fit the properties of the data.

Figure 20.12 is an extract of the parallel analysis syntax. In this figure, the number of cases was changed to 275, and the number of variables was changed to 37. Select “1” to choose principal component analysis and “2” for factor analysis. In this syntax, 1 was chosen. Finally, click “Run” and choose “All.”

SPSS then produces the outcomes of the parallel analysis to be used to decide the number of factors. For the purpose of this chapter, Table 20.6 was constructed to compare the eigenvalues from the sample data with those from the parallel analysis. The 95th percentile eigenvalues were chosen instead of the mean eigenvalues (i.e., the 50th percentile).

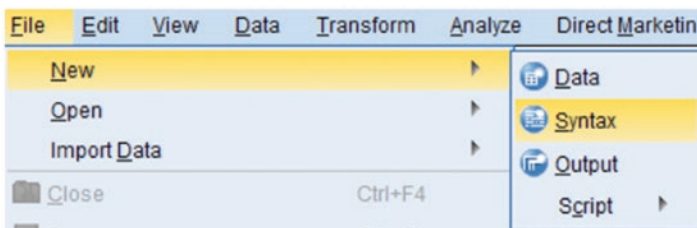


Fig. 20.11 Creating a parallel analysis syntax

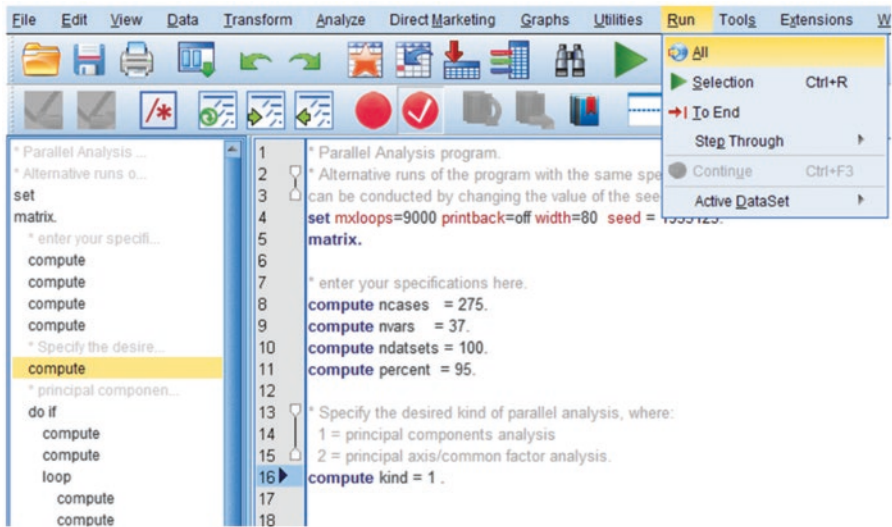


Fig. 20.12 Customising a parallel analysis syntax in SPSS

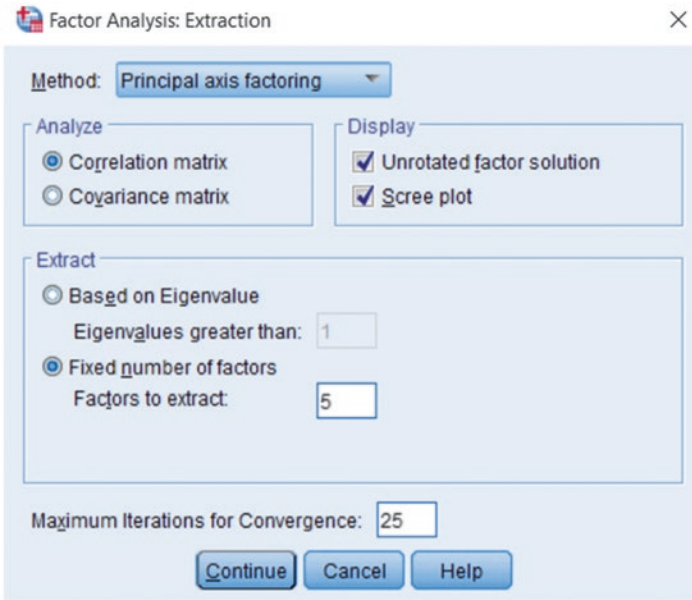
Table 20.6 A comparison of eigenvalues from the dataset with those from the parallel analysis

Factors	PCA eigenvalues	95th percentile eigenvalues
1	6.494	1.848463
2	3.636	1.725992
3	2.621	1.666542
4	1.949	1.588613
5	1.569	1.540972
6	1.343	1.481846
7	1.123	1.442494
8	1.102	1.396580

Here, we compare the eigenvalues from the real data (as computed to use the Kaiser criterion) and those from the random data. When the eigenvalue in the real data is lower than that produced by the parallel analysis, we can determine the number of factors to be further extracted. Based on Table 20.6, the eigenvalues of factors 1, 2, 3, 4 and 5 were lower than those of the parallel analysis. Therefore, the parallel analysis suggests that five common factors might be extracted from this dataset.

EFA can next be performed using this information to assist in the decision-making. Figure 20.13 presents the SPSS dialog box for EFA with the PAF extraction method. In this dialog box, tick “correlation matrix” and “fixed number of factors” (specify to 5). Then click “Continue” and “OK.”

The output from SPSS allows researchers to investigate whether there are items that have extraction values less than 0.3, and which can then be



**Fig. 20.13** Extracting factors using the principal axes factoring method with the fixed factor number = 5

**Table 20.7** KMO and Bartlett's test based on 25 items

KMO and Bartlett's test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.824
Bartlett's test of sphericity	Approx. chi-square	2593.969
	df	300
	Sig.	0.000

subsequently excluded. Researchers can rerun this analysis until all the items to be used have extraction values larger than (or just above) 0.3. In this illustration, a total of 12 items were excluded from further EFA (based on a series of three runs). Table 20.7 presents the results of the KMO and Bartlett's test, which suggest that factors are extractable in this dataset.

Table 20.8 presents the initial and extraction values based on the PAF method. Unlike PCA, the initial communalities in EFA are based on the variance for each item accounted for by all factors (Osborne, 2014).

Table 20.9 presents the total variance explained based on 25 items with five common factors being extracted (56.7% of the total variance explained).

Figure 20.14 presents the scree plot based on the PAF method, which suggests five common factors.



**Table 20.8** Initial and extraction values based on the PAF method

	Communalities	
	Initial	Extraction
Item1	0.499	0.577
Item2	0.462	0.492
Item3	0.557	0.569
Item4	0.627	0.692
Item5	0.561	0.535
Item6	0.420	0.399
Item8	0.448	0.567
Item10	0.417	0.454
Item11	0.437	0.650
Item12	0.499	0.411
Item13	0.503	0.368
Item14	0.541	0.577
Item15	0.561	0.578
Item16	0.391	0.315
Item17	0.537	0.588
Item18	0.400	0.340
Item20	0.384	0.395
Item29	0.367	0.364
Item30	0.456	0.391
Item31	0.379	0.400
Item32	0.336	0.312
Item35	0.378	0.323
Item36	0.435	0.436
Item37	0.467	0.510
Item25	0.397	0.404

Extraction method: principal axis factoring

**Table 20.9** Total variance explained (five-factor extraction)

Factor	Total variance explained					
	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	6.087	24.347	24.347	5.552	22.209	22.209
2	3.043	12.171	36.518	2.583	10.334	32.543
3	1.996	7.986	44.503	1.455	5.820	38.363
4	1.783	7.131	51.634	1.248	4.993	43.356
5	1.264	5.054	56.689	0.809	3.236	46.592

Extraction method: principal axis factoring

## Step 7: Choosing a Rotation Method

Rotation of factors does not change the amount of variance explained by the factors, but it helps redistribute the variance within a factor, so that the factor is more transparent and interpretable (Fabrigar & Wegener, 2012). There are

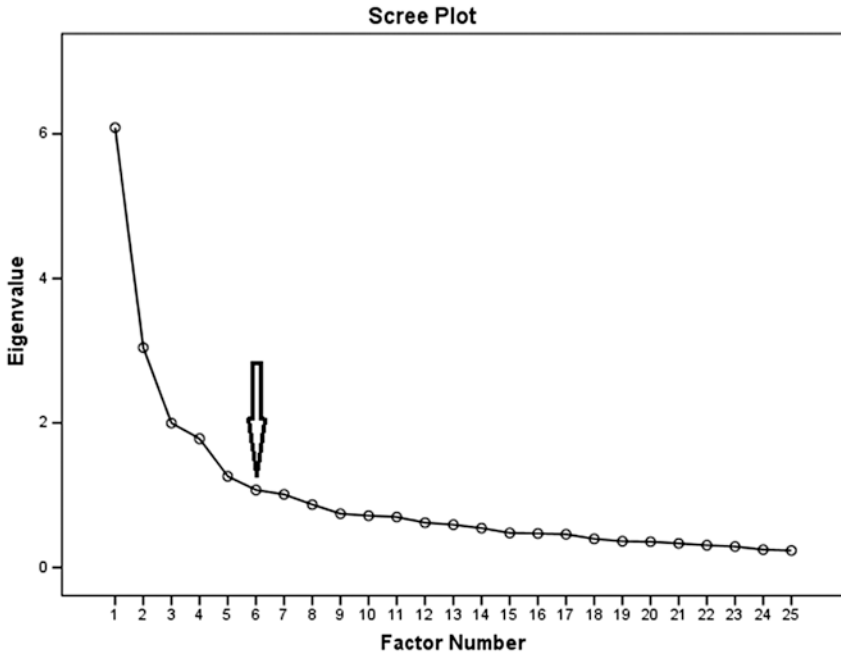


Fig. 20.14 Scree plot (PAF)

two types of rotation methods in EFA: *orthogonal rotation* and *oblique rotation* (see Fabrigar & Wegener, 2012). Orthogonal rotation (e.g., *Equimax*, *Varimax* and *Quartimax*) is chosen when the factors are assumed or expected to be unrelated, whereas oblique rotation (e.g., *Oblimin* and *Promax*) is chosen when the factors are assumed to be correlated. It is important to stress that whether factors are assumed to be related or not should be informed by theory and previous research.

It is recommended to begin by using an oblique rotation, so that we can check whether there is a sizable correlation (noted further below), and if not, an orthogonal method should be chosen for subsequent analysis. There are five rotation methods in SPSS: Varimax, Quartimax, Equamax, Direct Oblimin and Promax. “Varimax” is quite popular for the orthogonal method (although one should also investigate whether Equimax produces better rotation, because it was designed as a compromise between Varimax and Quartimax (Osborne, 2014)), and “Direct Oblimin” is commonly preferred for the oblique method.

To continue to rotate factors, follow the procedure in Fig. 20.5. Do not click “Reset” as the SPSS instruction for the descriptives and extraction method specified earlier will also be reset. Click “Rotation.” Figure 20.15 presents the

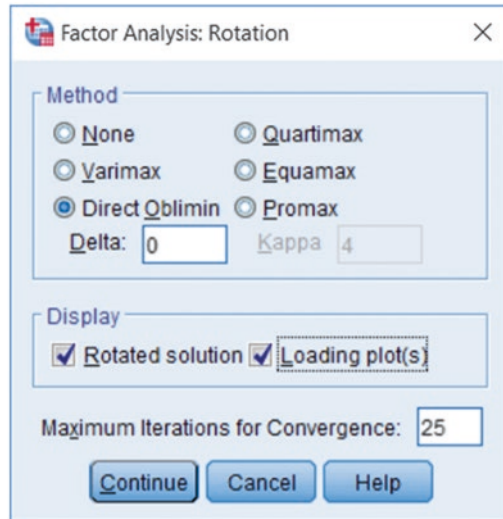


Fig. 20.15 Rotation dialog box (direct Oblimin method)

Table 20.10 Factor correlation matrix

Factor correlation matrix					
Factor	1	2	3	4	5
1	1.000	0.153	-0.202	-0.249	0.347
2		1.000	-0.048	-0.277	0.256
3			1.000	0.087	-0.236
4				1.000	-0.194
5					1.000

Extraction method: principal axis factoring.

Rotation method: Oblimin with Kaiser normalisation

rotation dialog box. In this dialog box, tick “Direct Oblimin,” “Rotated solution” and “Loading plot(s).”

At the end of the SPSS output, there is a factor correlation matrix table (Table 20.10) that allows researchers to determine whether an oblique method is suitable for the dataset. Table 20.10 suggests that all the five factors were not strongly correlated. A correlation coefficient of 0.5 or higher suggests that an oblique method is suitable for factor rotation (e.g., Osborne, 2014). On the basis of this finding, the Varimax rotation method will be chosen.

Figure 20.16 is the SPSS dialog box for the Varimax rotation. Click “Continue.”

Next click “Options.” Figure 20.17 presents the Options dialog box. This option is useful as it allows factor loadings (coefficients) to be sorted by size

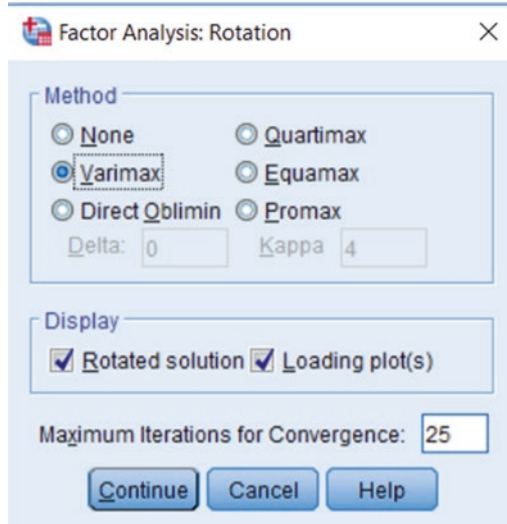


Fig. 20.16 Rotation dialog box (Varimax method)

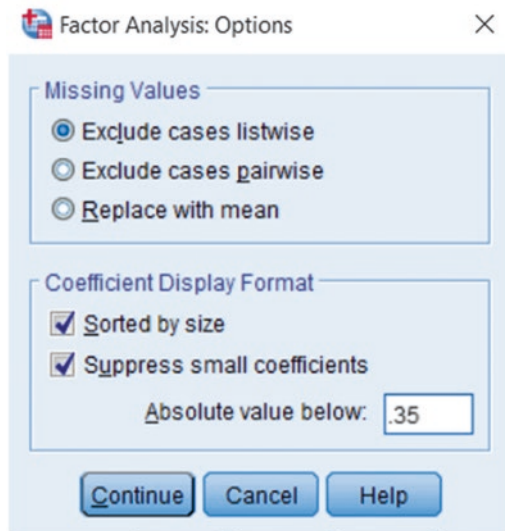


Fig. 20.17 Options dialog box

and allows the suppression of small factor loadings. In EFA, the decision to set the cutoff point for a factor loading to be included in a rotated factor matrix is largely arbitrary (see Osborne, 2014). However, generally a cutoff point is 0.30 (e.g., Fabrigar & Wegener, 2012; Field, 2013). Some authors may suggest 0.40 or 0.50 as the cutoff standard (Osborne, 2014). Ideally, a minimum

factor loading of 0.50 is preferred, because then at least 25% of the shared variance is obtained. However, since in EFA, all items have cross-factor loadings, it is not realistic to set a high cutoff value (e.g., 0.80) since then too many items will be lost. Researchers should explicitly articulate how they set the cutoff value for factor loadings, outlining theoretical, methodological and statistical considerations.

For the purpose of this chapter, 0.35 is chosen as the cutoff, as this value accounts just above 10% of a factor variance. Suppressing coefficients also allows us to investigate whether an item has large cross-factor loadings, which would prompt us to consider dropping an item from a subsequent EFA. In Fig. 20.17, tick “Sorted by size” and “Suppress small coefficient,” and in “Absolute small value,” type in “0.35.” It should be noted that in the Options, “missing value” is by default set to “Exclude cases listwise”—meaning that participants with missing data are excluded from the analysis. There is also an option for “Exclude pairwise” and “Replace with mean.”

Table 20.11 presents a rotated factor matrix that allows researchers to examine items clustering within a common factor.

Factor loadings in EFA are standardised correlation coefficients, which range from 0 to 1 (the total variances equal 1). In the case of oblique solutions, factor loadings are regarded as partial regression coefficients (Fabrigar & Wegener, 2012; Osborne, 2014).

## **Step 8: Rerun EFAs Until a Conclusion About the Factors Has Been Reached**

In Table 20.11, all the items, except item 25, clearly load on one factor. On this basis, we can try to exclude item 25 and rerun the EFA and then decide whether to retain or exclude it. Table 20.12 shows the rotated factor matrix obtained with item 25 excluded; it suggests a more transparent factor structure has been achieved. On the basis of this, a five-factor EFA model seems viable, subject to further reliability analysis of each factor.

## **Step 9: Computing a Reliability Coefficient for Each Factor**

Frequently, EFA may suggest a clear set of factors, but when items clustered within a particular factor are analysed for internal consistency (scale reliability), a Cronbach’s alpha may be found to be below 0.70. Table 20.13 summarises the Cronbach’s alpha coefficients for each factor. With the exception of factor 4, the Cronbach’s alpha coefficients are at least 0.70. For factor 4, the

**Table 20.11** Rotated factor matrix based on the Varimax method (based on 25 items)

	Rotated factor matrix <sup>a</sup>				
	Factor				
	1	2	3	4	5
Item17	0.758				
Item15	0.733				
Item14	0.715				
Item12	0.599				
Item13	0.575				
Item18	0.566				
Item16	0.553				
Item6	0.521				
Item25	0.380			-0.355	
Item30	0.379				
Item4		0.757			
Item1		0.714			
Item3		0.710			
Item5		0.650			
Item2		0.646			
Item37			0.656		
Item36			0.625		
Item31			0.616		
Item35			0.427		
Item29			0.396		
Item8				0.743	
Item20				0.595	
Item32				0.469	
Item11					0.750
Item10					0.606

Extraction method: principal axis factoring

Rotation method: Varimax with Kaiser normalisation

<sup>a</sup>Rotation converged in five iterations

“scale-if-item deleted” result suggests that when item 32 is removed, the coefficient would increase from 0.62 to 0.72.

A rerun of the EFA excluding item 32 suggests a similar factor pattern (see Table 20.14). It should be noted that only two items were explained by factor 4 or 5. Brown (2015, p. 34) suggests eliminating a factor that has only “two or three items.” That is, at least four to five items should form part of a factor. The current findings of having two factors on which only two items are included may be noted as a limitation and a caution in subsequent analysis (see Brown, 2015, p. 34). It may be worthwhile to rerun the EFA and specify the number of factors to be retained to three. However, this rerun will reduce the number of items to be used to examine the research constructs as well as to answer the research questions. In this chapter, given the high factor loading of factors 5 and 6, it was decided to accept the five-factor model.

**Table 20.12** Rotated factor matrix based on the Varimax method (based on 24 items)

Rotated factor matrix <sup>a</sup>					
	Factor				
	1	2	3	4	5
Item17	0.762				
Item15	0.735				
Item14	0.714				
Item12	0.600				
Item13	0.576				
Item18	0.563				
Item16	0.553				
Item6	0.522				
Item30	0.381				
Item4		0.757			
Item1		0.714			
Item3		0.710			
Item5		0.650			
Item2		0.646			
Item37			0.653		
Item36			0.628		
Item31			0.626		
Item35			0.430		
Item29			0.395		
Item8				0.803	
Item20				0.618	
Item32				0.416	
Item11					0.757
Item10					0.600

Extraction method: principal axis factoring

Rotation method: Varimax with Kaiser normalisation

<sup>a</sup>Rotation converged in five iterations

**Table 20.13** Cronbach's alpha coefficient of each factor (based on 24 items)

	Items	No of items	Cronbach's alpha	Cronbach's alpha (excluding Item 32)
Factor 1	17, 15, 14, 12, 13, 18, 16, 6, 30	9	0.85	
Factor 2	4, 1, 3, 5, 2	5	0.83	
Factor 3	37, 36, 32, 35, 29	5	0.73	
Factor 4	8, 20, 32	3	0.62	0.72
Factor 5	11, 10	2	0.70	
<b>Overall</b>		<b>24</b>	<b>0.81</b>	<b>0.83</b>

**Table 20.14** Rotated factor matrix based on the Varimax method (based on 23 items; final model)

Rotated factor matrix <sup>a</sup>					
	Factor				
	1	2	3	4	5
Item17	0.762				
Item15	0.742				
Item14	0.718				
Item12	0.606				
Item13	0.571				
Item18	0.550				
Item16	0.545				
Item6	0.532				
Item30	0.401				
Item4		0.751			
Item1		0.717			
Item3		0.707			
Item2		0.651			
Item5		0.647			
Item37			0.649		
Item36			0.644		
Item31			0.616		
Item35			0.431		
Item29			0.386		
Item8				0.863	
Item20				0.632	
Item11					0.815
Item10					0.569

Extraction method: principal axis factoring  
 Rotation method: Varimax with Kaiser normalisation  
<sup>a</sup>Rotation converged in six iterations

### Step 10: Naming a Factor

SPSS does not name or label factors automatically, so we need to examine the content of a cluster of items and decide on a name for each factor. Many books on factor analysis do not provide guidance on how to label a factor. A factor name should be indicative of the conceptual dimensions of an underlying construct. The process of naming a factor is similar to that used in content or thematic analysis in qualitative research. It is important to consider relevant theories and prior research as a factor name should be connected to a theory of interest and be similar to those used by other researchers, so that similarities or differences can be established. If an existing instrument is adopted, researchers should adopt factor names similar to those used previously. This consideration is important for a future meta-analysis. In the case of novel research,



EFA can help researchers theorise empirical evidence, thereby introducing a new significant facet to a well-established construct.

This naming step can be somewhat subjective as it requires researchers to interpret a common theme emerging from a group of items, and different researchers may give different names to the same factor. Table 20.15 presents factors, item statements and assigned factor names.

**Table 20.15** Rotated factor matrix based on the Varimax method (based on 23 items; final model)

		Factor	
	Item	Content	Factor name
Factor 1	Item17	I use background knowledge and previous information to help me understand lectures.	Comprehending strategies
	Item15	I use my prior experience and knowledge to help me understand lecture content.	
	Item14	I try to make sense of what I hear with what I have read or studied.	
	Item12	I link my prior personal experience with lecture content.	
	Item13	I relate what I hear to what I have understood earlier.	
	Item18	I try to simplify information for me to remember.	
	Item16	I use knowledge of words' prefixes and suffixes (e.g., un-, in-, -ness, -ion) to guess the meaning of an unknown word.	
	Item6	I guess the meaning of new vocabulary using familiar words.	
	Item30	When I find my mind is wandering, I try to regain focus on the lecture.	
Factor 2	Item4	I set up a general goal of what to achieve.	Planning strategies
	Item1	I have a general idea about what I am going to hear.	
	Item3	I have a clear plan of how I will proceed in the lecture.	
	Item2	I preview related readings to help me better understand the lecture.	
	Item5	I have specific purposes in mind.	
Factor 3	Item37	I consider how well I could comprehend and what I may do differently in the next lecture.	Appraisal strategies
	Item36	I evaluate my overall understanding.	
	Item31	I ask myself if I am satisfied with my level of understanding.	
	Item35	I compare what I understand with what I have known about the topic.	
	Item29	I check to see if I get the gist of the lecture.	

(continued)

Table 20.15 (continued)

		Factor	
	Item	Content	Factor name
Factor 4	Item8	I translate what I hear into my native language.	Translation strategies
	Item20	I translate word by word in my head as I listen.	
	Item 32	I check whether I have understood without always translating.	
Factor 5	Item11	I summarise information about the lecture.	Summary strategies
	Item10	I take notes or highlight important information.	

## Step 11: Forming Factor Scores or Composite Scores

This step deals with the process of translating factors into scores for each individual participant. DiStefano, Zhu, and Mîndrilă (2009) provide a comprehensive discussion of the various ways factor scores can be used. In this chapter, two methods are presented: *factor score estimates* and *composite scores*. First, SPSS can be instructed to create a factor score estimate for each participant. In SPSS, there are three options for factor score estimates, namely, *regression*, *Bartlett* and *Anderson-Rubin*. Regression is commonly used to create factor score estimates. Basically, this factor score estimate is the sum of the regression coefficients of items influenced by a common factor.

To obtain a factor score estimate in the SPSS data sheet, first follow the instructions in Fig. 20.5 and click “Scores.” In Fig. 20.18, tick “Regression.” It is not necessary to tick “Display factor score coefficient matrix.” Click “Continue” and then “OK.” The outputs will be the same as what have been previously used, but in the data sheet, new factor scores will have been added to each case (see Fig. 20.19). These new factor score estimates can then be used in subsequent analysis.

Second, we can instruct SPSS to create a composite (also known as a coarse score) of a factor for each participant. A composite is a raw score (or unweighted composite score) comprised of items loading on a factor. A composite score is accessible for readers as the value can be interpreted using a scale descriptor (e.g., 1 = not at all true of me to 5 = very true of me). A common term for this is *item parceling* (Brown, 2015), which helps create richness of data (i.e., a construct is represented by a number of items, rather than by one item). Table 20.14 can be used to create a composite score. To generate a composite score in SPSS, go to “Transform” and then “Compute variable.” Figure 20.20 is an example of how to create a composite score for

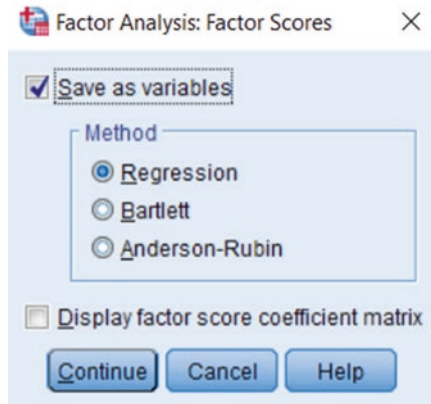


Fig. 20.18 Creating a factor score

	item36	item37	FAC1_1	FAC2_1	FAC3_1	FAC4_1	FAC5_1
1	2.00	2.00	.84038	-.63090	-1.00177	-1.30940	-.62687
2	3.00	2.00	-1.73820	.51104	-.25318	-1.31792	1.36136
3	4.00	3.00	-.21146	.75367	.84824	1.12557	.46582
4	4.00	3.00	-.27061	.81466	.68498	.36655	-.49110
5	1.00	1.00	1.52863	-1.05203	-1.37964	-1.33149	1.97029
6	1.00	1.00	.53980	-.94364	-2.10665	1.88644	.59027
7	3.00	3.00	-.43765	1.47804	.49292	.48575	-1.08413
8	1.00	1.00	-1.71437	-.76635	-.97644	-.16718	.25701
9	4.00	4.00	.86501	1.54521	.71539	-.60222	-.32701
10	5.00	4.00	.26322	1.27959	1.21928	.38119	-.42897

Fig. 20.19 Factor scores in the SPSS data sheet

“Comprehending Strategies.” It should be noted that the items are within brackets, and the sum is divided by the number of items added to scale the score to a value in the range between one and five. Once “OK” is clicked, a new variable will be added to the SPSS data sheet. Repeat the same process for other composites.

The factor score estimate and composite score of the same factor should be strongly correlated as they are based on the same information. When possible, performing such correlational analysis should be done as a matter of course. A strong correlation provides evidence that either the factor or the composite scores to be used in subsequent analysis are appropriate. Table 20.16 presents the correlations between the factor scores and the composite scores.

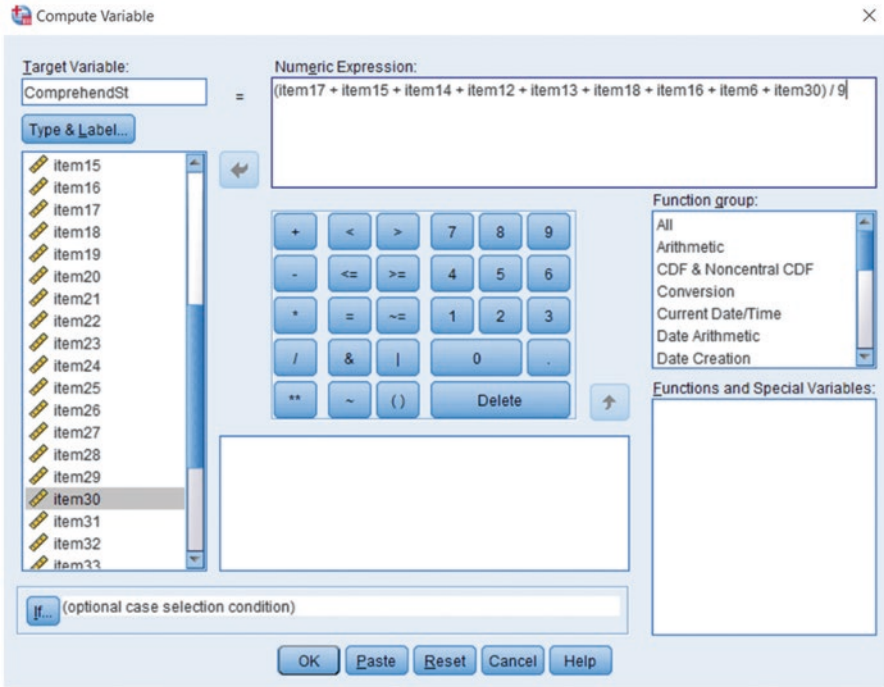


Fig. 20.20 Creating a composite score for comprehending strategies

Table 20.16 Correlations between factor scores and composite scores ( $N = 275$ , \*\* =  $p < 0.01$ )

	Comprehending	Planning	Evaluating	Translation	Summary
Factor1	0.959**				
Factor2		0.980**			
Factor3			0.919**		
Factor4				0.947*	
Factor5					0.910**

### Step 12: Performing Subsequent Analysis to Address Research Questions

EFA helps us ensure that the variables being used to address the research questions are sufficiently and appropriately representative of the constructs of interest. For example, variables representing extracted factors can be used to correlate with one another or variables from other instruments or sources, to compare independent variables, such as proficiency groups, first language background,

gender and so on (see, e.g., Dörnyei & Chan, 2013; O'Brien, 2014; Phakiti, 2003a). For example, Phakiti (2003b) employed EFA to examine a strategy use questionnaire ( $N = 384$ ) and found a two-factor solution (i.e., metacognitive and cognitive strategy factors). He then used Pearson correlation to examine the relationship between cognitive and metacognitive strategy use using a composite score based on the EFA results ( $r = 0.61$ ,  $p < 0.05$ ). The relationships of cognitive and metacognitive strategy use to reading test performance were also statistically significant ( $r = 0.391$  and  $0.469$ , respectively). Furthermore, three groups of test-takers (unsuccessful, moderately successful and highly successful) were compared in terms of their reported cognitive and metacognitive strategy use through multivariate analysis of variance (MANOVA). Statistical significance was detected among these test-taker groups. Phakiti (2003a) used the EFA results to examine gender differences. Peng and Woodrow (2010) first employed EFA to extract questionnaire items on willingness to communicate in English among Chinese learners of English. The researchers found a two-factor solution (meaning-focused and form-focused engagement) and subsequently performed CFA and SEM. In these sample studies, the researchers used EFA to reduce a set of data, obtain the convergent and discriminant validity of the constructs of interest, form a single score for each factor and employ subsequent statistical analysis to answer their research questions.

## Conclusions

Conducting EFA may present several challenges. These include dealing with missing data, achieving an adequate sample size, determining the number of factors to be extracted, attaining comparably sized cross-factor loadings, deciding on the cutoff factor loading values, dealing with low reliability estimates after EFA and finalising a final factor pattern. Perhaps one of the most controversial issues in EFA is that despite the number of EFA statistical methods available to help researchers determine the number of factors, patterns of factors and labels of factors, many decisions can be subjective, requiring researcher judgement and expertise theoretically and methodologically (see Fabrigar & Wegener, 2012; Osborne, 2014).

The intent for using EFA is to explore whether common factors influence observed responses. This chapter has provided an illustration of how to conduct EFA in applied linguistics. Despite being meticulous in EFA, it is important to keep in mind that “there is an indefinite number of equally fitting solutions for the parameter estimates” in EFA (Fabrigar & Wegener, 2012, p. 87). That is, there is a possibility that other patterns of factors (i.e., plausible rival EFA models) from the same dataset exist. Therefore, researchers cannot conclude with

absolute confidence that their EFA model has been proven because the fit of a single tested EFA model may always be an artefact of having tested too few models (see Brown, 2015).

Researchers are obliged to replicate their EFA models. For example, they can replicate their final EFA results using an independent sample (e.g., new data), develop a tentative CFA model or a second-order CFA based on the final EFA model, and/or with a new dataset, perform CFA models. Additionally, Rasch Item Response Theory (IRT) can be employed to investigate the level of difficulty to answer each item as well as item fitting (see, e.g., Bond & Fox, 2007, Chap. 11). Using multiple analyses allows researchers to obtain different information to enrich the understanding of a construct of interest. EFA is only useful when researchers have a solid understanding of its methodological principles. Finally, “An EFA is only as good as the data on which it is based” (Fabrigar & Wegener, 2012, p. 145).

## Resources for Further Reading

Bandalos, D. L., & Boehm-Kaufman, M. R. (2009). Four common misconceptions in exploratory factor analysis. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 61–87). New York: Routledge.

This chapter discusses the common misconceptions about EFA (viz., the choice between component and common factor extraction procedures, orthogonal versus oblique rotation, minimum sample size, the “eigenvalues greater than one” rule).

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10(7). Retrieved from <http://pareonline.net/getvn.asp?v=10&n=7>

This article provides a useful guideline for performing EFA, particularly focusing on issues in extraction, rotation, the number of factors to interpret and issues in sample size.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Los Angeles: Sage.

Chapter 17 in this book discusses key issues related to EFA and illustrates how to perform EFA in IBM SPSS.

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In Plonsky, L. (Ed.), *Advancing quantitative methods in second language research* (pp. 182–212). New York: Routledge.

This chapter gives a clear explanation of the principles and procedures behind both EFA and PCA. They also present a systematic review of factor analytic practices as observed in published second language research.

## References

- Asención-Delaney, Y., & Collentine, J. (2011). A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, 32(3), 299–322.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York and London: Guilford Press.
- Cheng, L., Andrews, S., & Yu, Y. (2010). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing*, 28(2), 221–249.
- Davis, J. M. (2016). Toward a capacity framework for useful student learning outcomes assessment in college foreign language programs. *Modern Language Journal*, 100(1), 377–399.
- DiStefano, C., Zhu, M., & Míndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20). Retrieved from <http://pareonline.net/getvn.asp?v=14&n=20>
- Dörnyei, Z., & Chan, L. (2013). Motivation and vision: An analysis of future L2 self images, sensory styles, and imagery capacity across two target languages. *Language Learning*, 63(3), 437–462.
- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford: Oxford University Press.
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (3rd ed.). Los Angeles: SAGE.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.
- Kozaki, Y., & Ross, S. J. (2011). Contextual dynamics in foreign language learning motivation. *Language Learning*, 61(4), 1328–1354.
- Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out Parallel Analysis.

- Practical Assessment Research & Evaluation*, 12(2). Retrieved from <http://pareonline.net/getvn.asp?v=12&n=2>
- Mizumoto, A., & Takeuchi, O. (2011). Adaptation and validation of self-regulating capacity in vocabulary learning scale. *Applied Linguistics*, 33(1), 83–91.
- Murray, J. C., Riazi, A. M., & Cross, J. L. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW, Australia. *Language Testing*, 29(4), 577–595.
- O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, 64(3), 715–748.
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, 32(3), 396–402.
- Osborne, J. W. (2014). *Best practices in exploratory factor analysis*. Lexington, KY: CreateSpace Independent Publishing Platform.
- Peng, J. E., & Woodrow, L. (2010). Willingness to communicate in English: A model in the Chinese EFL classroom context. *Language Learning*, 60(4), 834–876.
- Phakiti, A. (2003a). A closer look at gender differences in strategy use in L2 reading. *Language Learning*, 53(4), 649–702.
- Phakiti, A. (2003b). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading comprehension test performance. *Language Testing*, 20(1), 26–56.
- Phakiti, A. (2006). Modeling cognitive and metacognitive strategies and their relationships to EFL reading test performance. *Melbourne Papers in Language Testing*, 11, 53–102.
- Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, 65(S1), 9–36.
- Roever, C., & Phakiti, A. (2018). *Quantitative methods for second language research: A problem-solving approach*. New York: Routledge.





# 21

## Confirmatory Factor Analysis and Structural Equation Modeling

Aek Phakiti

### Introduction

Confirmatory factor analysis (CFA) and structural equation modeling (SEM) are multivariate statistical techniques that are used to test a hypothesis or theory. While CFA identifies and confirms a common factor (or set of factors) that underlies an ability or psychological construct, such as motivation or anxiety (in this way it is similar to exploratory factor analysis, or EFA), SEM establishes relationships (linear or causal, direct or indirect) among common factors or other observed variables. When CFA and SEM are used, theories, previous research and hypotheses inform researchers about a particular factor and its relationship to others. Typically, researchers develop a hypothesised CFA or SEM model, which is then tested using an empirical dataset, bearing in mind that it may be possible to test rival alternative or competing models that might well be theoretically or statistically acceptable.

Compared to EFA, CFA is not as widely used by applied linguists, perhaps because it requires a deeper technical knowledge of a number of underlying methodological concepts and techniques. Similarly, SEM is not as widely used as other standard statistics (e.g., correlations, multiple regression and analysis of variance; Plonsky, 2013). Apart from the fact that both CFA and SEM

---

A. Phakiti (✉)

Sydney School of Education and Social Work, The University of Sydney,  
Sydney, NSW, Australia

e-mail: [aek.phakiti@sydney.edu.au](mailto:aek.phakiti@sydney.edu.au)



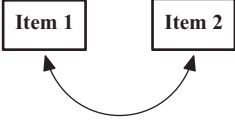
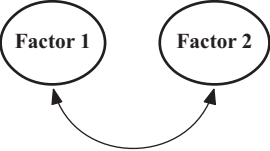
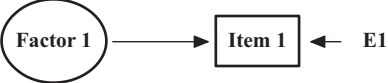
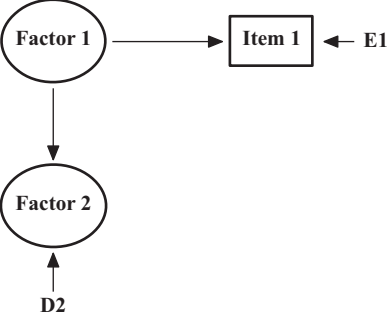
require a large sample size, which is not common in many applied linguistics projects, a lack of an accessible gateway to CFA and SEM methodology may impede their application. This chapter, therefore, presents the methodological basics of CFA and SEM techniques and provides some practical guidelines for how to perform CFA. Resources for further reading on CFA and SEM are also suggested. This chapter should be seen as complementary to recently published articles or chapters in applied linguistics (e.g., Hancock & Schoonen, 2015; Ockey, 2014; Ockey & Choi, 2015; Schoonen, 2015; Winke, 2014).

## What Is CFA?

CFA is a multivariate method for researchers to confirm that a hypothesised latent variable can be inferred from observed variables. In CFA, *the extent to which the items used to measure a latent variable are related not only to one another but also to the latent variable* can be determined by examining the loadings of observed variables. Like EFA, CFA is an *effect indicators model* (see Phakiti, Chap. 20), but unlike EFA, researchers believe that a hypothesised latent variable manifests individuals' observed behaviours or thoughts. Such a belief may be informed by a theory or previous research. Although researchers primarily aim to confirm whether a set of items is manifested in a hypothesised factor, they are also able to explore whether a particular item statistically fits in the hypothesised CFA model (i.e., exploratory CFA or ECFA). In other words, conducting CFA is not merely about confirming or rejecting a factor model. It is also about revising, refining and retesting a CFA model using a well-defined set of criteria. A diagrammatic representation of a CFA model is commonly found in CFA or SEM research reports. Table 21.1 presents the common symbols used in CFA and SEM, with explanations, which are elaborated further throughout the chapter.

Higher-order factor analysis is testable in CFA. Phakiti (2008), for example, employed CFA to test a fourth-order factor model of strategic competence based on four five-point Likert-type scale questionnaires (two trait and two state) collected over two occasions (two-month interval;  $N = 561$ ). Based on this study, strategic competence was confirmed as a higher-order factor that governs both declarative strategic knowledge and procedural strategic regulation. In this study, CFA started from a first-order factor, which is based on primary observed variables. A second-order factor is one that is inferred from first-order factors and a third-order factor is one that is based on second-order factors. Figure 21.1 is an example of a third-order CFA model (note that the observed items in the first-order factors are excluded).

Table 21.1 Common symbols used in CFA and SEM

Symbols	Explanation
	<ul style="list-style-type: none"> <li>A circle or ellipse indicates a latent variable or factor.</li> </ul>
	<ul style="list-style-type: none"> <li>A rectangle or box indicates an observed variable or indicator.</li> </ul>
	<ul style="list-style-type: none"> <li>A double-headed arrow implies a linear or non-directional relationship between two observed variables.</li> <li>A value on a double-headed arrow is a correlation coefficient.</li> </ul>
	<ul style="list-style-type: none"> <li>A double-headed arrow between circles or ellipses implies a linear or non-directional relationship between two latent factors.</li> </ul>
	<ul style="list-style-type: none"> <li>A single-headed or unidirectional arrow indicates a causal relationship from a factor to an observed variable or dependent variable.</li> <li>E1: Measurement error associated with Item 1 (observed variable 1). E1 is an error in prediction of Item 1 by Factor 1. Some SEM programs may use "e" instead of E. Error is an independent variable.</li> <li>A value on the single-headed arrow is a regression or path coefficient.</li> </ul>
	<ul style="list-style-type: none"> <li>An example of a basic structural model of two factors and an observed variable.</li> <li>An independent factor is called exogenous (Factor 1). A dependent factor is called endogenous (Factor 2).</li> <li>D2: Residual error (disturbance/error) in prediction of the dependent latent factor (i.e., Factor 2) by Factor 1.</li> </ul>

While EFA only produces a standardised factor loading in the form of a correlation or regression coefficient, CFA and SEM can produce “completely standardised,” “standardised” and/or “unstandardised” solutions for factor loadings and parameter estimates. The *completely standardised solution* refers to all parameter estimates that have a value between 0 and 1 (i.e., the total of the factor variances equals one). This is the same as that used in EFA. Based

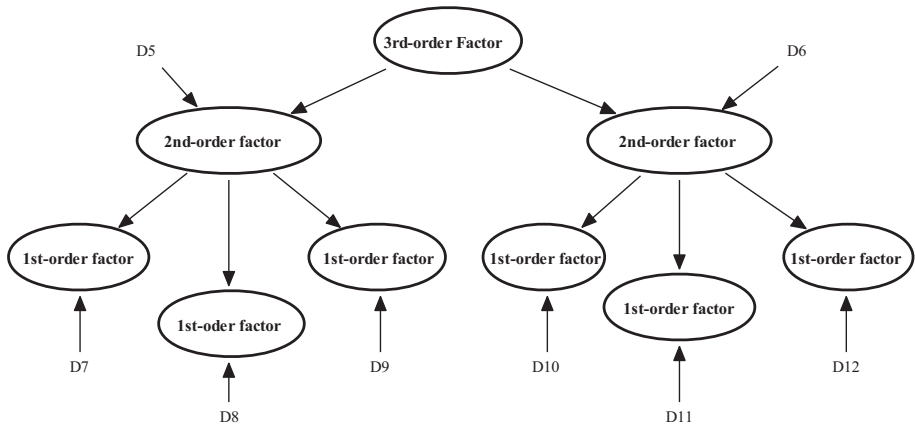
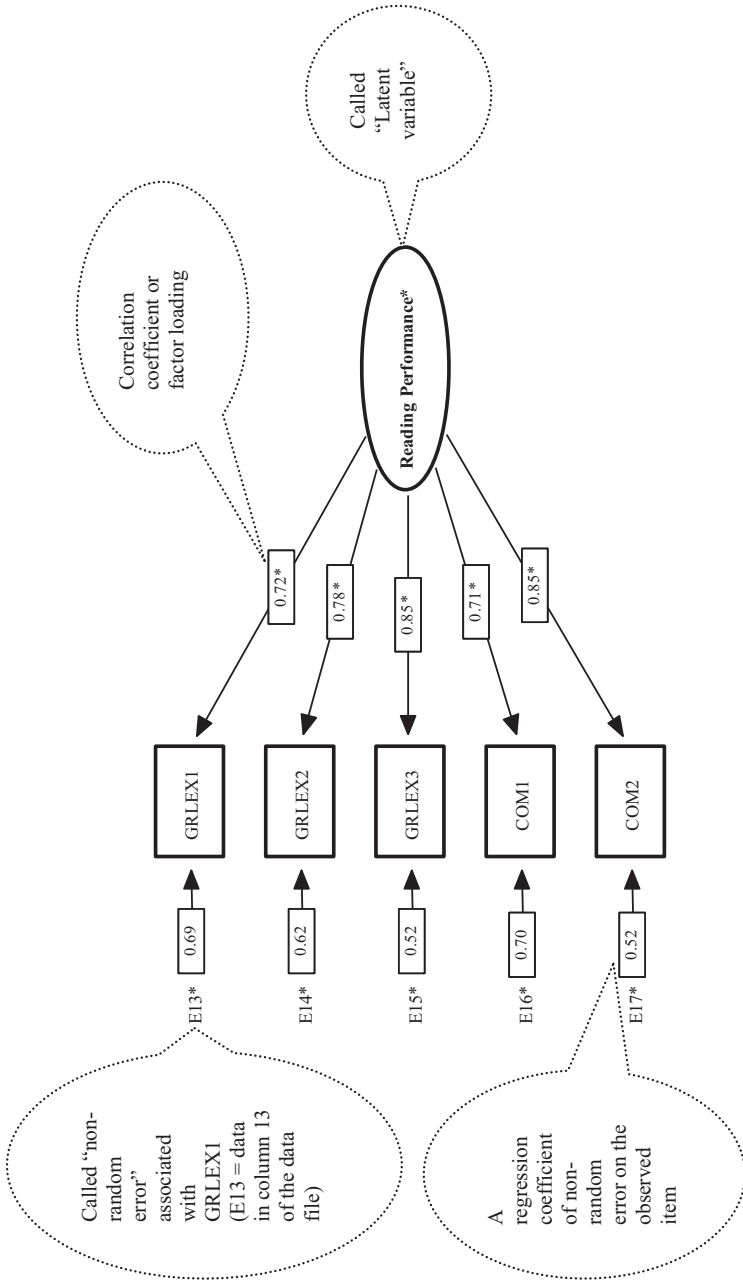


Fig. 21.1 A third-order factor CFA model

on Bentler (2006), the standardised solution makes it easy to interpret the variables in the linear structural equation system. The *unstandardised factor solution* refers to that which uses parameter estimates from the original matrix of the raw data, so that the values can be outside the range from 0 to 1. The *standardised solution* uses parameter estimates that express the relationships among standardised and unstandardised parameter estimates. In CFA, unstandardised parameter estimates are used when researchers aim to compare CFA models among different groups of participants (i.e., testing of parameter invariance, which is similar to how analysis of variance, for example, is used).

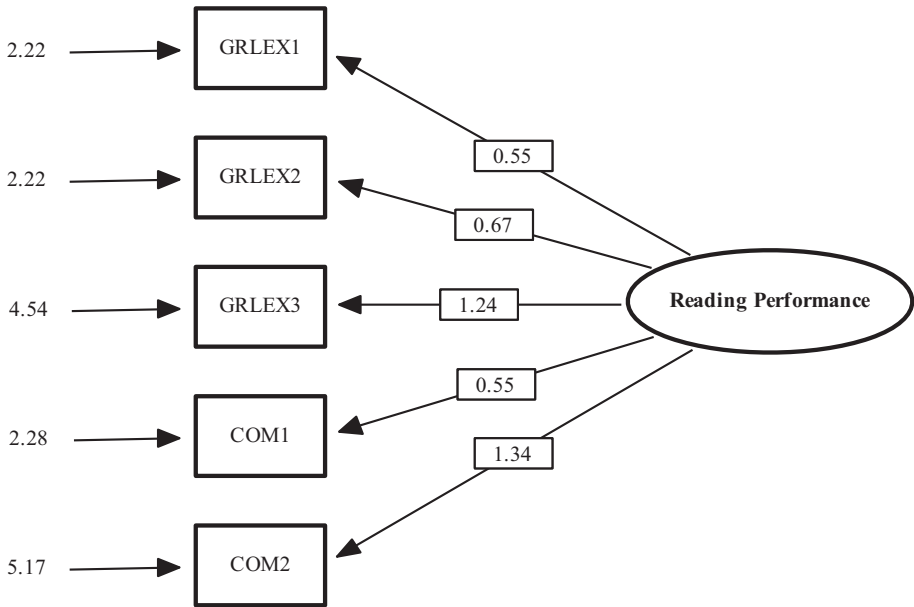
Figure 21.2 is a CFA model on reading performance by Thai English as a foreign language (EFL) test-takers ( $N = 651$ ). Dash bubbles are added to explain the key components in the CFA model. The parameter estimates are based on the standardised solution in which all observed variables (Vs), factors (Fs) and non-random errors (Es) are rescaled to have a variance of 1.0. All factor loadings were statistically significant at the 0.05 level. These factor loadings were large (values above 0.70). An error variance (E) is computed as follows:  $\sqrt{(1 - r^2)}$ , where  $\sqrt{\quad}$  = squared root and  $r$  = factor loading. For example, in the case of GRLEX1, E13 is computed as:  $\sqrt{(1 - 0.72^2)} = \sqrt{(1 - 0.52)} = \sqrt{0.48} = 0.69$ . One may wonder why the error variance is calculated this way. A simple answer to this is that SEM is *an equation-based approach*, so that  $r^2 + e^2 = 1$  (i.e.,  $0.52 + 0.48$ ). Note that in some SEM programs such as LISREL and AMOS,  $e^2$  may already be included in the model, so the reader needs to be careful when interpreting what researchers say about errors in their models.



Chi-square ( $\chi^2_{(4)} = 5.07, p = 0.28; CFI = 1.00; RMSEA = 0.02 (90\% CI = 0.00, 0.07)$ )

GRLEX = Grammatical-Lexical Ability COM = Reading Comprehension

Fig. 21.2 CFA model of reading performance (Standardised solution; N = 651)



**Fig. 21.3** CFA model of reading performance (Unstandardised solution;  $N = 651$ )

In an attempt to understand how well the observed variables measure a latent variable in CFA, researchers can compute the total common factor variance ( $h^2$ ) as *the sum of squared factor loadings, divided by the number of variables*.  $1 - h^2$  is then the amount of *unexplained variance*, or the extent to which the latent variable is not defined by the observed variables. In Fig. 21.3,  $h^2$  was 0.606 (i.e.,  $(0.72^2 + 0.75^2 + 0.85^2 + 0.71^2 + 0.85^2) \div 5 \rightarrow 3.03 \div 5 = 0.606$ ). The total common factor variance provides evidence of the *convergent* and *discriminant validity* of the latent model of interest (see Phakiti, Chap. 20). Figure 21.3 presents an unstandardised solution of Fig. 21.2. Unstandardised solutions are used for testing parameter invariance (e.g., simultaneous comparisons among groups).

## What Is SEM?

SEM is a multivariate statistical technique that helps researchers evaluate the validity of a theory or hypothesis through the use of empirical data. CFA, which is normally performed using SEM software, is a typical measurement model in SEM. However, SEM can combine a single observed variable as a measurement model. Like CFA, SEM provides researchers with a comprehen-

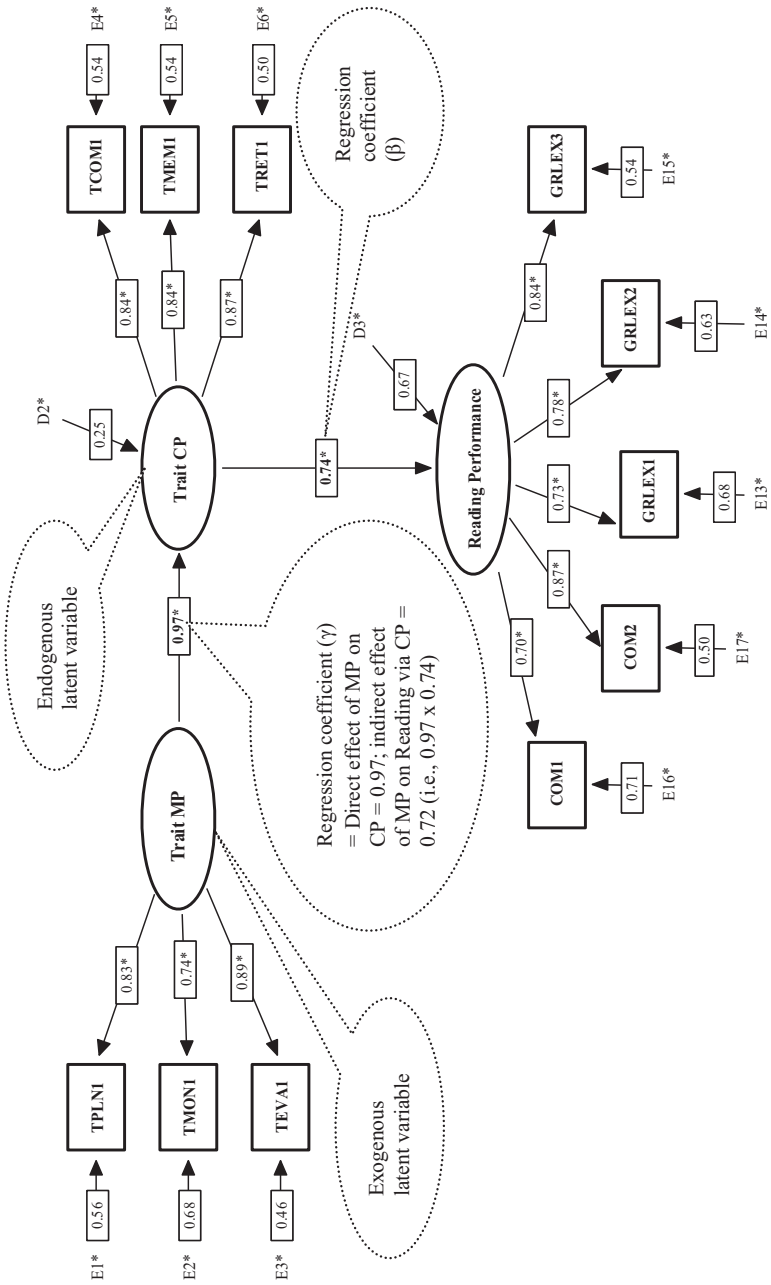
sive method for testing theories and examining data fit (illustrated further below). SEM can help researchers test a substantive theory (hypothesis testing) and organise concepts about data analysis into scientific models.

If all the variables in SEM are observed variables, the analysis is typically known as *path analysis* because parameter estimates are based on single observed ones (see, e.g., Rubinfeld & Clément, 2012; Vandergrift & Baker, 2015). That is, SEM models use latent variables (and/or observed variables), whereas path analysis uses observed variables. If one understands CFA and SEM, one will also understand the way in which path analysis is performed and interpreted. Unlike SEM, many path models do not include interaction effects, and the observed variables are assumed to be perfectly measured (Bohrnstedt & Carter, 1971). Pedhazur and Schmelkin (1992) pointed out that a measurement model in SEM should derive from multiple observed variables rather than from a single variable because using a number of observed variables to represent one construct can reduce the effect of measurement error.

Researchers usually present a SEM diagram, which allows readers to understand the complex (direct and indirect) relationship among various constructs. It is important to remember that SEM models are not automatically generated. Researchers need to draw them up based on a set of theoretical assumptions or hypotheses. An SEM diagram is usually accompanied by standardised parameter estimates (discussed above). When researchers wish to determine whether a parameter estimate is statistically significant at 0.05, they use the unstandardised portion of the output and examine whether the coefficient is significant (i.e., the associated *z*-score should be  $\geq 1.96$ ). The following is an example of an EQS output (Item 6) on measurement equations with standard errors and test statistics; the significance value at the 0.05 level is marked with @. In this example, the associated *z*-score is 7.487, which is larger than 1.96, suggesting that the parameter estimate for this item is statistically significant.

```
ITEM6    =V7    =          .721*F1    +          1.000 E7
          .096
          7.487@
```

Figure 21.4 presents an SEM model of the influences of trait metacognitive (MP) and cognitive processing (CP) on reading performance (also see Fig. 21.1, which is the hypothesised SEM model). This SEM model had good model fit (e.g., CFI = 0.98 and RMSEA = 0.06; model fit is explained further below). Note that an explanatory bubble is added in this figure. The observed variables for latent trait MP (i.e., TPLN1, TMON1 and TEVA1) and trait



Chi-square ( $\chi^2_{(39)} = 128.63, p = 0.00; CFI = 0.98; RMSEA = 0.06 (90\% CI = 0.05, 0.07)$

Fig. 21.4 A hypothesised SEM model of the influences of trait cognitive and metacognitive processing on reading performance (N = 651)



CP variables (i.e., TCOM1, TMEM1 and TRET1) were derived using an *item parcelling* technique. For example, items used to measure trait planning (i.e., items 1–6) were combined and averaged to create one single observed score (i.e., a composite score; see also Phakiti, Chap. 20). Item parcelling was adopted for practical reasons (more complex SEM models could be more easily built on at a later stage). Prior to this, a measurement model for each of these parcelled observed scores was tested and the models were found to have good fit.

Technically, in SEM, a latent variable that is used to predict another latent variable is called an *exogenous* (or independent) variable, but the latent variable being predicted is called the *endogenous* (or dependent) variable (Schumacker & Lomax, 2016). In Fig. 21.4, there is one exogenous latent variable (i.e., metacognitive processing), and there are two endogenous latent variables (i.e., cognitive processing and reading performance). The error variances associated with the observed variables are labelled as E, and error variances associated with endogenous variables are referred to as disturbance (D). In SEM, we are interested not only in model fit but also in *parameter estimates* (i.e., relationships among variables) and their meaningful interpretation in light of the associated theory. A regression coefficient of an exogenous latent variable on an endogenous latent variable is represented by *gamma* ( $\gamma$ , e.g., MP to CP), whereas that of an endogenous latent variable on another endogenous variable is represented by *beta* ( $\beta$ , e.g., CP to reading performance).

According to Fig. 21.4, trait MP had a strong direct effect on trait CP ( $\gamma = 0.97$ ;  $R^2 = 0.94$ ). This value of  $R^2$  means that trait MP accounted for 94% of the trait CP variance. This strong association implies that a large proportion of variance between the two constructs is shared. However, it may suggest a *unidimensional model* of trait strategic processing, which accounts for both MP and CP; a second-order CFA could subsequently be tested. Simply put, it may be plausible that trait MP and CP variables had *multicollinearity* (i.e., the measured variables were so highly related that they were essentially redundant; see Weston & Gore, 2006). In Fig. 21.4, CP had a large direct effect on reading performance ( $\beta = 0.74$ ;  $R^2 = 0.55$ ). Based on this parameter estimate, trait CP accounted for 55% of the reading performance variance. This relationship, however, is partially mediated by trait MP. That is, in this SEM model, trait MP had an *indirect* positive effect on reading performance via trait CP ( $\gamma = 0.72$ ;  $R^2 = 0.52$ , large ES). It could be inferred that during reading test tasks, trait MP is essential, as it directs the way test-takers use certain cognitive strategies, thereby enhancing reading performance.

We can determine the total amount of explained variance in the reading performance by subtracting the squared disturbance error ( $D^2$ ) from 1 (i.e.,  $1 - D^2$ ).

For the reading performance,  $D3$  is the disturbance error, so the total amount of the reading performance explained by both trait MP and CP was 0.55 (i.e.,  $1 - D^2$ ;  $1 - 0.67^2 \rightarrow 1 - 0.45 = 0.55$  or 55%). An SEM model is also convenient to examine the indirect effects of other observed variables on endogenous latent variables. For example, using the parameter estimates in Fig. 21.4, observed trait planning strategy use (TPLN1) had a positive, moderate indirect effect on reading performance (i.e.,  $\gamma = 0.60 [= 0.83 \times 0.97 \times 0.74]$ ;  $R^2 = 0.36$ ). The degree to which trait planning strategy use indirectly affected lexico-grammatical performance 1 (LEXGR1) was 0.43 (i.e.,  $\gamma = 0.83 \times 0.97 \times 0.74 \times 0.83$ ;  $R^2 = 0.18$ ).

## Critical Requirements and Statistical Assumptions for CFA

Many requirements and statistical assumptions for CFA and SEM are the same as those required for EFA, as discussed in Phakiti (Chap. 20). For example, researchers need to report how the raw data have been handled (e.g., how missing data and outliers are dealt with). In this chapter, four critical requirements are highlighted.

1. *Sample size.* The issue of sample size is important in CFA and SEM given that both are inherently large sample techniques. There is no cut-and-dry rule of thumb to say how many participants are required in order for a CFA or SEM model to be reliable. However, most SEM researchers would agree that at least 200, if not 300, participants are needed. Kaplan (1995) suggests that the evaluation of models should primarily be based on power considerations, rather than on sample size. However, successively larger samples are needed when many measured variables are to be used in a CFA model or a complex SEM model because the number of parameters to be evaluated increases accordingly (Thompson, 2000). Caution should be exercised when performing an SEM study that has a sample size lower than 200.
2. *Reliability and validity of observed variables.* Researchers need to provide evidence of the reliability and validity of their data. Reliability is a critical consideration for CFA and SEM because if some behaviours, thoughts or skills cannot be observed consistently, the data associated with them cannot be valid. In a Likert-type scale questionnaire, a reliability estimate (e.g., Cronbach's alpha) of at least 0.70 should be found. In language tests, the reliability estimate (e.g., KR20; Cohen's kappa) should be at least 0.80, if not 0.90.

3. *Univariate normality*. Descriptive statistics are useful to determine whether the univariate normality requirement is met. Univariate normality is related to the extent to which data for each variable has a bell-shaped distribution. *Skewness* (indicating an asymmetrical distribution) and *kurtosis* (the peak and tails of the distribution) statistics are normally used for this purpose (ideally, values should be within  $\pm 1$ , but it should not exceed  $\pm 3$ ).
4. *Multivariate normality*. Multivariate normality of the observed variables indicates that the observations are independent and identically distributed. In CFA and SEM, the multivariate kurtosis statistic should be zero or near zero. *Mardia's coefficient* is used to evaluate whether the data are multivariately normal. Strictly speaking, Mardia's coefficient should not exceed 3. For the maximum likelihood estimation method, multivariate normality is critical to the success of model estimation. When the data are non-normally multivariately distributed, the model fit indices can be suspect (Bentler, 2006). When the multivariate normality assumption is not met, a correction method (e.g., Elliptical or AGLF) may be employed. It is useful to know that a SEM program (e.g., EQS) normally allows the researcher to check multivariate outliers (participants who have two or more extreme scores). Researchers need to consider removing *multivariate outliers* before CFA or SEM analysis is performed.

## Software

Popular software for conducting CFA and SEM includes LISREL (linear structural relationships), EQS (Equations), MPLUS and AMOS (Analysis of Moment Structure), which is an added SPSS module (see Kline, 2016, for a range of SEM programs). In this chapter, EQS Version 6.3 (Bentler, 1985–2018) is used to illustrate how to perform CFA and SEM (see Byrne, 2006).

## Major Steps in CFA and SEM

Chapter 20 provides details of the key EFA steps in CFA and SEM, especially when EFA is performed prior to CFA (e.g., examining descriptive statistics and reliability analysis). For the purpose of this chapter, eight major steps that researchers go through when they perform CFA and/or SEM are presented in Fig. 21.5. Despite the differences between the various SEM programs, these steps are common to all of them.

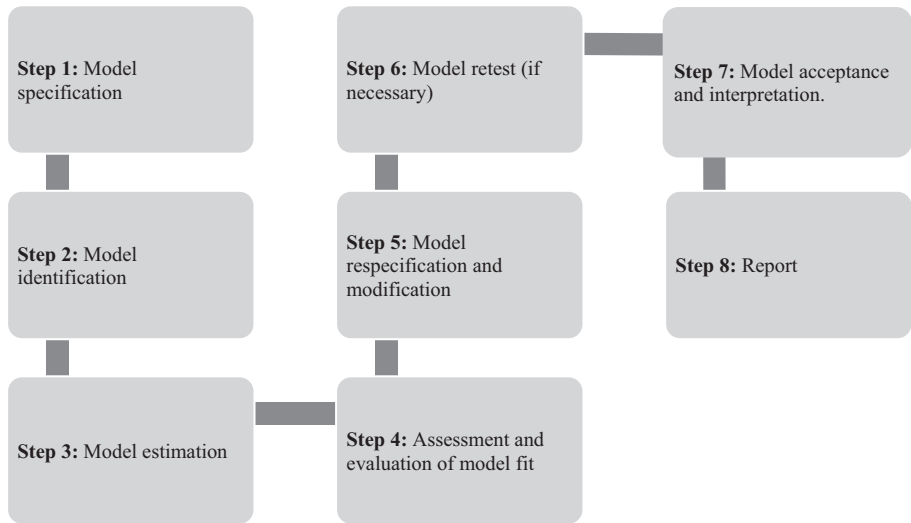


Fig. 21.5 Eight essential steps in CFA or SEM

## Step 1: Model Specification

This step is related to how researchers develop a CFA or SEM model to be tested. A hypothesised CFA or SEM model is typically generated from an extensive review of the relevant literature and theory. A CFA or SEM model cannot be valid or worth pursuing if it is based on a poor understanding of the nature of the construct of interest. A CFA or SEM is typically an *a priori* design which researchers have carefully considered before they start to collect data. However, in some research projects, researchers may use *archival data* (e.g., school examinations in Grant, McDonalds, Phakiti & Cook, 2014). Using archival data can, however, restrict model specifications because there may not be enough observed variables for a latent variable to be identified. When specifying a CFA or SEM model, researchers need to consider the measurement of both independent variables (that are assumed to have an influence) and dependent variables (that are assumed to be an outcome). In CFA, dependent or observed variables will then be used to form latent variables. In SEM, most observed variables are defined as being measured on a *linear continuous scale* (with the occasional *dichotomous scale* (e.g., taking on the value 1 or 0)). A CFA or SEM model is typically built from raw data that are then converted to a correlation or variance-covariance (i.e., unstandardised correlation) matrix. Researchers should make their variance-covariance matrix available for readers to view or use to retest or replicate the CFA or SEM model being reported (see In'nami & Koizumi, 2010).

## Step 2: Model Identification

Model identification involves considering *degrees of freedom* in parameter estimates. In classical univariate analysis, the degree of freedom total equals  $N - 1$ , where  $N$  is the sample size. For example, if there are scores of only two participants on only two variables, then the *model degree of freedom* is one (i.e.,  $2 - 1$ ) and the *degree of freedom error* is zero (i.e.,  $2 - 2$ ). In this case, no matter what the scores are on the two variables, the  $r^2$  value can only be one. The results are, however, trivial because the  $r^2$  analysis involves the scores of only two people on two measured variables. This scenario also applies to CFA and SEM in regard to model identification. In CFA and SEM, each parameter to be estimated (e.g., factor loading, regression weight, path coefficient, variance and covariance) needs at least *one degree of freedom*. It is, therefore, preferable to have not only models that have more degrees of freedom than the number of estimated parameters but also models that yield parameters that still do reasonably well at reproducing the matrix of associations (Brown, 2015; Kline, 2016; Schumacker & Lomax, 2016). The main reason for this preference is that there are “more ways” in which these models are potentially falsifiable.

In CFA and SEM, there are three levels of model identification: *under-identified*, *just-identified* and *overidentified*. A model is considered to be *under-identified* when one or more parameters cannot be uniquely determined because there is not enough information in the matrix  $S$ , the sample covariance matrix of the observed or indicator variables. Based on Ullman (1996), this means that there are fewer data points than parameters to be estimated (e.g., a case of two participants and two variables). It is essential to note that having more than one degree of freedom is a *necessary*, but *not always a sufficient condition* to satisfy model identification (Thompson, 2000). The parameter estimates of an under-identified model cannot be trusted (Schumacker & Lomax, 2016).

A model is *just-identified* when all the parameters can be uniquely determined because there is “just enough” information in the data matrix. Ullman (1996) argued that the analysis of just-identified models is uninteresting because there is the same number of data points as parameters to be estimated. In a just-identified model, the estimated parameters perfectly reproduce the sample covariance matrix, and the chi-square and degrees of freedom are equal to zero. Hypotheses based on just-identified models are not rigorous enough to be statistically tested in CFA or SEM.

Finally, a model is *overidentified* when there is more than enough information in the matrix  $S$ . For model identification, in CFA and SEM, the number of known variables (KVs; computed as  $\frac{1}{2} O(O + 1)$ , where  $O$  is the number of observed variables in the model) and the number of unknown variables (UVs) need to be compared (Schumacker & Lomax, 2016). The model is overidenti-

fied (i.e., desirable) when the number of UVs is less than that of KVs. For example, in a CFA model, with four observed variables, the number of KVs is 10 (i.e.,  $\frac{1}{2} 4(4 + 1)$ ) and the number of UVs is 9 (i.e., 4 errors, 4 factor loadings and 1 latent variable), thereby being overidentified. An overidentified model can be achieved through the use of many observed variables. For CFA and SEM, overidentified models are estimatable and testable, and therefore desirable.

### Step 3: Model Estimation

An explanation of CFA or SEM model estimation is complex and cannot be addressed fully here as it requires a discussion of formulas and explanations of the data matrices to be used (see also parsimony fit indices below). However, in brief, in this step, researchers aim to obtain estimates for each of the parameters specified in the model that produce the covariance matrix  $\Sigma$  in such a way that the parameter estimates are as close as possible to those in the sample covariance matrix of the observed or indicator variables  $S$  (see e.g., Schumacker & Lomax, 2016). This estimation method consists of the use of a particular fitting function to minimise the difference between  $\Sigma$  and  $S$ . In CFA or SEM, model estimation procedures (e.g., generalised least squares (GLS) and maximum likelihood (ML) are normally used.

### Step 4: Assessment and Evaluation of Model Fit

In CFA and SEM, a number of goodness-of-fit (GOF) criteria have been developed to assist researchers in evaluating a structural equation model under differing model-building assumptions. Table 21.2 summarises the key goodness-of-fit criteria that are useful for assessing model fit. Since different indices reflect different aspects of an SEM model, it should be expected that researchers report a variety of indices.

There are three kinds of fit indices that should be reported or examined in a CFA or SEM report: *absolute fit*, *parsimony fit* and *relative or incremental fit* indices.

*Absolute fit indices* are used to assess how well the a priori model fits the sample data. They indicate the extent to which the proposed theory or hypothesis fits the data. Fit indices for this category are chi-square ( $\chi^2$ ), root mean square residual (RMR), standardised root mean square residual (RMSEA) with 90% confidence intervals (indicating the precision of the estimate of fit), goodness-of-fit index (GFI) and the adjusted goodness-of-fit index (AGFI). The acceptable levels of these indices are quite straightforward (see Table 21.2). It should be noted that chi-square ( $\chi^2$ ) has been noted in terms of its sensitivity to sample size since it assumes multivariate normality. This assumption makes it

Table 21.2 Summary of the key goodness-of-fit criteria for CFA and SEM (based on, e.g., Brown, 2015; Byrne, 2006; Kline, 2016; Schumacker & Lomax, 2016)

GOF criteria	Acceptable level	Meaning and evaluation
<i>Absolute fit indices</i>		
Chi-square ( $\chi^2$ )	$\chi^2/df$ should be $\leq 3$	Compare obtained $\chi^2$ value with table value given <i>df</i> (degree of freedom). <i>p</i> -value should be larger than 0.001. <i>p</i> < 0.001 indicates that the event occurs less than one time in a thousand. A <i>nonsignificant</i> $\chi^2$ test implies that the data fit the model (unlike other standard statistics). A well-fitting model has a value of 0.05 or lower.
Root mean squared residual (RMR)	Indicates the closeness of $\Sigma$ to <i>S</i> matrix. 0 indicates perfect model fit.	A value lower than 0.05 indicates a good model fit. 90% confidence intervals should be used and provided.
Root mean squared error of approximation (RMSEA)	Value should be < 0.05, but not > 0.10	Value close to 0.95 reflects good model fit. Value close to 0.95 reflects good model fit.
LISREL goodness-of-fit index (GFI)	0 (no fit) to 1 (perfect fit)	
LISREL adjusted goodness-of-fit index (AGFI)	0 (no fit) to 1 (perfect fit)	
<i>Parsimony fit indices</i>		
Akaike information criterion (AIC)	Small values indicate well-fitting, parsimonious models	This index compares the value of the hypothesised model AIC with that of the independence AIC (i.e., null model). This index compares the value of the hypothesised model AIC with that of the independence AIC (i.e., null model).
Consistent Akaike information criterion (CAIC)	Small values indicate well-fitting, parsimonious models	
<i>Relative or incremental fit</i>		
Norm fit index (NFI)	0 (no fit) to 1 (perfect fit)	Value close to 0.95 reflects good model fit.
Non-norm fit index (NNFI)	0 (no fit) to 1 (perfect fit)	Value close to 0.95 reflects good model fit.
Comparative fit index (CFI) and Bollen (IFI)	0 (no fit) to 1 (perfect fit)	Value close to 0.95 reflects good model fit.



difficult for researchers to obtain non-statistical significance, particularly when a large sample size is used. Researchers should report the  $\chi^2$  statistic. This information allows the calculation of Wheaton, Muthen, Alwin and Summers' (1977) relative normed  $\chi^2$  as  $(\chi^2/df)$ . The value of  $\chi^2/df$  should be less than 3.0, but no larger than 5.0, which then suggests good model fit (despite a non-significant  $\chi^2$ ). In a CFA or SEM study, it is recommended that RMR and RMSEA are reported to complement  $\chi^2$  (Bentler, 2006).

*Parsimony fit indices* are fit statistics for addressing the issue of *parsimony* (i.e., the number of estimated coefficients required to achieve a specific level of fit). The AIC (Akaike information criterion) indicates that both model fit ( $S$  and  $\Sigma$ ) elements are similar and that the model is not overidentified (i.e., the model is parsimonious). The AIC and consistent AIC (CAIC) measures are commonly used to compare models with differing numbers of latent variables. Small values of AIC and CAIC indicate a parsimonious model with good fit (see Table 21.2).

*Relative or incremental fit indices* do not use  $\chi^2$  in its raw form. Rather, incremental fit indices compare the  $\chi^2$  value to a baseline model. Incremental fit indices include normed fit index (NFI), non-normed fit index (NNFI) and comparative fit index (CFI). NFI rescales  $\chi^2$  into a 0 (not at all fit) to 1 (perfect fit) range. An adjustment to the NFI to incorporate the degree of freedom in the model is known as the NNFI. Bentler (1990) developed a new coefficient of comparative fit in the context of specifying a population parameter and distribution to solve a problem of the NFI for nested models (i.e., models that are subsets of one another). NFI, NNFI and CFI values approaching or greater than 0.95 are indicative of a model with good fit. Such values indicate that the relative overall fit of the hypothesised model is about 95% better than that of the null model estimated with the same sample data. Values of 0.90 and above, however, are deemed sufficient to accept the tested model (Bentler, 2006). Note that NNFI may exceed the normed index in magnitude and can be *outside* the 0.00–1.00 range (Bentler, 2006).

After researchers have examined the model parsimony and gathered adequate evidence to support the fit of the hypothesised model, they may move on to the model interpretation stage (Steps 7 and 8) and attempt to answer their research questions. Note, however, that if the model does not satisfy the statistical criteria, the SEM researchers will need to, for example, think of a strategy to respecify and modify the model (i.e., Steps 5 and 6). The process of respecification and modification is briefly discussed below. It is important to stress that model fit is only part of the whole CFA or SEM process. Researchers “should not only aim to obtain the best model fit” through model respecification (e.g., adding new paths into a model to merely increase the values of fit indices). Doing so may be problematic because the parameter estimates obtained may be meaningless (e.g., some parameters may become statistically non-significant) and it may make a similar CFA or SEM model non-replicable by other researchers.



## Step 5: Model Respecification and Modification

This step may or may not be necessary in a CFA or SEM analysis, but CFA or SEM reports should make it clear whether the researcher has been through this process or not. This step takes place when a poor fit is found. If the true model is not consistent with the model being tested, then the model is assumed to be *misspecified*. There are a number of procedures for the detection of specification errors that SEM researchers can use to assist them in the respecification of their models. Researchers often use significance tests of parameter estimates for nested models to help them respecify a model (e.g., the *likelihood ratio* test (LR test), the *Lagrange multiplier* test (LM test; for adding parameters) and the *Wald* test (W test; for dropping parameters)). It is beyond the scope of this chapter to explain and illustrate these tests. See Schumacker and Lomax (2016) for further details. In CFA, a model respecification may be achieved by adding a correlation between two non-random errors (i.e., error covariance) or by dropping a variable. In SEM, a model respecification may include adding a new regression coefficient from one factor to another factor or adding a higher-order latent variable.

## Step 6: Model Retest

After a hypothesised model has been respecified, the model is retested following Steps 1 to 4. Step 5 may still be needed if the fit indices suggest poor model fit or unmeaningful parameter estimates.

## Step 7: Model Acceptance and Interpretation

After researchers have undertaken all the SEM statistical procedures outlined above and believe that the hypothesised model has met the necessary criteria, they then accept the model and interpret CFA or SEM, for example, by looking at the factor loadings, regression coefficient estimates ( $\beta$ ), squared multiple correlation ( $R^2$ ) and the decomposition of parameter total effects (direct and indirect effects on independent factors on dependent factors). It is important to look beyond statistical significance and good model fit. CFA and SEM findings that are statistically significant are not always worthy in a practical sense (Kline, 2016). Researchers should report and discuss the *effect sizes* of their statistical results. A statement in the *Publication Manual of the American Psychological Association* (APA, 6th edition) emphasises that a statistical significance *p*-value is not an acceptable index of effect size because it

depends on sample size. Another phrase for effect size is *magnitude-of-effect* (ME) estimates. An ME estimate that is related to CFA or SEM is shared variance ( $R^2$ ), which explains how much of the variability in the dependent variable(s) is associated with variation in the independent variable(s). Cohen (1992, p. 175) provides the following ES criteria to help interpret the product moment correlation coefficients:  $r = 0.10$  as small effect size;  $r = 0.30$  as medium effect size;  $r = 0.50$  as large effect size. It is worth noting, however, that these benchmarks generally underestimate the magnitude of correlations observed in applied linguistics. Based on a synthesis of 346 primary studies and 91 meta-analyses, Plonsky and Oswald (2014) suggest the following as very general guidelines for interpreting the magnitude of correlation coefficients:  $r = 0.25$  as small (corresponding roughly to the 25th percentile of correlations in their sample);  $r = 0.4$  as medium (~50th percentile);  $r = 0.6$  as large (~75th percentile).

## Step 8: Report

This step involves an explanation of steps 1 to 7, as well as other key components, such as a review of the literature and research methodology.

### Writing or Reading a CFA or SEM Report

The key steps for CFA and SEM outlined above can be used when writing or reading a CFA or SEM report. Such reports require a step-by-step, detailed presentation of how researchers have analysed data and dealt with issues arising during the procedures prior to accepting their CFA or SEM model. It is important for authors to report on the descriptive statistics of the data (including skewness and kurtosis statistics, which provide evidence that the data meets univariate assumptions) and reliability estimates of the measures or instruments used to establish the hypothesised CFA or SEM model. SEM researchers need to confirm whether the multivariate sample statistics for sample normality have been violated in their dataset.

The reports also need to assess whether individual parameter estimates (e.g., regression coefficients, factor loadings) are adequate or meaningful. The question of whether parameter estimates have the correct sign (plus or minus) and are within an expected reasonable range of values (e.g., in a standardised estimate, the value should be between 0 and  $\pm 1$ ) should be addressed. As Schumacker and Lomax (2016) pointed out, occasionally parameter estimates can take on impossible values, such as correlations that exceed 1, and there are

cases in which a negative variance is encountered. There are potential causes for such incidences, including outliers, an insufficient sample size and a lack of the required numbers of indicators per latent variable. When writing or reading a CFA or SEM report, the following questions should be addressed:

1. Is the review of the literature adequate to explain the hypothesised CFA or SEM model? Have a substantive theory, previous research and/or reasonable rationale been used to form the hypothesised CFA or SEM model?
2. Is the sample size used large enough (e.g.,  $N > 200$  or 300)? In a multi-group SEM, each group should be at least 200.
3. How have research instruments been developed? Were the data sufficiently reliable (e.g., Cronbach's alpha  $> 0.70$ )?
4. Are the descriptive statistics of observed variables reported? Were the data being used for CFA or SEM univariately and multivariately distributed? If the data were not multivariately distributed, what estimation method was employed?
5. Have all the CFA or SEM stages been clearly discussed? For example, was there sufficient evidence of good model fit? Were multiple fit statistics used to examine model fit?
6. In CFA, what was the total common factor variance ( $h^2$ )?
7. In SEM, was each measurement model (i.e., CFA) individually evaluated prior to evaluating a structural equation model?
8. Have other plausible rival/competing models been considered? In CFA or SEM, a model is only one of the many plausible models that may fit the dataset, so reasons for the adoption of the particular model being adopted need to be given.
9. Has the CFA or SEM model been adequately explained and interpreted in relation to the research questions, an underlying theory/hypothesis and previous relevant research?
10. Is the conclusion logical, reasonable and evidence-based?
11. Are the limitations of the study made clear and have directions for further research been provided?

## Doing CFA in EQS

For the purpose of this chapter, a CFA is illustrated through the use of EQS (Version 6.3). The outcomes of EFA presented in Chap. 20 (Table 20.14) are used to illustrate how CFA can be performed. Factor 1 as a first-order CFA is

illustrated. It is important to stress that to produce a robust SEM model, all individual CFA models need to be tested individually (and respecified if needed) before they are connected in a full SEM, as a model misspecification may have occurred at the CFA model level. A parsimonious process is arguably the best approach for SEM because before a final CFA or SEM model is achieved, several individual analyses and respecifications will have been performed, modified or corrected. Comprehensive instructions for this program can be found in Byrne (2006) and Bentler and Wu (2006).

## Opening and Naming a Data File

EQS can import an SPSS file and automatically convert it to the “.ess” file format for EQS use. Figure 21.6 presents an EQS dialog box, which is seen when opening a data file. In this dialog box, there are options for opening different types of EQS files (e.g., “.eds” for a saved diagram file and “.out” for saved EQS outputs). When naming EQS files, keywords describing the model should be used to make it easier to retrieve at a later stage (e.g., comprehending and comprehending\_revised2).

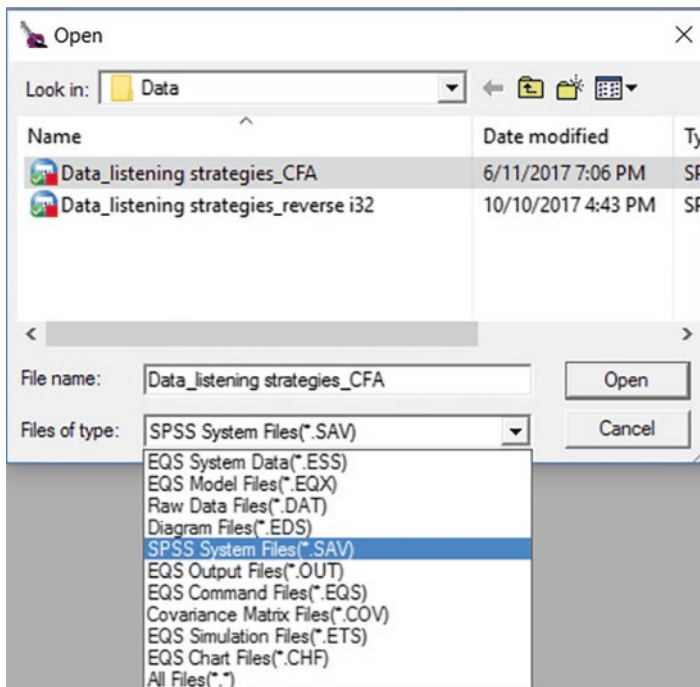


Fig. 21.6 Open a data file in EQS

EQS 6.3 for Windows - [data\_listening\_strategies\_cfa.ess]

File Edit View Data Analysis Data Plot Build\_EQS Window Help



	CODE	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6
1	1	3.0000	3.0000	2.0000	1.0000	2.0000	5.0000
2	2	4.0000	3.0000	3.0000	2.0000	4.0000	1.0000
3	3	3.0000	3.0000	3.0000	4.0000	4.0000	4.0000
4	4	4.0000	3.0000	4.0000	3.0000	3.0000	4.0000
5	5	4.0000	3.0000	1.0000	1.0000	1.0000	5.0000
6	6	3.0000	3.0000	1.0000	1.0000	1.0000	3.0000
7	7	4.0000	3.0000	4.0000	4.0000	4.0000	3.0000
8	8	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
9	9	5.0000	4.0000	3.0000	4.0000	4.0000	5.0000

Fig. 21.7 EQS spreadsheet

Figure 21.7 is a view of the EQS spreadsheet after clicking “Open.”

## Drawing a Hypothesised CFA

While syntax can be created for EQS to analyse CFA, it is more user-friendly for applied linguists to utilise a drawing tool and then ask EQS to create the corresponding syntax for the analysis. To do this, a data file must be opened, so that EQS can connect to the variables to be drawn.

1. Click  to draw a CFA model in EQS. A new page will appear as in Fig. 21.8).
2. There are four options here. Choose “Diagram Window: Open an empty diagram window.” A blank drawing canvas will appear (Fig. 21.9).
3. In the drawing canvas, click  from the several options on the left-hand side. Place your mouse on the blank canvas and click on it to open up a dialog box to add items of a CFA model (see Fig. 21.10). Here, move Items 6, 12, 13, 14, 15, 16, 17, 18 and 30 to the right-hand pane. At this stage, the default setting is appropriate. There is no need to label this factor just yet. All the labels of variables in a diagram should be added once the CFA model has been established. Click “OK.”
4. A CFA model will appear in the canvas (Fig. 21.11). By default, latent variables are in grey, observed variables are in yellow and arrows are normally in red. However, these colours can be edited (e.g., in black and white, bold or italicised text) in the final accepted model.

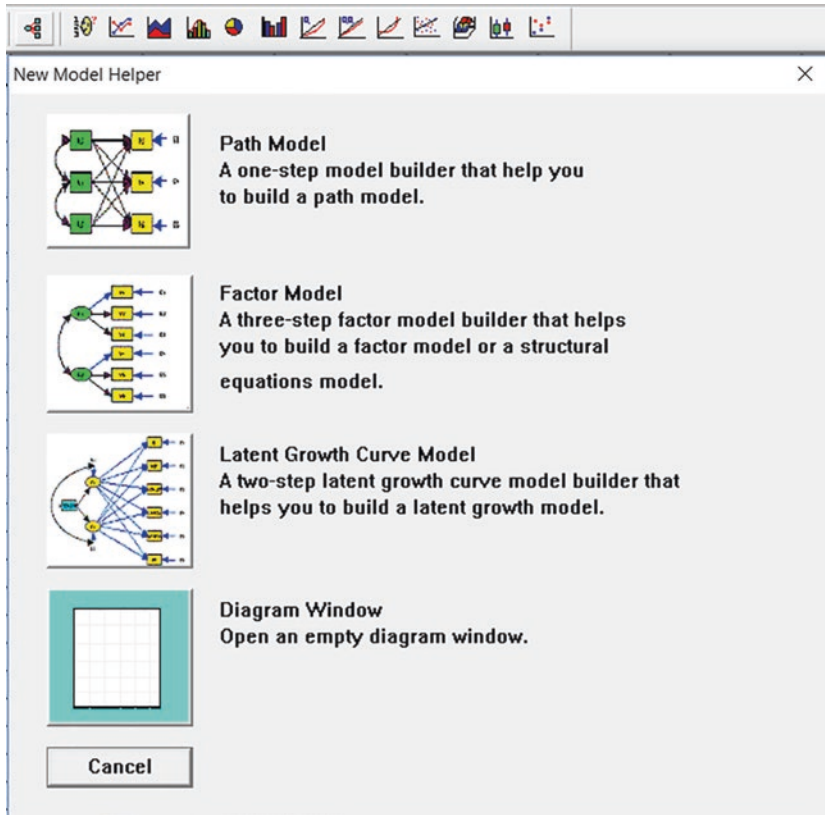



Fig. 21.8 EQS diagram drawing tool

5. On the top menu, click “save” (  ) and name the file as “comst” in a designated folder in which the EQS files are to be saved.
6. Next click “Build EQS” and choose “Title/Specifications.” A new dialog box will appear (Fig. 21.12), in which EQS is instructed to create the syntax for analysis. Options for SEM analysis include multisample analysis and multilevel analysis, which are beyond the scope of this chapter. Options for non-normal estimators corrections are also provided (when the multivariate normality assumption is not met; this will be revisited below). Here, make sure that ML (maximum likelihood) is chosen. Finally click “OK.” The text of a syntax file will appear (Fig. 21.13).
7. Click “Build EQS” again to run the analysis (Fig. 21.14). In this figure, we see that there are a number of active analysis options available for CFA and SEM. For example, “constraints” is used when a constraint is imposed on a particular model or parameter estimate; “Inequality” is used to test if two or

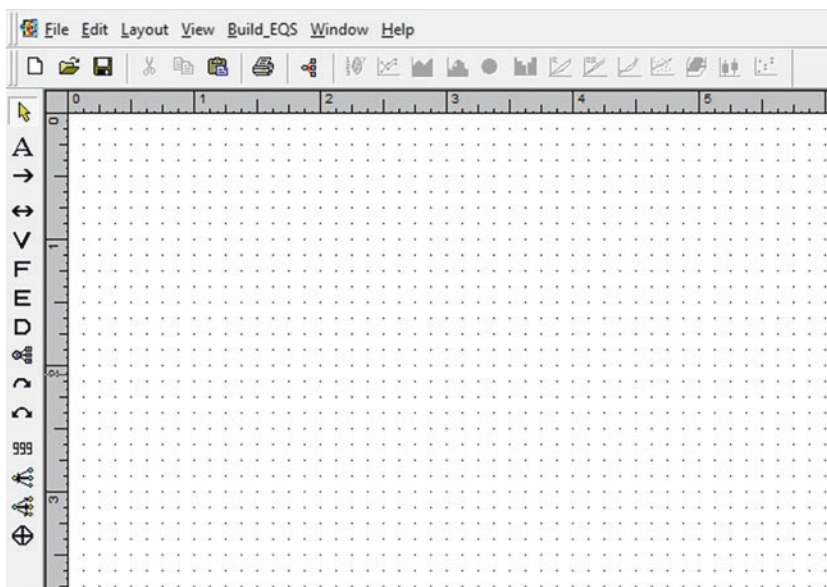


Fig. 21.9 EQS diagram drawing canvas

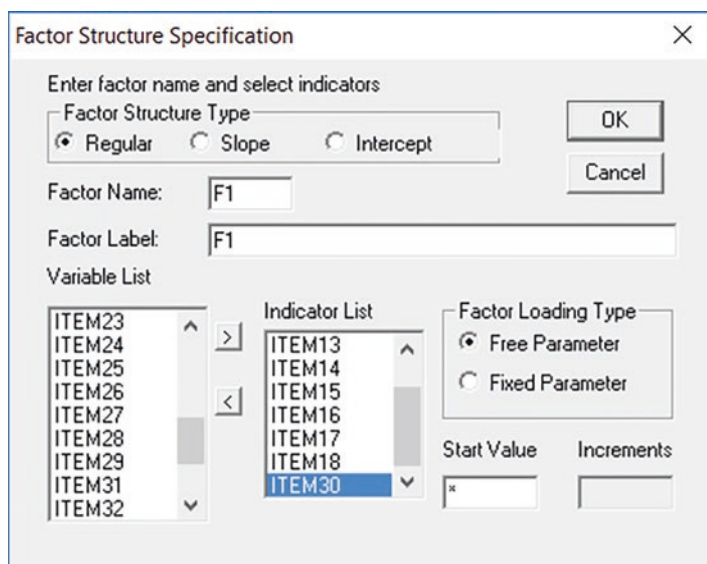


Fig. 21.10 Factor structure specification



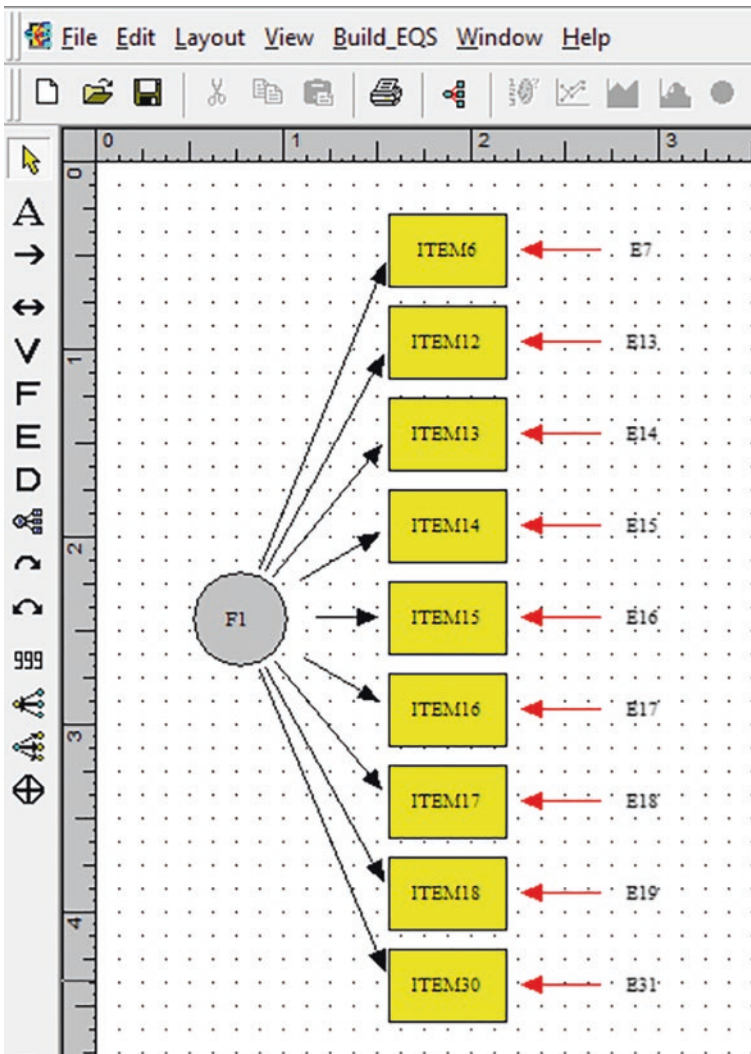


Fig. 21.11 A hypothesised CFA of comprehending strategies

more groups have model invariance; “LMTEST” is used to assess whether some parameters can be added to improve the model; “Wald Test” is used to assess whether some parameters may be dropped; “Print” is used to request more outputs (e.g., full decompositions of parameter estimates and fit indices). For now, we click “Run EQS.” A new dialog box will prompt us to save the file (the file name is already “comst”).

8. A message stating “Estimates have been inserted in the diagram file” appears (discussed below). Simply click “OK.” EQS will have produced three files



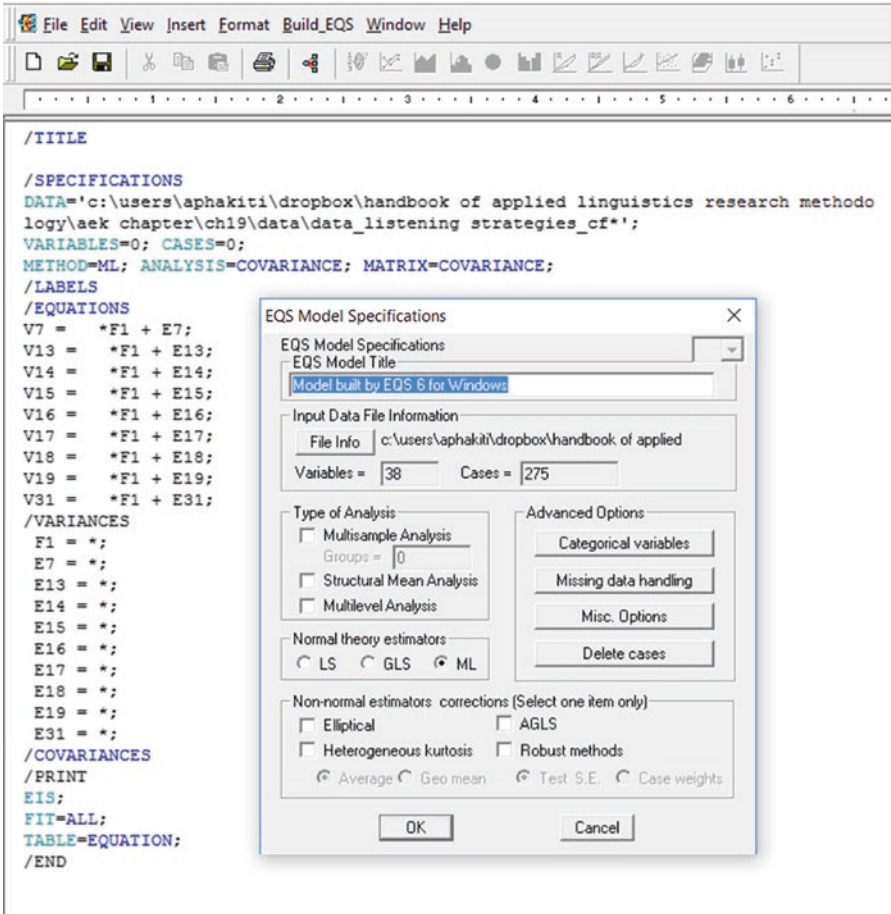
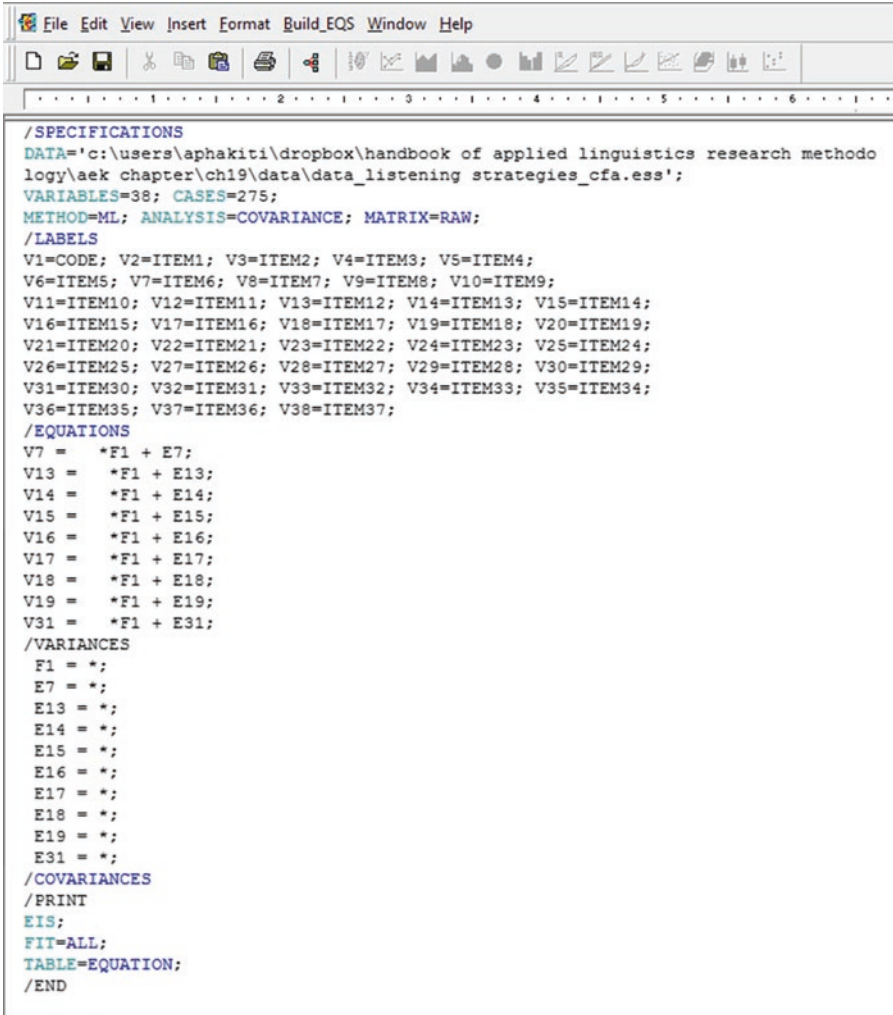


Fig. 21.12 EQS model specifications

including the *output file* (comst.out), which is currently displaying on screen, the *EQS syntax file* generated earlier (comst.eq; can be closed), and a *diagram file* (comst.eds) that was created at the beginning.

## EQS Outputs

The output file provides a number of important pieces of information to help us evaluate statistical assumptions: a covariance matrix (which can be reported in an appendix), model fit statistics, parameter estimates and statistical significance. The evaluation of model adequacy should be based on an inspection of the values of standardised residuals, the chi-square statistic and other



```

File Edit View Insert Format Build_EQS Window Help
[Icons]
... 1 ... 2 ... 3 ... 4 ... 5 ... 6 ...
/SPECIFICATIONS
DATA='c:\users\aphakiti\dropbox\handbook of applied linguistics research methodology\
aek chapter\ch19\data\data_listening_strategies_cfa.ess';
VARIABLES=38; CASES=275;
METHOD=ML; ANALYSIS=COVARIANCE; MATRIX=RAW;
/LABELS
V1=CODE; V2=ITEM1; V3=ITEM2; V4=ITEM3; V5=ITEM4;
V6=ITEM5; V7=ITEM6; V8=ITEM7; V9=ITEM8; V10=ITEM9;
V11=ITEM10; V12=ITEM11; V13=ITEM12; V14=ITEM13; V15=ITEM14;
V16=ITEM15; V17=ITEM16; V18=ITEM17; V19=ITEM18; V20=ITEM19;
V21=ITEM20; V22=ITEM21; V23=ITEM22; V24=ITEM23; V25=ITEM24;
V26=ITEM25; V27=ITEM26; V28=ITEM27; V29=ITEM28; V30=ITEM29;
V31=ITEM30; V32=ITEM31; V33=ITEM32; V34=ITEM33; V35=ITEM34;
V36=ITEM35; V37=ITEM36; V38=ITEM37;
/EQUATIONS
V7 = *F1 + E7;
V13 = *F1 + E13;
V14 = *F1 + E14;
V15 = *F1 + E15;
V16 = *F1 + E16;
V17 = *F1 + E17;
V18 = *F1 + E18;
V19 = *F1 + E19;
V31 = *F1 + E31;
/VARIANCES
F1 = *;
E7 = *;
E13 = *;
E14 = *;
E15 = *;
E16 = *;
E17 = *;
E18 = *;
E19 = *;
E31 = *;
/COVARIANCES
/PRINT
EIS;
FIT=ALL;
TABLE=EQUATION;
/END

```

Fig. 21.13 EQS model specifications

fit indices. The researcher's knowledge of the data and theoretical and conceptual aspects of the constructs under study should be considered together with these statistics (Jöreskog, 1971). In the analysis, four iterations were needed to reach convergence. According to Byrne (2006), the best scenario is a situation in which few iterations are needed to reach convergence.

Univariate and multivariate statistics are also provided. Based on the outputs, we learn that the multivariate normality assumption was not met (*Mardia's coefficient* ( $G_2, P$ ) = 29.4609, which is larger than 3, with an associated  $z$  statistic of 17.36). Accordingly, when we retest a revised CFA model,

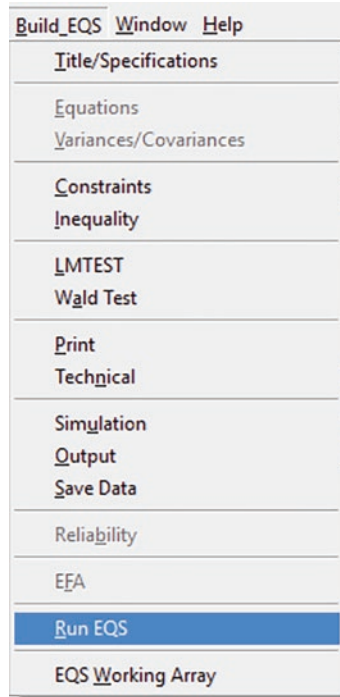


Fig. 21.14 Analysis in EQS

we can choose an option from the non-normal estimators corrections (see Fig. 21.12): *Elliptical*, AGLS (adjusted generalised least squares), *heterogeneous kurtosis* or *robust methods* (see Byrne, 2006). Elliptical is usually robust enough for this correction and will be chosen. The average absolute residual was 0.0362; the average off-diagonal absolute residual was 0.0452; the average absolute standardised residual was 0.0426; and the average off-diagonal absolute standardised residual was 0.0533. These values should be close to 0.

In CFA, the *standardised residuals* provide useful information for model respecification; these indicate the shared error variance between two variables. Any pair with a value above 0.10 should be considered for a model respecification. This information indicates some redundant content of a pair of items (i.e., issue of method effects); this can be corrected by simply correlating the errors of the pair. The EQS output suggests four pairs of variables have large standardised residuals (V14, V13 = 0.214; V19, V17 = 0.170; V31, V7 = 0.146; and V17, V7 = 0.125). In the model respecifications, the errors of these variable pairs will be specified as correlated and retested (see Fig. 21.20).

DISTRIBUTION OF STANDARDIZED RESIDUALS

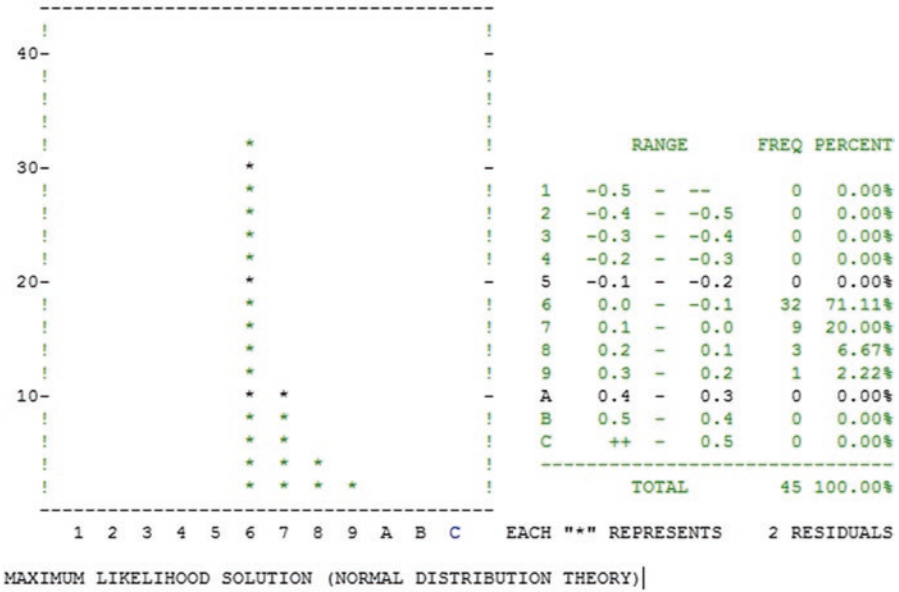


Fig. 21.15 Distribution of standardised residuals

Figure 21.15 presents the distribution of standardised residuals. A review of the frequency distribution reveals that all residual values (91%) are between  $-0.1$  and  $0.1$  and 8% are between  $-0.2$  and  $0.2$ . This suggests that the distribution was non-symmetric and was not quite centred around zero.

Turning to the goodness-of-fit summary (Table 21.3), the first of these indices was the independence chi-square statistic ( $\chi^2_{(36)} = 877.96$ ), which is reported for the likelihood ratio test of the Bentler and Bonett (1980) null hypothesis. The null model is typically used as a baseline against which alternative models can be compared for the purpose of evaluating the gain in improved fit (Byrne, 2006). Given a sound hypothesised model, the  $\chi^2$  value for the null model needs to be high. The chi-square of the alternative model, on the other hand, needs to be much smaller than that of the null model. If the fit of the hypothesised CFA model were to be close to that of the null model, there would be a question concerning the soundness of the hypothesised model.

As can be seen in Table 21.3, while the chi-square statistic was smaller than that of the independence chi-square ( $\chi^2_{(26)} = 122.912$ , with the normal theory

RLS chi-square for this ML solution of 122.404), it remains quite large, suggesting a poor model fit. The probability value for the chi-square statistic was significant ( $p = 0.000$ ). As discussed in Table 21.2, ideally the researcher needs to obtain a *non-significant*  $\chi^2$  value with the associated degree of freedom so that the model can be taken as fitting the data (a non-significant  $\chi^2$  test implies that the data fit the model). Hence, this information confirms that the hypothesised CFA model was not supported. In addition to the use of  $\chi^2$  statistics for the independent (null) and hypothesised CFA model, Akaike's (1987) information criterion (AIC) and Bozdogan's (1987) consistent version of the AIC (CAIC) can be used to assess model parsimony (i.e., to determine the number of estimated coefficients required to achieve a specific level of fit) in the assessment of model fit. As can be seen in Table 21.3, the AIC and CAIC statistics for both the null and hypothesised models had values close to those of the independent  $\chi^2$  statistics, suggesting a lack of model parsimony.

**Table 21.3** Model fit indices

GOODNESS OF FIT SUMMARY FOR METHOD = ML			
INDEPENDENCE MODEL CHI-SQUARE	=	877.966 ON	36 DEGREES OF FREEDOM
INDEPENDENCE AIC =	805.966	INDEPENDENCE CAIC =	639.762
MODEL AIC =	70.912	MODEL CAIC =	-49.124
AKAIKE INFORMATION CRITERION (AIC) BASED ON LOG LIKELIHOOD	=	5815.773	
BAYESIAN INFORMATION CRITERION (BIC) BASED ON LOG LIKELIHOOD	=	5884.492	
CHI-SQUARE =	122.912	BASED ON	26 DEGREES OF FREEDOM
PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS	0.00000		
THE NORMAL THEORY RLS CHI-SQUARE FOR THIS ML SOLUTION IS	122.407.		
FIT INDICES			
-----			
BENTLER-BONETT	NORMED FIT INDEX =	0.860	
BENTLER-BONETT	NON-NORMED FIT INDEX =	0.841	
COMPARATIVE FIT INDEX (CFI)	=	0.885	
BOLLEN'S	(IFI) FIT INDEX =	0.886	
MCDONALD'S	(MFI) FIT INDEX =	0.838	
JORESKOG-SORBOM'S	GFI FIT INDEX =	0.910	
JORESKOG-SORBOM'S	AGFI FIT INDEX =	0.844	
ROOT MEAN-SQUARE RESIDUAL (RMR)	=	0.054	
STANDARDIZED RMR	=	0.063	
ROOT MEAN-SQUARE ERROR OF APPROXIMATION (RMSEA)	=	0.117	
90% CONFIDENCE INTERVAL OF RMSEA (	0.096,	0.137)	
RELIABILITY COEFFICIENTS			
-----			
CRONBACH'S ALPHA	=	0.849	
RELIABILITY COEFFICIENT RHO	=	0.850	
MAXIMAL WEIGHTED INTERNAL CONSISTENCY RELIABILITY	=	0.871	

In Table 21.3, the NFI, NNFI and CFI that provide a measure of complete covariation in the data had values of 0.86, 0.84 and 0.88, respectively. Other indices, such as Bollen (IFI), LISREL goodness-of-fit index (GFI), root mean squared residual (RMR) and root mean squared error of approximation (RMSEA) also indicate a poor model fit. In summary, all the statistics discussed provide strong evidence not to accept this model but to respecify and retest it.

## Determining Whether Items Are Statistically Significant

We can examine the test statistic, known as a  $z$  statistic (i.e., the parameter estimate divided by its standard error) to determine whether items included in the CFA model are statistically significant (i.e., whether the factor loadings are significant). An alpha level is normally set at 0.05, and as discussed earlier, the test statistic needs to be *greater than 1.96 or less than -1.96* for the hypothesis that the estimate equals zero to be rejected. Table 21.4 presents the unstandardised solutions of all items for this CFA model. A review of the output file suggests that all estimates were statistically significant at the 0.05 level.

## CFA Diagram in EQS

In the .eds file (Fig. 21.16), click “View” and then “Estimates,” which has four options (Start Values, Parameter Estimates, Standardised Solution and Parameter Labels). “Start Values” indicates parameter estimates prior to the analysis, “Parameter Estimates” are unstandardised parameter estimates and “Parameter Labels” indicate if a variable is latent, observed or error. In the standardised solution, all V, F and E (D for later models) variables are rescaled to have a variance of 1.0.

Figure 21.17 is the first-order CFA model on comprehending strategy use. To evaluate parameter estimates, we examine (1) the appropriateness or feasibility of parameter estimates and (2) the statistical significance of the parameter estimates (e.g., Byrne, 2006; Kline, 2016). First, we determine whether any estimates fall outside the admissible range. Parameters that signal unreasonable estimates are, for example, correlations greater than 1.00, a negative variance and standard errors abnormally large or small (see e.g., Byrne, 2006).

All the factor loadings were statistically significant at the 0.05 level and ranged from 0.47 for Item 30 to 0.76 for Item 15.

Table 21.4 Test statistics

MEASUREMENT EQUATIONS WITH STANDARD ERRORS AND TEST STATISTICS				
STATISTICS SIGNIFICANT AT THE 5% LEVEL ARE MARKED WITH @.				
ITEM6	=V7	=	.726*F1 .113 <b>6.409@</b>	+ 1.000 E7
ITEM12	=V13	=	.758*F1 .086 <b>8.789@</b>	+ 1.000 E13
ITEM13	=V14	=	.697*F1 .085 <b>8.225@</b>	+ 1.000 E14
ITEM14	=V15	=	.869*F1 .073 <b>11.973@</b>	+ 1.000 E15
ITEM15	=V16	=	1.056*F1 .085 <b>12.464@</b>	+ 1.000 E16
ITEM16	=V17	=	.714*F1 .104 <b>6.889@</b>	+ 1.000 E17
ITEM17	=V18	=	.996*F1 .084 <b>11.925@</b>	+ 1.000 E18
ITEM18	=V19	=	.744*F1 .094 <b>7.910@</b>	+ 1.000 E19
ITEM30	=V31	=	.600*F1 .092 <b>6.522@</b>	+ 1.000 E31

### Need for a Model Respecification

On the basis of the CFA analysis results, this current CFA should be respecified given the poor model fit. A respecified CFA model is presented in Fig. 21.18 (with start values to be estimated). The LM test was employed, but it is beyond the scope of this chapter to discuss this test. The LM test results support a decision to revise the original model to by correlating the errors of four variable pairs (i.e., V14, V13; V19, V17; V31, V7; V17, V7). Bentler (2006) advises that it may be necessary to modify the structural antecedents



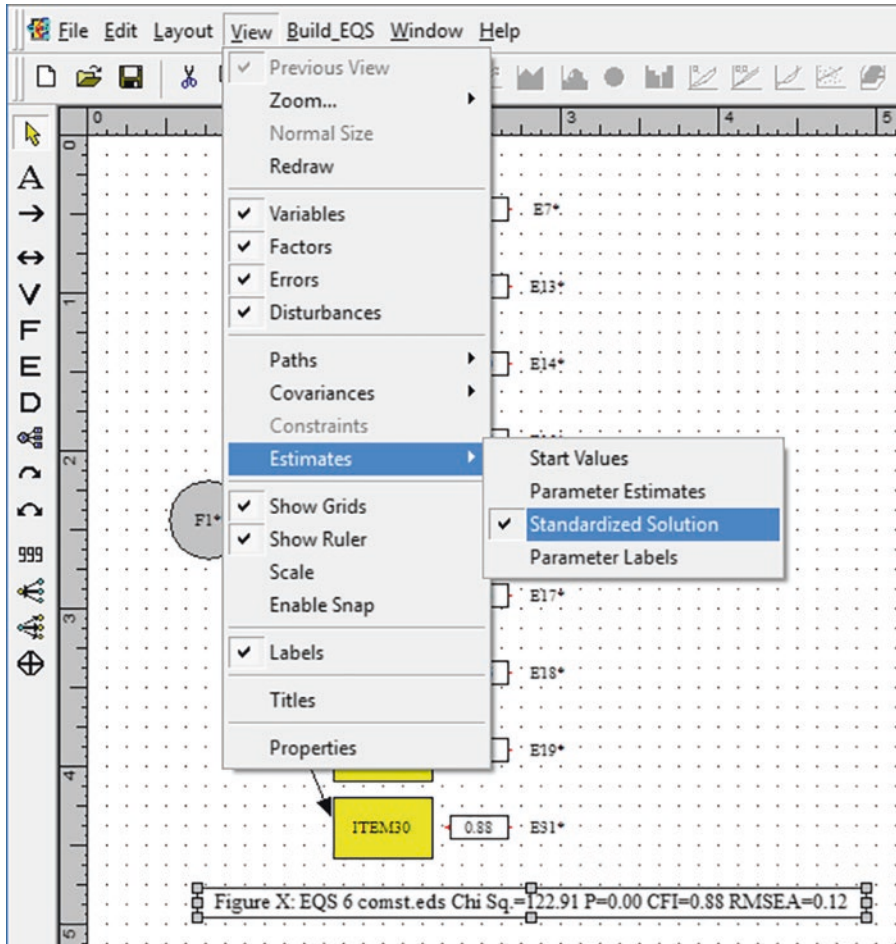


Fig. 21.16 Parameter estimates options

or correlates of these variables. If these variables are observed variables (which is the case here), the residuals associated with them are antecedents to the variables and hence can be correlated, so that they may be better explained in the model.

## Model Retest

The steps for analysing CFA outlined above were repeated, but cannot be included here for reasons of space. Note that in the EQS model specifications (see Fig. 21.12), “Elliptical” should be ticked.



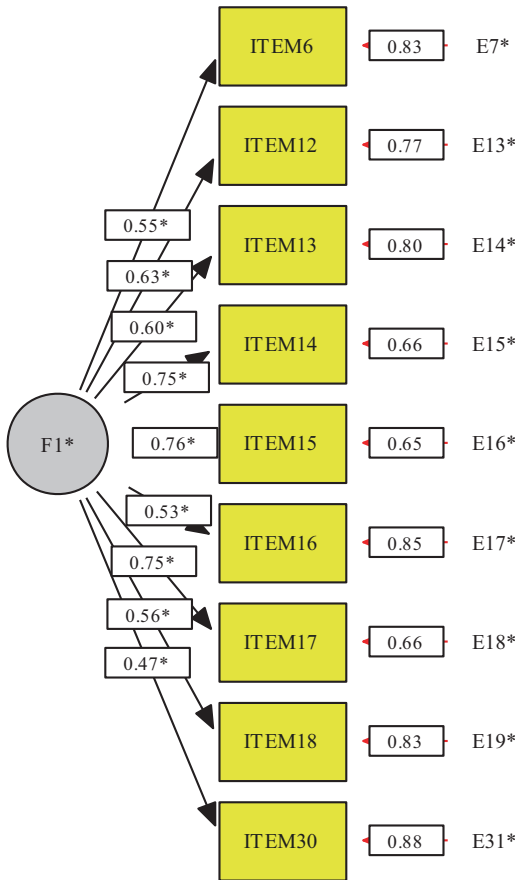


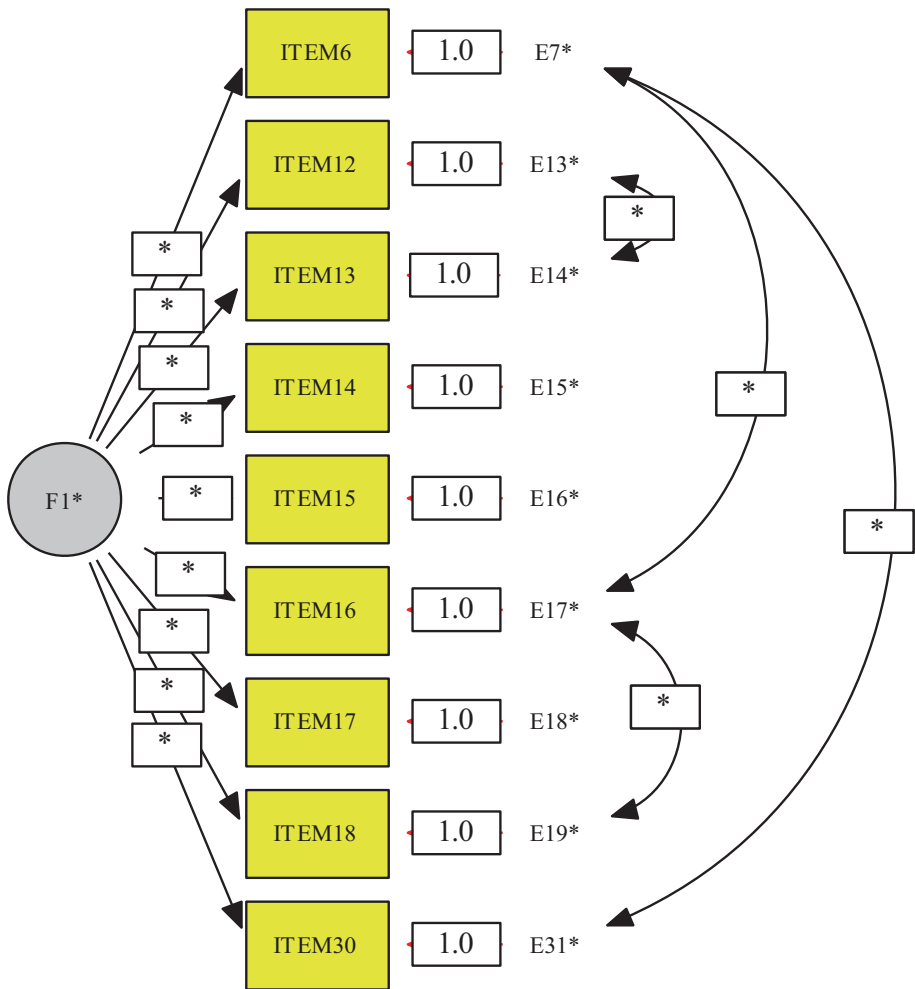
Figure X: EQS 6 comst.eds Chi Sq.=122.91 P=0.00 CFI=0.88 RMSEA=0.12

Fig. 21.17 First-order CFA for comprehending strategy use (unedited version)

### Outcomes of the Revised CFA Model

In this retest, two iterations were needed to reach convergence. A review of the frequency distribution reveals that all residual values are between  $-0.1$  and  $+0.1$  (see rows 6 and 7 in Fig. 21.19). The revised CFA model indicates a significant improvement on the first run.

According to Bentler (2006), an out-of-range coefficient suggests an inadequacy in the model. The values of the individual parameters were inspected and it was found that all were within the expected ranges ( $\pm 1$  range,  $\pm SE$  for standardised solution) and there were no anomalies. Table 21.5 presents and



**Fig. 21.18** Revised first-order CFA model for comprehending strategy use

compares the goodness-of-fit criteria and summary. In this table, the revised CFA model has much better fit indices, and all the indices suggest the revised CFA had good model fit. The chi-square statistic was also *non-significant* in the revised CFA model ( $p < 0.001$ ), which is uncommon as the chi-square statistic is sensitive to sample size (Byrne, 2006). As can be seen in Table 21.5, it should be noted that model modification and respecification do not improve the multivariate kurtosis statistic.

Figure 21.20 presents the revised CFA model of comprehending strategy use. Instructions on how to edit the model cannot be presented in this chapter for reasons of space (see e.g., Byrne, 2006, Chap. 4). All factor loadings and error correlations were statistically significant at the 0.05 level.

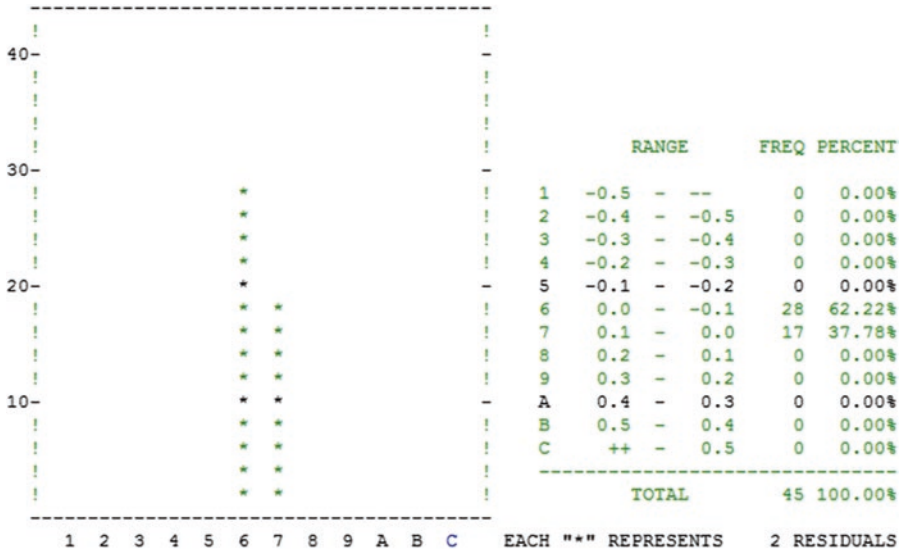


Fig. 21.19 Distribution of standardised residuals (revised model)

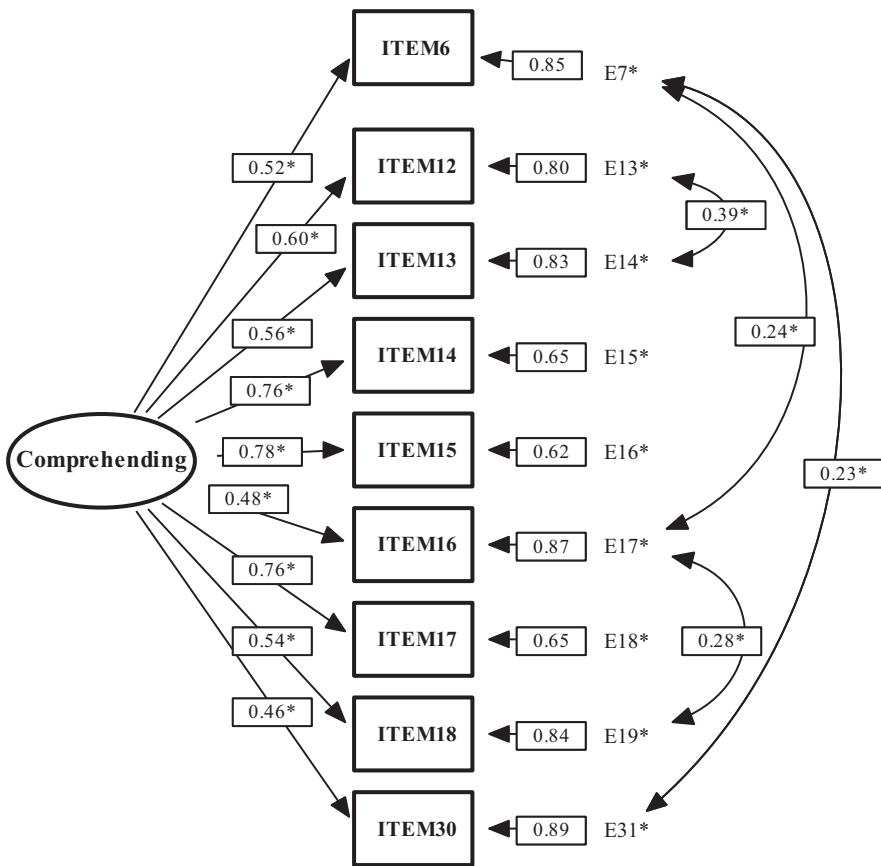
Table 21.5 The results of the revised first-order CFA model of comprehending strategy use

Goodness-of-fit criteria	Original CFA	Revised CFA
<i>Multivariate kurtosis</i>		
• Mardia's coefficient (G2,P)	29.4609	29.4609
• Normalised estimate	17.3600	17.3600
<i>Elliptical theory kurtosis estimates</i>		
• Mardia-based kappa		0.2976
• Mean scaled univariate kurtosis		0.1244
<i>Standardised residuals</i>		
• Average absolute standardised residuals	0.0426	0.0221
• Average off-diagonal absolute standard residuals	0.0533	0.0275
<i>Goodness-of-fit summary</i>		
<i>Null hypothesised model</i>		
• Independence model chi-square	877.966 (df = 36)	1235.987 (df = 36)
• Independence AIC	805.966	1163.987
• Independence CAIC	639.762	997.784
<i>One-factor hypothesised CFA</i>		
• Model AIC	70.912	-16.555
• Model CAIC	-49.124	-118.124
• Chi-square	122.912 (df = 26)	27.445 (df = 22)
• p-value for the chi-square statistic	0.00000	0.19480
<i>Fit indices</i>		
• Bentler-Bonett normed fit index (NFI)	0.860	0.978
• Bentler-Bonett non-normed fit index (NNFI)	0.841	0.993

(continued)

Table 21.5 (continued)

Goodness-of-fit criteria	Original CFA	Revised CFA
• Comparative fit index (CFI)	0.885	0.995
• Bollen (IFI)	0.886	0.996
• Joreskog-Sorbom's goodness-of-fit index (GFI)	0.910	0.972
• Root mean squared residual (RMR)	0.054	0.028
• Standardised RMR	0.063	0.032
• Root mean squared error of approximation (RMSEA)	0.117 (90% CI = 0.096, 0.137)	0.030 (90% CI = 0.000, 0.062)



Chi-square ( $\chi^2_{(22)} = 27.445, p = 0.195$  CFI = 1.00 RMSEA = 0.03

Fig. 21.20 Revised CFA model of comprehending strategy use

**Table 21.6** EFA and CFA factor loadings for the comprehending factor

Items	EFA	CFA
Item17	0.76	0.76
Item15	0.74	0.78
Item14	0.72	0.76
Item12	0.61	0.60
Item13	0.57	0.56
Item18	0.55	0.54
Item16	0.55	0.48
Item6	0.53	0.52
Item30	0.40	0.46

The total common factor variance ( $h^2$ ) for this model was 0.38 (i.e.,  $0.52^2 + 0.60^2 + 0.56^2 + 0.76^2 + 0.78^2 + 0.48^2 + 0.76^2 + 0.54^2 + 0.46^2$ ) $\div$ 9. This indicates that 38% of the factor variance was defined by the nine variables. The residual variance, which could be due to unique or specific sources of error, indicates the amount of variance not explained (Schumacker & Lomax, 2016). The unique (residual) factor variance ( $1 - h^2$ ) is 0.62. In practice, this CFA model will be considered together with other CFA models (based on the EFA results discussed in Chap. 20). Finally, Table 21.6 compares the factor loadings between the EFA (as presented in Table 20.14 in Chap. 20) and the CFA here. Although the statistical procedures used in EFA and CFA are different in the two methods, the factor loadings were found to be quite similar. In this chapter, it can be said that the current CFA model can be used to validate the EFA results presented in Chap. 20.

In practice, the next step is to test a CFA of the rest of the EFA factors and then subsequently to test whether a second-order CFA model may hold. However, it is beyond the scope of this chapter to illustrate all of these analyses. For an overview of higher-order CFA, see Brown (2015) and Phakiti (2008).

## Conclusion

CFA and SEM are robust analytical methodologies that allow researchers to perform analysis to investigate complex research questions and relationships. This chapter has offered some basic principles of CFA and SEM by describing how a CFA and SEM model is typically developed, evaluated and interpreted without referring to statistical formulas. This chapter has left out a number of CFA and SEM areas, such as ECFA (exploratory CFA), simultaneous multigroup analysis to test factorial invariance, boot-

strapping and jackknife, cross-validation approaches, growth trajectories over time through repeated measures and analysis of hierarchical data with multilevel SEM modeling, to name but a few (see Resources for Further Reading).

A SEM approach is far from perfect because certain goodness-of-fit indices, for example, are dependent upon sample size. We need to be aware that no one index serves as a definite criterion for testing a hypothesised CFA or SEM model. Furthermore, a perfect statistical model fit does not imply that a theory or hypothesis is proven. There may be other SEM models that could fit any given dataset (Kline, 2016). As Thompson (2000) stressed, the fit of a single tested model may always be an artefact of having tested too few models. It is also important to note that relationships among latent and observed variables cannot be interpreted as *causal*, although a directional arrow is used in CFA and SEM (see Cliff, 1983; Pearl, 2000). Theoretically, causality can only be inferred in cases in which a study is longitudinal or the data are from a strictly controlled experimental design.

## Resources for Further Reading

A number of resources here are oriented towards SEM, rather than CFA, because CFA is normally treated under SEM. Only Brown (2015) has comprehensively treated CFA.

SEM Journal: *Structural Equation Modeling: A Multidisciplinary Journal* is a refereed journal devoted to CFA and SEM.

The SEMNET FAQs website (frequently asked questions): <http://www2.gsu.edu/~mkteer/>

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York and London: The Guilford Press.

This book is the only comprehensive book on CFA available. It is highly recommended for those who would like to know more about CFA and the kinds of CFA analysis that may be applied in applied linguistics research (e.g., CFA of multitrait-multimethod matrices and CFA with equality constraints, multiple groups and mean structures).

Hancock, G. R., & Schoonen, R. (2015). Structural equation modelling: Possibilities for language learning researchers. *Language Learning*, 65(S1), 160–184.

This article discusses a range of possibilities to apply advanced SEM methods to L2 research (e.g., covariance structure models, measured variable path models, latent path models, multisample covariance structure models and latent growth models).

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York and London: The Guilford Press.

This fourth edition is more comprehensive than its previous edition and presents new topics in SEM analysis (e.g., Pearl's graphing theory and path models for longitudinal data).

Ockey, G. J. (2014). Exploratory factor analysis and structural equation modeling. In A. J. Kunnan (Ed.), *The Companion to language assessment* (pp. 1224–1444). Oxford: John Wiley & Sons.

This chapter presents an application of EFA and SEM in language assessment research. It provides a good overview of what is involved in both types of statistical analysis.

Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, 12(3), 305–319.

This article provides a comprehensive guideline of how to appropriately prepare a report of a study employing SEM, so that readers and journal reviewers can be assisted with the interpretation of the results established by the researcher.

Schoonen, R. (2015). Structural equation modelling in L2 research. In L. Plonsky (Ed), *Advancing quantitative methods in second language research* (pp. 213–242). New York: Routledge.

This chapter provides extensive discussion with published studies of essential SEM methodology, including a discussion of its advantages and caveats in L2 research. Schoonen illustrates a step-by-step SEM procedure through the use of LISREL (and AMOS).

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). New York and London: Routledge.

This fourth edition provides details of current developments and issues in SEM through an application of LISREL. It provides links among correlation, regression, path and confirmatory factor models to SEM accessible for people new to SEM and CFA.

Winke, P. (2014). Testing hypotheses about language learning using structural equation modeling. *Annual Review of Applied Linguistics*, 34, 102–122.

This article reviews research areas and published studies in second language acquisition research that apply SEM methodology. Winke also presents some of the main problems faced by SLA researchers when applying SEM.

## References

- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3), 317–332.
- Bentler, P. M. (1985–2018). *EQS Version 6 for Windows [Computer software]*. Encino, CA: Multivariate Software.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238–246.
- Bentler, P. M. (2006). *EQS structural equation program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588–606.
- Bentler, P. M., & Wu, E. J. C. (2006). *EQS for Windows user's guide*. Encino, CA: Multivariate Software.
- Bohrstedt, G. W., & Carter, T. M. (1971). Robustness in regression analysis. In H. L. Costner (Ed.), *Sociological methodology* (pp. 118–146). San Francisco: Jossey-Bass.
- Bozdogan, H. (1987). Model selection and Akaike's information criteria (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345–370.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York and London: Guilford Press.
- Byrne, B. M. (2006). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. New York and London: Psychology Press.
- Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, 18(1), 115–126.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
- Grant, R., MacDonald, R., Phakiti, A., & Cook, H. (2014). The importance of writing in mathematics: Quantitative analysis of U.S. English learners' academic language proficiency and mathematics achievement. In E. Stracke (Ed.), *Intersections:*



- Applied linguistics as a meeting place* (pp. 208–232). Newcastle upon Tyne). London: Cambridge Scholars Publishing.
- Hancock, G. R., & Schoonen, R. (2015). Structural equation modelling: Possibilities for language learning researchers. *Language Learning*, 65(S1), 160–184.
- In'nami, Y., & Koizumi, R. (2010). Can structural equation models in second language testing and learning research be successfully replicated? *International Journal of Testing*, 10, 262–273.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 57, 239–257.
- Kaplan, D. (1995). Statistical power in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 100–117). Thousand Oaks: SAGE.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York and London: The Guilford Press.
- Ockey, G. J. (2014). Exploratory factor analysis and structural equation modeling. In A. J. Kunnan (Ed.), *The Companion to language assessment* (pp. 1224–1444). Oxford: John Wiley & Sons.
- Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, 12(3), 305–319.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
- Pedhazur, E. J., & Schmelkin, L. P. (1992). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Holt, Rinehart and Winston.
- Phakiti, A. (2008). Strategic competence as a fourth-order factor model: A structural equation modeling approach. *Language Assessment Quarterly*, 5(1), 20–42.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35(4), 655–687.
- Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912.
- Rubinfeld, S., & Clément, R. (2012). Intercultural conflict and mediation: An inter-group perspective. *Language Learning*, 62(4), 1205–1230.
- Schoonen, R. (2015). Structural equation modelling in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 213–242). New York: Routledge.
- Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). New York and London: Routledge.
- Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261–283). Washington: American Psychological Association.
- Ullman, J. B. (1996). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (3rd ed., pp. 709–811). New York: Harper Collins College Publishers.

- Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning, 65*(2), 390–416.
- Weston, R., & Gore, P. A. (2006). A brief guide to structural equation modeling. *The Counselling Psychologist, 34*(5), 719–751.
- Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. (1977). Assessing reliability and stability in panel models. *Sociological Methodology, 8*, 84–136.
- Winke, P. (2014). Testing hypotheses about language learning using structural equation modeling. *Annual Review of Applied Linguistics, 34*, 102–122.



# 22

## Analyzing Group Differences

Luke Wander Amoroso

### Introduction

The relatively young field of quantitative second language (L2) research has historically had an inconsistent approach to the use and reporting of descriptive and inferential statistics. Most L2 research studies do not report meeting or even checking the underlying assumptions required by inferential statistical procedures (as revealed by Plonsky, 2013, 2014; Plonsky & Ghanbar, *in press*). Indeed, in L2 research studies generally, the quality of statistical choices and reporting has been suboptimal and has resulted in a sustained call for better and more consistent practices across the field (see e.g., Larson-Hall & Plonsky, 2015; Mackey & Goo, 2007; Marsden, Thompson, & Plonsky, *in press*; Norris, 2015; Norris & Ortega, 2000; Norris & Ortega, 2006; Plonsky, 2013, 2014; Plonsky & Gass, 2011; for examples of meta-analyses and critiques of SLA/L2 research methods).

This chapter focuses specifically on comparing group means and the statistical procedures used to assess group differences in L2 research contexts, with the primary emphasis being on analysis of variance (ANOVA) testing. Given the prevalence of this technique in L2 research, it is critical that we as a field understand and apply it appropriately. In addition to ANOVA, two other general linear model (GLM) procedures, multiple regression and *t*-tests, are examined, albeit briefly. The ANOVA family also includes analysis of covariance

---

L. W. Amoroso (✉)

Truman State University, Kirksville, MO, USA

e-mail: [lamoroso@truman.edu](mailto:lamoroso@truman.edu)

(ANCOVA), multivariate analysis of variance (MANOVA), and multivariate analysis of covariance (MANCOVA). See Larson-Hall (2016), Roever and Phakiti (2018), and Lomax and Hahs-Vaughn (2012) for detailed explanations of the different ANOVA procedures. Also, because L2 researchers often deal with small data sets containing non-normally distributed data, there is often a need to employ nonparametric statistical analyses, with Kruskal-Wallis, Welch, Brown-Forsythe, James, and bootstrapping procedures. See Lomax and Hahs-Vaughn (2012) and Hollander and Wolfe (1999) for a thorough and comprehensive explanation of nonparametric statistics, Turner (2014) and Roever and Phakiti (2018) for a book-length treatment of these tools and procedures in the context of L2 research.

The chapter maintains a focus on how to deal with small sample and group sizes, which are common to L2 research studies (Plonsky, Egbert, & LaFlair, 2015). This chapter is meant to be partly an explanation of the conceptual rationale for means-based analyses in the context of L2 research and partly an instruction manual for conducting such analyses. Prior to working through a hypothetical example, two very different applications of ANOVA in L2 research will be briefly noted.

## Two Examples of ANOVA in L2 Research

Larson-Hall's (2006) study of how L1 Japanese speakers' length of residence in an English-speaking country affected their English pronunciation is an established L2 research topic addressed with both parametric and nonparametric procedures.

### Sample Study 22.1

Larson-Hall, J. (2006). What does more time buy you? Another look at the effects of long-term residence on production accuracy of English /ɹ/ and /l/ by Japanese speakers. *Language and Speech*, 49(4), 521–548.

This early study by Larson-Hall clearly describes and justifies her statistical choices, checks assumptions, and interprets results clearly. Furthermore, she uses both parametric (repeated measures and two-way ANOVA, linear regression) and nonparametric (bootstrapping) statistics to analyze her data. She even subjects some of her data to parametric and nonparametric procedures and then compares the results of the two procedures for the reader, in order to highlight the fact that the choice of statistical test used to compare group means matters. This study is dense with statistics, but they are all central to comparing group differences and are clearly and thoroughly explained, and the area of research should prove familiar to many L2 researchers.

Unlike Larson-Hall's (2006) paper, which focused heavily on statistics, Oshita's (2014) study of how intransitive verbs are acquired by Japanese learners of English is notable for its application of simple ANOVA procedures to analyze an unusual data pattern found in multiple areas of linguistics.

### Sample Study 22.2

Oshita, H. (2014). U-shaped development: Definition, exploration, and falsifiable hypotheses. In L. W. Amoroso, & J. Connor-Linton (Eds.), *Measured language: Quantitative studies of acquisition, assessment, and variation* (pp. 155–170). Washington, DC: Georgetown University Press.

Oshita's study of interlanguage development hypothesized that L2 learners would exhibit U-shaped development in accurately judging whether or not English sentences were grammatically correct. Oshita's independent variable (language proficiency, or ability) had 172 learners distributed across five levels. The dependent variable was the degree to which a learner's grammaticality judgments were correct. Having a five-level independent variable, along with a reasonably large sample and a continuous dependent variable, the data appear to be suitable for regression analysis. However, a typical regression analysis would not allow for evaluation of a clearly nonlinear relationship such as a U-shaped pattern, as the procedure tries to fit the data to a straight line to model the relationship between the two main variables of interest. By using ANOVA, Oshita was able to show that there were clear (and statistically significant) differences between the five groups. Post-hoc testing was then used to show how grammaticality judgments changed as proficiency increased. As predicted, grammaticality judgments were better with the high and low proficiency groups, and not as good in the middle proficiency groups. Oshita's study is an example of how unusual (nonlinear) relationships can be fruitfully investigated with ANOVA techniques, and this is especially applicable to smaller data sets commonly found in L2 research.

Having provided the reader with examples of how ANOVA and other procedures have been used to analyze particular linguistic phenomena, we now turn to a general description of ANOVA and use it to work through a simple hypothetical research question: Do different teaching methods (we will call them Methods 1, 2, and 3 for brevity) affect L2 vocabulary test scores? The data for this hypothetical study can be found in Table 22.1.

## When to Use ANOVA

Before using an ANOVA, the question of whether or not you are interested solely in comparing multiple groups' performance on a common measure must be answered. If you are interested in making claims about whether or

**Table 22.1** Teaching method and vocabulary test score data

ID	Method	Score
1	1	56
2	1	79
3	1	83
4	1	82
5	1	65
6	1	90
7	1	95
8	1	75
9	1	88
10	1	78
11	2	89
12	2	37
13	2	65
14	2	89
15	2	70
16	2	65
17	2	69
18	2	82
19	2	55
20	2	70
21	3	83
22	3	82
23	3	85
24	3	79
25	3	80
26	3	81
27	3	82
28	3	90
29	3	75
30	3	82

not different teaching methods affect vocabulary test scores, then ANOVA may be a good choice, but if you want to know the relative contribution of a number of different variables to vocabulary test score differences then you will probably want to look outside of ANOVA.

A good candidate for ANOVA analyses would be the aforementioned scenario, where there is a single independent group variable (teaching method) and a single continuous dependent variable (vocabulary test score). Teaching method is clearly a categorical variable, with each group corresponding to a different teaching method. The dependent variable (vocabulary test score) is continuous, with possible scores ranging from zero to 100. In this case, the mean vocabulary test scores for each teaching method group can be compared against each other with an ANOVA. This initial step of matching the ANOVA procedure to an appropriate research question and data (comparison of group

mean scores, single categorical independent variable, single continuous dependent variable, an interest in whether and to what extent there may be differences between group mean scores) is essential before proceeding with ANOVA (or a  $t$ -test, if there are only two groups to compare).

### **ANOVA Assumptions: Independence of Observations**

Prior to conducting an ANOVA, three basic statistical assumptions must be checked and reported. The assumption of independent observations means that the vocabulary test scores must be independent of each other within a single teaching method group and also across groups. This assumption will not be violated if you keep each group from receiving more than one teaching method during the course of the study and if participants have only one test score each and do not collaborate on the test.

### **ANOVA Assumptions: Normally Distributed Data**

The second assumption is that the dependent variable is normally distributed within groups. In other words, each teaching method group should have normally distributed vocabulary test scores. This assumption is easiest to satisfy by having a large number of participants in each group. Checking this assumption can be done graphically, with histograms indicating normality of test score distribution for each group. The histogram should exhibit a bell curve for the data to be considered normally distributed. Normality can also be assessed statistically with a Kolmogorov-Smirnov or Shapiro-Wilk test. For both of these tests, if the  $p$ -value is greater than 0.05 then the data are assumed to be normally distributed. Like all null hypothesis tests, these methods of assessing normality are dependent upon sample size, so you may wish to present statistical test results and graphical evidence of normality (histograms) together.

### **ANOVA Assumptions: Homogeneity of Variance**

The third assumption that must be met for ANOVA is homogeneity of variance. In our example, this means that the amount of variation in test scores must be the same for each group. This assumption can be checked with Levene's test, with  $p$ -values greater than 0.05 indicating that variation is roughly equal across groups.

In addition to checking the ANOVA assumptions and reporting them, sample sizes, means, and standard deviations (SDs) must be reported for each of the groups. Reporting SDs allows the reader to see, at a glance, how similar variations in test scores are across groups. Reporting SDs alone is sufficient to check the homogeneity of variance assumption in many cases. Also, as recommended by researchers in other social sciences (Cumming, 2012; Harlow, Mulaik, & Steiger, 1997) and by L2 research methodologists (Larson-Hall & Plonsky, 2015; Norris, 2015; Norris, Plonsky, Ross, & Schoonen, 2015; Plonsky, 2015), reporting 95 percent confidence intervals (CIs) for each group mean is an excellent way to help readers understand the level of certainty to attach to your findings.

## Checking Assumptions: An Example

### Before Checking Assumptions

Table 22.2 provides descriptive statistics from the data.

CIs are an indication of how variable group means are. If the ten students who received *Teaching Method 2* averaged 69.1 points and had a 95 percent CI of 57.86 to 80.34, we could conclude that the mean score for a larger sample of students from the same population and receiving the same instructional method is likely to fall between 58 and 80 points. In other words, we should not have too much confidence that a mean vocabulary score of 69.1 is reflective of what the actual mean score from a much larger sample of participants receiving *Teaching Method 2* would be, as the 69.1 mean is simply one possible outcome in the 58 to 80 range. The small sample size and wide range of individual vocabulary test scores for *Teaching Method 2* contribute in this case to the relatively broad CI. On the other hand, the CI for *Teaching Method 3* test scores is much narrower, ranging from 79 to 85 points, with the mean being roughly 82. We can be slightly more confident that as we apply *Teaching Method 3* to different

**Table 22.2** Descriptive statistics for hypothetical ANOVA data

	Group size	Mean vocabulary test score	Standard deviation	95% CI
Teaching Method 1	10	79.1	11.67	70.75–87.45
Teaching Method 2	10	69.1	15.72	57.86–80.34
Teaching Method 3	10	81.9	3.9	79.11–84.69



samples within the same population, their mean scores on the vocabulary test will be relatively close to 82 points. In short, there is a lot of variation in the test scores of the *Teaching Method 2* group but less variation in the scores of the *Teaching Method 3* group. Given these variations in group test scores, we must check the normality of their distributions and then check that the variation within each group is similar enough to proceed with an ANOVA.

### Checking Assumptions: Normally Distributed Data

A review of the Kolmogorov-Smirnov (K-S) test for normality of the unstandardized residuals suggested that normality was a reasonable assumption for *Teaching Method 1* ( $D(10) = 0.163, p = 0.200$ ), *Teaching Method 2* ( $D(10) = 0.197, p = 0.200$ ), and *Teaching Method 3* ( $D(10) = 0.190, p = 0.200$ ).

While the K-S test indicated that distributions were normal, the reader should evaluate the K-S statistical test results in light of the very small sample size (10 participants) and SDs for each group. For example, the histogram for the test scores of *Teaching Method 2* shows a distribution that is not exactly shaped like a bell curve (as illustrated in Fig. 22.1).

It is important to provide a thorough description of the data so that the readers can understand how your data vary and the range of that variability. This is especially true with smaller sample sizes.

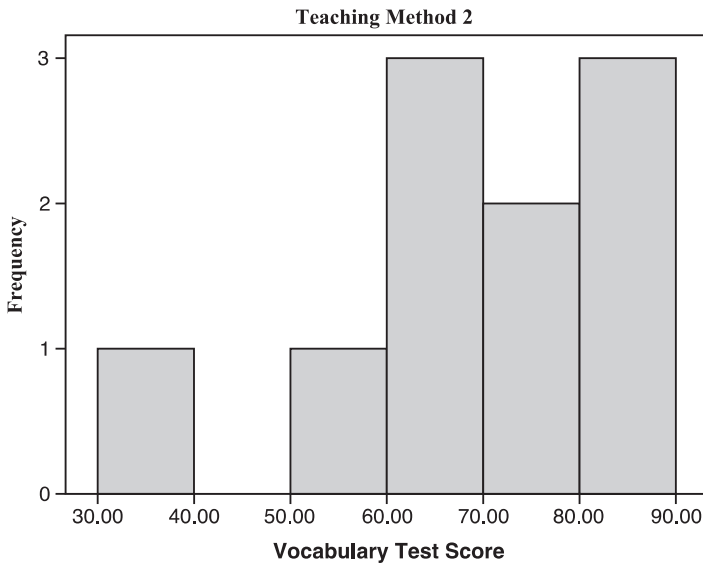


Fig. 22.1 Histogram of vocabulary test scores for Teaching Method 2

**Table 22.3** Standard deviations for each group of vocabulary test scores

	Group size	Mean vocabulary test score	SD
Teaching Method 1	10	79.1	11.67
Teaching Method 2	10	69.1	15.72
Teaching Method 3	10	81.9	3.9

## Checking Assumptions: Homogeneity of Variance

The third assumption that was checked was homogeneity of variance. Levene's test for equality of variances ( $F(2,27) = 3.122$ ,  $p = 0.060$ ) indicated that the assumption of equal variances across groups was satisfied. Although the result of Levene's test was larger than the  $p$ -value of 0.05, a glance at the SDs of the three groups shows that there was more variation in the test scores of *Teaching Method 1* and *Teaching Method 2* than in those of *Teaching Method 3* (Table 22.3).

Displaying descriptive statistics, in the form of histograms, SDs, and CIs, instead of displaying only the statistical test results associated with each of the ANOVA assumptions, allows readers to evaluate how much faith they want to place in the heterogeneity or normality of your group data. This is especially important when working with small sample sizes, as much L2 research has traditionally used group sizes below 20 (for aggregated sample and group sizes in L2 research, see Lyster & Saito, 2010; Mackey & Goo, 2007; Norris & Ortega, 2000; Plonsky, 2013; Plonsky & Gass, 2011).

Before using ANOVA procedures, you must check the aforementioned statistical assumptions. In the event that the ANOVA assumptions cannot be met, you may wish to proceed with descriptive statistics only. When presented and interpreted thoroughly, descriptive statistics, including means, SDs, CIs, and effect sizes (e.g.,  $d$ , eta-squared—see below) are often just as informative as the results of statistical tests, if not more so. Another option is to proceed with nonparametric statistical tests such as Kruskal-Wallis, Welch, or Brown-Forsythe tests, which do not carry the same assumptions as  $t$ -tests and ANOVA.

## Significance Testing with ANOVA

Having chosen ANOVA as the best fit for the research question and the data, the paradigm of null hypothesis significance testing (NHST) must be addressed. Are the questions of interest yes/no kinds of questions? Or, to situate NHST within our L2 vocabulary learning context, is the goal of the

research to show that groups with different teaching methods had different average test scores? Even if this is all that you want to accomplish, null hypothesis significance testing may still not be a good paradigm under which to proceed.

NHST answers a simple question: how confident can you be that groups behaved differently? The null hypothesis was that the groups would all behave identically. The groups were then tested and the mean test scores of the groups were found to be different. NHST will provide you with a measure of how probable it is that the groups are statistically indistinguishable from each other. The probability that the observed differences among three groups would be found if there was no true difference between them in an entire population of learners is expressed as a  $p$ -value. L2 researchers, as in many other fields, have been comfortable concluding that the difference between groups is statistically significant if  $p < 0.05$ . Although this threshold is arbitrary and controversial (Plonsky, 2015), it is generally accepted across the field (cf. Larson-Hall, 2016).

## A Note on Sample Size

Whether you choose ANOVA and  $t$ -test, or regression and correlation procedures (see Chap. XX), having a sample large enough to detect the effect that you are looking for is essential. Prior to conducting any research, it is recommended that you look at the published effect sizes for the types of effects that you are going to investigate. You can do so by examining meta-analyses or primary studies in the domain. You can then conduct a power analysis to determine how many participants you will need in each group to be able to state with confidence that the effect you were hoping to examine does in fact exist. For a concise overview of how to conduct a power analysis with freely available software, consult Plonsky (2015, p. 29).

## ANOVA Output and Interpretation of Results

Null hypothesis significance testing indicates with some degree of confidence (roughly 95 percent confidence) that the mean test scores in our hypothetical vocab study were not equivalent for all three teaching methods. Table 22.4 provides the relevant output from an ANOVA.

In this example, the conclusion is that the three groups of students (each group taught with a different method) did not all have the same average test

**Table 22.4** ANOVA output for mean differences in vocabulary test score

	Sum of squares	Degrees of freedom	Mean square	$F$	Sig.
Between groups	905.600	2	452.800	3.410	0.048
Within groups	3584.700	27	132.767		
Total	4490.300	29			

score. A typical reporting of ANOVA results is: There was a statistically significant effect of *Teaching Method* on *vocabulary test score*,  $F(2, 27) = 3.410$ ,  $p < 0.05$ . Another way to understand this result is to reject the null hypothesis that there are no differences in scores for the three conditions. Furthermore, even these conclusions must be qualified in terms of sample size and the variation within each group's test scores. As sample size increases and variation within a group's test scores decreases,  $p$ -values will drop (statistical significance will be reached). Conversely, smaller samples and greater within-group score variation (i.e., larger SDs) will lead to larger  $p$ -values (statistical significance will not be reached). In short, NHST of any kind really tells you whether or not you have collected enough data and whether or not the data are not too variable to be compared to each other, as statistically significant differences will always be found if the sample is large enough. The null hypothesis *can always be rejected* with a large enough sample.

An interpretive step that is often taken, and that is absolutely not licensed by null hypothesis testing, is that one method of teaching vocabulary is better than another. In fact, ANOVAs are not even directional. The omnibus  $F$ -test statistic that is output by an ANOVA merely indicates that there is a statistically significant difference somewhere among the mean test scores of at least two groups. With our hypothetical data, the ANOVA results indicate that there is a statistically significant difference in mean vocabulary test scores between the students of at least two different teaching methods. In order to determine which group performed better than another, post hoc tests can be used. Tukey's honestly significant difference (HSD) test will determine where the mean differences between the three groups lie. Tukey's test is similar to a  $t$ -test, in that it compares two groups at a time, but with the crucial difference that it controls for the family-wise error (i.e., the increased likelihood of a false positive when running multiple tests) that is present when there are more than two groups being compared. A post hoc Tukey HSD test showed that the *Teaching Method 2* and *Teaching Method 3* groups differed significantly at  $p = 0.05$  but that there was not a significant difference in mean vocabulary test scores between the *Teaching Method 1* and *Teaching Method 2*, or the *Teaching Method 1* and *Teaching Method 3* groups.

**Table 22.5** Tukey HSD post hoc comparisons of mean differences in vocabulary test score

(I) Teaching method	(J) Teaching method	Mean difference (I - J)	Standard error	Sig.
1	2	10.00000	5.153	0.147
	3	-2.80000	5.153	0.851
2	1	-10.00000	5.153	0.147
	3	-12.80000	5.153	0.050
3	1	2.80000	5.153	0.851
	2	12.80000	5.153	0.050

A second, complementary approach for understanding the direction and magnitude of differences among group scores involves simply going back to the descriptives. We can see from the mean test scores that the *Teaching Method 2* (69.1) group performed better than *Teaching Method 3* (81.9) group. In this way post hoc testing allows us to see which teaching method resulted in higher average vocabulary test scores (Table 22.5). Note that in this example the post hoc test indicated that there was precisely a 5 percent probability that the two means were not significantly different, despite the ANOVA  $F$ -test having a  $p$ -value of less than 0.05. This data set was designed to highlight the issues that arise when working with small samples and the importance of post hoc testing when dealing with such samples.

Having completed the ANOVA and post hoc tests, we can show that different teaching methods resulted in statistically significantly different test scores and that one method resulted in higher test scores than another method, but we still have no idea if those differences are meaningful. For example, while there is a statistically significant difference (if  $p \leq 0.05$ ) between the test scores of the *Teaching Method 2* group and the *Teaching Method 3* group, we do not know how important this difference is. We have reached the limits of null hypothesis testing with ANOVA at this point and must turn to measures of *effect size* to determine whether or not teaching method had any practical effect on vocabulary test scores.

## ANOVA and Effect Size

### Why Report Effect Sizes?

Some researchers will be content to have conducted an ANOVA and determined the direction of differences in test scores across the groups. It is a mistake, however, to conclude the analysis here. We have answered a very basic

question, namely, whether there are differences across conditions. We have not determined the magnitude of those differences and, practically speaking, the relative value of the different teaching methods. Moreover, if your goal is to determine *how effective* different teaching methods are, you must look beyond null hypothesis testing, to calculating, reporting, and interpreting effect sizes appropriate for ANOVA. Very few L2 research questions can be adequately addressed with NHST alone, and the remedy for overuse of NHST is reporting and interpreting effect size measures.

There have been a number of calls for reporting of effect sizes in L2 research in recent years (e.g. Norris, 2015; Plonsky & Gass, 2011; Plonsky & Oswald, 2014). This recent focus on effect size is a direct result of the lack of information provided by NHST. The field appears less willing to accept a statistically significant result as valuable in and of itself and has been moving toward *analysis of* statistical differences and their importance to L2 practitioners and researchers.

## ANOVA Effect Size Measures

Effect sizes, unlike  $p$ -values, are standardized measures of the extent to which variables are related to one another. In our example, effect sizes can describe the magnitude of the influence of teaching method on test scores. Effect sizes can be directly compared across studies and can also be used to indicate practical significance. For example, if differences in teaching method accounted for 85 percent of the variation in test scores, we could confidently say that teaching method will likely have a strong effect on test scores.

Appropriate effect size measures for ANOVA are eta-squared ( $\eta^2$ ) and, less commonly, epsilon-squared ( $\epsilon^2$ ).  $\eta^2$  provides a liberal estimate of the proportion of variation in test scores accounted for by differences in teaching method.  $\epsilon^2$  is an adjusted, and more conservative, form of  $\eta^2$ . (See Norouzian & Plonsky, 2018, for an examination and discussion of another related effect sizes,  $\eta_p^2$  (i.e., partial eta<sup>2</sup>), which is often confused with  $\eta^2$  and, consequently, misreported and/or misinterpreted).

SPSS will provide  $\eta^2$ , but the computation of  $\eta^2$  is simple and can be completed with a glance at the SPSS ANOVA output.

$$\eta^2 = \frac{\text{Sum of Squares Between Groups}}{\text{Sum of Squares Total}} = \frac{905.600}{4490.300} = 0.202$$

The computation of  $\varepsilon^2$  is also possible with SPSS ANOVA output but requires calculation of the *Mean Square Total* variable. *Mean Square Total = Sum of Squares Total / Degrees of Freedom Total*.

$$\varepsilon^2 = 1 - \left( \frac{\text{Mean Square Within Groups}}{\text{Mean Square Total}} \right) = 1 - \frac{132.767}{154.838} = 1 - 0.857 = 0.143$$

The  $\varepsilon^2$  value of 0.143 indicates that teaching method accounts for 14 percent of the variation in vocabulary test scores. As the values above illustrate, there can be large differences in the positively biased ( $\eta^2$ ) and unbiased ( $\varepsilon^2$ ) effect size measures when the groups are small (only ten participants per group in our example). Sample size matters, and as it increases, the two effect size measures will begin to approach parity. These two indices of association both range from 0 (no association) to 1 (perfect association). The latter would indicate, for example, that 100 percent of the variance in test scores between groups could be explained by teaching method alone. While we can be confident that an effect size of 1 is categorically significant and an effect size of 0 is not, how are we to interpret values between these two extremes?

## Interpreting ANOVA Effect Sizes

The  $\eta^2$  and  $\varepsilon^2$  values indicate that teaching method accounts for between 20 and 14 percent of the variance in vocabulary test scores, respectively. To interpret these effect size measures, we must look to the L2 research literature as a guide. There are no published guidelines for interpreting  $\eta^2$  and  $\varepsilon^2$  values in L2 research, but Plonsky and Oswald (2014) did make recommendations for interpreting  $r$  correlation coefficients. They recommended “that  $r$ s close to 0.25 be considered small, 0.40 medium, and 0.60 large” (Plonsky & Oswald, 2014, p. 889). When  $r$  correlation coefficients are squared, they also yield measures of the amount of variation in the dependent variable accounted for by the independent variable (see Norouzian & Plonsky, Chap. 19). So, by extending Plonsky and Oswald’s (2014) recommendation to squared correlation coefficients,  $r^2$  values of 0.06 should be considered small, 0.16 should be considered medium, and 0.36 should be considered large, generally speaking. These same guidelines for interpreting  $r$ , and by extension  $r^2$ , can be used to evaluate  $\eta^2$  and  $\varepsilon^2$  effect sizes for L2 research. By

applying Plonsky and Oswald's (2014) rules of thumb, we can conclude that the effect of teaching method on vocabulary test scores is a small one but nearing the threshold for a medium effect ( $\epsilon^2 = 0.143$ ). Using this indirect method of determining the magnitude of  $\eta^2$  and  $\epsilon^2$  effect size measures is justified because ANOVAs,  $t$ -tests, and multiple regression are all analyzing variance in the same way.

## ANOVA, Regression, $t$ -tests, and the GLM

Cohen (1968) clarified the relationship between ANOVA and multiple regression by asserting that the latter provided the same information as the former and that multiple regression could indeed be used in place of ANOVA in many instances.

### The $t$ -test as a Special Case of ANOVA

Another member of the GLM family is the  $t$ -test, which is an ANOVA with only two groups. If you are only interested in whether or not there is a mean difference between two groups, then a  $t$ -test is a good statistical choice. As with an ANOVA, there should be a single continuous dependent variable and a single categorical grouping variable with only two levels or groups. A  $t$ -test will tell you whether or not the means of the two groups are statistically significantly different from one another, reported as a  $p$ -value.

### Effect Size for $t$ -tests

As noted earlier, whether or not there is a statistically significant difference is not valuable information in and of itself; what is really of import is the magnitude of the treatment effect or difference between groups. The appropriate effect size measure when comparing two groups is Cohen's  $d$ , Glass's  $\delta$ , or Hedges's  $g$ . There are a number of free  $t$ -test effect size calculators online:

<https://www.uccs.edu/~lbecker/>

<http://www.socscistatistics.com/tests/Default.aspx>

Plonsky and Oswald have made data-supported recommendations for interpreting Cohen's  $d$  values in the L2 research context: "For mean differences



between groups,  $d$  values for comparisons of independent groups in the neighborhood of 0.40 should be considered small, 0.70 medium, and 1.00 large” (Plonsky & Oswald, 2014, p. 889). However, the authors also caution against overreliance on these rules of thumb, proposing a number of additional factors to consider when interpreting effect sizes.

Having examined the conceptual and procedural foundations of one-way ANOVA and  $t$ -test procedures, along with interpretation of their effect sizes, the discussion will now turn to appropriate uses and misuses of ANOVA.

## When Should I Use or Not Use ANOVA?

As mentioned earlier, one-way ANOVA is a particular type of regression. However, there are times when an ANOVA is not just a different way of conducting a multiple regression, but an incorrect procedure to use. Language is, by its nature, multivariate. The example that has been used throughout this chapter, how teaching methods affect vocabulary test scores, is a purposefully reductive one, created for illustrative purposes only. To claim that the 14 percent of variation in test scores attributed to teaching method is the result solely of the teaching method is very likely not accurate. What if intrinsic and extrinsic motivation had been measured in the 30 participants? What if months of L2 study had been recorded? What if gender had also been a variable? The point is that it is very difficult if not impossible to reduce a linguistic phenomenon to categorical variables affecting a continuous variable. Many factors ought to be examined as contributors to vocabulary test scores, as they may have a greater effect than teaching method.

As noted by Plonsky and Oswald (2017) in their article-length treatment of the subject, multiple regression is much more amenable to analyzing the profusion of measures, data types, and variables that exist and coexist in L2 research. In addition to accommodating continuous and categorical data, multiple regression allows for the influence of each variable to be seen *in concert with*, not *isolated from*, its fellow variables. Furthermore, interpretation of regression results, coupled with accessible effect size measures ( $R^2$ ), is far easier than attempting to explain the effects of the same variables in a factorial ANOVA or ANCOVA framework.

While there are times when ANOVA can and should be used, it is overused in L2 research. What follows is a description of three incorrect uses of ANOVA that have been noted by L2 researchers and possible alternatives for each.

## When ANOVA Is Inappropriate: Related Independent Variables

If there are multiple independent variables in a study (e.g., low vs. high proficiency, treatment vs. control, controlled vs. spontaneous production measure), then researchers will often run a series of single-variable (i.e., one-way) ANOVAs to assess the contribution of each of the independent variables to the dependent variable. However, when adding up the results of the multiple ANOVAs, there may be more explained variance than actual variance in the dependent variable. That is, the sum of the  $\eta^2$  values associated with each independent variable might exceed 1, which is neither intuitive nor possible. In contrast, a multiple regression distributes all of the variation in the dependent variable among the various independent variables at the same time; there is never more explained variance than actual variance. A more detailed explanation of this misuse of ANOVA can be found in Plonsky and Oswald (2017).

When dealing with a number of related independent variables, it is incorrect to conduct multiple univariate ANOVAs. The solution is to analyze all of the independent variables simultaneously with a multiple regression, taking care to conduct an a priori power analysis to determine how many participants will be needed to support such regression analyses.

## When ANOVA Is Inappropriate: Showing Group Equivalence

This misuse occurs when a researcher uses an ANOVA to show that the mean scores of groups are not statistically significantly different from one another. This argument is a restatement of a phenomenon identified and explained by Norris (2015).

The idea behind this use of ANOVA is that groups will change over time and that the nature of these differential changes from the same starting point will illuminate some aspect of L2 learning. Researchers, knowing that an ANOVA with an *F*-test *p*-value of less than 0.05 means that there are differences between the groups, extend this thinking to the belief that if an ANOVA *F*-test *p*-value is greater than 0.05 there are no differences between the groups. Cumming (2012) refers to this logical fallacy as the “slippery slope of non-significance” (p. 29).

An ANOVA *F*-test *p*-value of 0.04 indicates that there is a 4 percent probability that the groups are not different from one another. This probability is

small enough for us to comfortably assume that the groups are different from one another. However, an ANOVA  $F$ -test with a  $p$ -value of 0.17 indicates that there is a 17 percent probability that the groups are not different from one another. A less than one-in-five chance is not valid support for the equivalence of the groups. If an ANOVA is to be used to assess equivalence of group means prior to treatment, then the author should only accept ANOVA  $F$ -tests with  $p$ -values of greater than 0.95 as indicative of group equivalence. Though outside the scope of this chapter, I will simply note that a Bayesian approach to data analysis can allow for researchers to substantiate claims regarding group equivalence (Dienes, 2014; Norouziyan, de Miranda, & Plonsky, [in press](#); see also Lakens, [in press](#)).

### **When ANOVA Is Inappropriate: Forcing Continuous Data to Be Nominal or Ordinal**

This misuse of ANOVA results from researchers taking continuous data (e.g., years of study, composite scores on a motivation questionnaire, working memory test scores) and splitting them into “groups” for use in an ANOVA. This practice was noted by Plonsky and Oswald (2017), in their thorough discussion of the merits of multiple regression for L2 statistical analyses.

There are a number of problems with this approach to data. First, information is lost. Differences between individuals (i.e., variance) have been obliterated in order to use an ANOVA, which is a mistake, as the individual data points could be used in a regression to provide a detailed picture of how subtle changes in the continuous independent variable affect the dependent variable. The second issue with shoehorning continuous data into a few categories is that cut-points will have to be determined for those categories. What if a score of 78 gets put into the “high” group and a score of 77 gets put into the “low” group? You run the risk of not seeing real differences because of how you have arbitrarily grouped the participants. The rejoinder to this is “but I only use the people at the top and the bottom of the score range, so that my groups are clearly ‘high’ and ‘low.’” This is the third problem with this approach: You actively remove valuable data points from your analysis for the sole purpose of using an ANOVA. In other words, you limit your own potential understanding of the phenomenon you are studying.

The solution to this problem is obvious: Do not impose an artificial and misleading structure on your data. Instead, opt for a correlational or multiple regression analytical framework that can accommodate the data in their more

“natural” form and that has the additional benefit of providing a more complete picture of the phenomenon of interest.

## Conclusion

This chapter has attempted to describe some of the different methods that L2 researchers may use to analyze mean differences when comparing groups, and to help them choose the appropriate statistical framework for their data. ANOVA is the most prevalent statistical procedure used in L2 research today (Plonsky and Oswald, 2017), and it is overused, and even abused, by some researchers. This chapter argued for a judicious use of ANOVA and other parametric statistics that are based on an evaluation of data types and research questions. Furthermore, this chapter has attempted to provide enough guidance and support to allow researchers to conduct means-based comparisons of their own, through the use of hypothetical data sets and L2 variables.

While this chapter was unable to address many of the issues integral to ANOVA, *t*-test, or regression analyses (power analyses, sample size considerations, multivariate ANOVA, ANCOVA, logistic regression), I urge you to look into them (e.g., Larson-Hall, 2016; Lomax & Hahs-Vaughn, 2012) to help improve the methodological quality of L2 research and, more importantly, to put our field on firmer theoretical footing that will allow us to solidify and build upon our current knowledge of applied linguistics.

## Statistical Resources

All analyses in this chapter were conducted with the SPSS statistical package, but results obtained with other software will be the same or very similar. There are a number of freely accessible software packages that offer tools for conducting regression, ANOVA, and *t*-tests. Two such programs that are especially user-friendly are:

- <https://jasp-stats.org/>—a free interface that is quite similar to SPSS in appearance
- <https://www.jamovi.org/>—an easy-to-use statistical spreadsheet that provides clearly marked effect sizes

## Resources for Further Reading

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York, NY: Routledge.

This is an updated version of Larson-Hall's authoritative 2010 text and is one of the small but growing number of statistics texts written specifically for linguists. It covers a full range of statistical procedures (ANOVA, ANCOVA, regression, and others) and specifics of how to conduct analyses with either SPSS or R (free and open-source) software. The book takes the reader from the initial stages of creating variables and testing hypotheses through reporting of statistical results. Seven chapters, each covering a single statistical test, are structured around actual research studies (one study per chapter) from the diverse field of L2 research.

Lomax, R., & Hahs-Vaughn, D. L. (2012). *Statistical concepts: A second course* (4th ed.). New York: Routledge.

This comprehensive book-length treatment of ANOVA covers the entire family of related tests, detailing conceptual underpinnings, and worked examples (with data sets) of single- and multiple-factor ANOVA, ANCOVA, and regression procedures. Some nonparametric statistical tests are also explained in detail. While the book is not written specifically for the L2 researcher, it is written for social science researchers. The book provides SPSS screenshots and has helpful sections on power analysis and writing up statistical results.

## References

- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70(6), 426–443.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Quantitative Psychology* 402 and. *Measurement*, 5, 781.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

- Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.
- Lakens, D. (in press). Equivalence tests: A practical primer for *t*-tests, correlations, and meta-analyses. *Social Psychology and Personality Science*.
- Larson-Hall, J. (2006). What does more time buy you? Another look at the effects of long-term residence on production accuracy of English /ɹ/ and /l/ by Japanese speakers. *Language and Speech*, 49(4), 521–548.
- Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York: Routledge.
- Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(S1), 127–159.
- Lomax, R., & Haahs-Vaughn, D. L. (2012). *Statistical concepts: A second course* (4th ed.). New York: Routledge.
- Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, 32(2), 265–302.
- Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–449). Oxford: Oxford University Press.
- Marsden, E., Thompson, S., & Plonsky, L. (in press). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*.
- Norouzian, R., de Miranda, M. A., & Plonsky, L. (in press). The Bayesian revolution in L2 research: An applied approach. *Language Learning*.
- Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, 34, 257–271.
- Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, 65(S1), 97–126.
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528.
- Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3–50). Philadelphia: John Benjamins.
- Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, 65(2), 470–476.
- Oshita, H. (2014). U-shaped development: Definition, exploration, and falsifiable hypotheses. In L. W. Amoroso & J. Connor-Linton (Eds.), *Measured Language: Quantitative studies of acquisition, assessment, and variation* (pp. 155–170). Washington, DC: Georgetown University Press.
- Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition*, 35(4), 655–687.

- Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, 98(1), 450–470.
- Plonsky, L. (2015). Statistical power, p values, descriptive statistics, and effect sizes: A “back-to-basics” approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–45). New York: Routledge.
- Plonsky, L., Egbert, J., & LaFlair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, 36(5), 591–610.
- Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, 61(2), 325–366.
- Plonsky, L., & Ghanbar, H. (in press). Multiple regression in L2 research: A methodological synthesis and guide to interpreting  $R^2$  values. *Modern Language Journal*.
- Plonsky, L., & Oswald, F. L. (2014). How big is ‘big’? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912.
- Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, 39, 579–592.
- Roever, C., & Phakiti, A. (2018). *Quantitative methods for second language research: A problem-solving approach*. New York: Routledge.
- Turner, J. L. (2014). *Using statistics in small-scale language education research: Focus on non-parametric data*. New York, NY: Routledge.



# 23

## Statistics for Categorical, Nonparametric, and Distribution-Free Data

Jesse Egbert and Geoffrey T. LaFlair

### Introduction

This chapter deals with data that do not look like the clean, well-behaved example data sets that are often used in introductory statistics courses. In other words, this chapter is about nontraditional data. It is crucial for researchers to make a distinction between bad data and nontraditional data. Bad data are data that were collected using poor sampling practices, unreliable/invalid instruments, or careless data entry and management practices. Nontraditional data, on the other hand, are data that cannot be analyzed using statistical models that are familiar or widely used in applied linguistics (AL). Note that the notion of nontraditional data is relative; data labeled as *nontraditional* in one research community such as AL or a subset thereof may be the default data type in another. If a particular data set was appropriately sampled using reliable and valid instruments and careful data entry and management practices, then it is likely that there is nothing inherently *bad* about the data at all. Nevertheless, data deemed to be nontraditional within a research community are often stigmatized. This stigma can lead researchers to use questionable practices, such as eliminating outliers or transforming data without adequate justification, in an

---

J. Egbert (✉)

Northern Arizona University, Flagstaff, AZ, USA

e-mail: [Jesse.Egbert@nau.edu](mailto:Jesse.Egbert@nau.edu)

G. T. LaFlair

University of Hawai'i at Mānoa, Honolulu, HI, USA

e-mail: [glafair@hawaii.edu](mailto:glafair@hawaii.edu)



effort to fit their data into a traditional mold. Or, worse yet, this stigma could even lead researchers to use statistical techniques that are inappropriate for their data. Rather than trying to change their data or analyze them with traditional, but inappropriate statistics, researchers should focus on trying to understand nontraditional data and find appropriate methods for analyzing them.

Thus, the primary objective of this chapter is to help applied linguists understand and analyze nontraditional data. Specifically, this chapter is dedicated to three particular types of nontraditional data:

- categorical data
- non-normal data
- data with small sample sizes

In this section, we briefly introduce each of these three data scenarios, as well as their prevalence in the field of AL, and outline the remainder of the chapter.

Despite being quite prevalent in AL research, researchers are often unsure how to measure statistical patterns in categorical data. This could be due, at least in part, to a lack of emphasis on nominal data in courses on research methods and statistics. It could also be related to the strong bias in the AL literature in favor of measurement data and the use of parametric statistical techniques such as *t*-tests and ANOVAs (see, e.g., Plonsky & Gass, 2011; Plonsky, 2013). Regardless of the reason, categorical data are often viewed as nontraditional data and misunderstood in AL research. In section [Analyzing Categorical Data](#), we introduce different types of categorical data. We then give an overview of different statistical techniques for modeling categorical data to answer a variety of research questions about frequency-based data.

Another type of data often viewed as nontraditional in the AL community is data that do not follow a normal (Gaussian) distribution. The irony is that non-normal data are evidently quite prevalent in AL research. Plonsky, Egbert, and LaFlair (2015) reanalyzed 26 data sets from studies previously published in major AL journals and found that the samples in 18 (69%) of the 26 studies were non-normally distributed. Other studies have shown that much of the time statistical assumptions, such as normality, are either not checked or not reported. Plonsky (2013) reported that authors checked statistical assumptions, such as normality, and reported the results in only 101 (17%) out of 606 studies (see similar results in Ghanbar & Plonsky, [under review](#)). As mentioned above, *t*-tests and ANOVAs together constitute the overwhelming majority of all statistical tests performed in AL. However, both of these techniques assume that sample data are normally distributed. There is nothing inherently wrong with data that are non-normally distributed, but parametric statistical tests are inappropriate

for this type of data. Section [Nonparametric Tests](#) offers alternatives to traditional parametric statistics that can be applied to data that are non-normal.

The final type of nontraditional data is that of small sample sizes. Many parametric statistical tests are not appropriate for small sample sizes. As with non-normal data, data sets with small sample sizes are often stigmatized in the AL research community despite their prevalence. Plonsky and Gass (2011) found an average sample of only 22 participants in a methodological synthesis of 174 interactionist second language acquisition (SLA) studies, and Plonsky (2013) found an average group  $N$  size of 19 in his research on study quality in 606 primary L2 studies. There are a number of factors that contribute to small sample sizes in AL research. These include practical data collection constraints like limited classroom sizes and participant attrition in longitudinal studies, as well as research decisions such as subdividing overall participant samples into several small groups. Small sample sizes in AL research may at times result from a lack of emphasis on or awareness of the debilitating effects of low statistical power. However, it is quite clear that in many cases researchers are genuinely limited in their ability to collect large data sets. Section [Nonparametric Tests](#) addresses statistical methods for dealing with small samples or samples that are not normal.

## Analyzing Categorical Data

Categorical data—sometimes referred to as frequency data, nominal data, or data with nonquantitative outcomes—consist of simple frequencies that represent how many observations fall into each of two or more categories. Categorical data are distinguished from measurement data, in which each of  $N$  observations contributes a score along a scale, making it possible to compute measures of central tendency (e.g., mean, median) and dispersion (e.g., standard deviation). Measurement data can also be described in terms of its underlying distribution. In contrast, measures of central tendency and dispersion cannot be computed for categorical data, which can only be measured in terms of frequency counts in each category.

Common examples of categorical data in AL studies include variables such as sex, race, L1/L2 background, education level, and other demographic variables. Another example of categorical data in AL that is becoming increasingly common is frequency data from language corpora. Corpus data are often measured using frequencies or rates of occurrence of a particular linguistic feature. While it is possible to collect measurement data from corpora, this can only be done if rates of occurrence for linguistic features are calculated for each of many texts in a corpus (as opposed to a single rate for the entire corpus) (see Biber & Jones, 2009).

## Simple Frequencies

A common approach to analyzing categorical data is to simply report frequency data using tables and figures such as bar plots and pie charts. Say, for example, we are interested in comparing the frequency of the word *said* between online news reports and other online registers. We could use the Corpus of Online Registers of English (CORE) (<http://corpus.byu.edu/core/>) to make this comparison. The results of this query are displayed in Table 23.1. Because the two sub-corpora are of different sizes, it is not fair to use raw counts to compare linguistic features. Therefore, we have also provided normalized rates of occurrence for *said* in both sub-corpora.

We can also display these normalized frequencies in the form of a bar plot (see Fig. 23.1). We can see from these results that online news reports use *said* more than other online registers.

## Measuring Frequency Differences

In many cases, researchers are satisfied with reporting simple frequency information for their categorical variables. However, we might look at a frequency difference between two categories and wonder whether a difference is big enough to be statistically significant. In other words, we might be interested in estimating the probability that a particular difference occurred due to random chance alone. As we have already discussed, the categorical nature of our data renders it impossible to answer this type of question using a means-based statistic, such as an independent-samples *t*-test. Fortunately, there are statistical tests designed to answer this question with frequency data. This section introduces three common tests used to measure statistical differences in categorical data, including a discussion of their strengths and limitations. We conclude this section with a brief introduction to two measures of effect size for measuring differences in categorical data.

The statistical tests introduced in this section—chi-square test, likelihood ratio test, and Fisher’s exact test—draw on frequencies from a contingency table to estimate the likelihood that a particular difference occurred by chance. A contingency table contains the levels of the first variable as the columns and

**Table 23.1** Frequency of *said* in online news and other CORE registers

	News reports	Other online registers
Raw frequency of <i>said</i>	31,333	45,873
Normed <sup>a</sup> frequency of <i>said</i>	3.69	2.04

<sup>a</sup>Per 1000 words

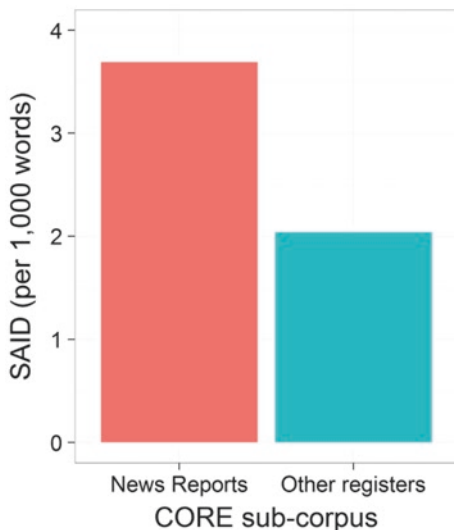


Fig. 23.1 Bar plot displaying normed frequency of *said* in news and other CORE registers

Table 23.2 Contingency table of frequencies for *said* in the CORE

	News	Other registers	TOTAL
Frequency of <i>said</i>	31,333	45,873	77,206
Frequency of other words	8,460,163	44,396,174	52,856,337
TOTAL	8,491,496	44,442,047	52,933,543

the levels of the second variable as the rows, typically with a column on the far right and row on the bottom for totals. The corpus data for the use of *said* in the CORE is displayed as a contingency table in Table 23.2.

The most common test for measuring statistical differences in categorical data is Pearson’s chi-square test. The chi-square test is an inferential statistic that must be calculated using raw frequencies rather than ratios or percentages. According to Gries (2013), the chi-square test has the following three assumptions:

- 80% of the expected frequencies are greater than 4;
- All expected frequencies are greater than 1;
- All observations are independent of each other.

In addition, sample size is an important consideration in chi-square tests. While suggested cutoff points vary, researchers should be aware that the risk of committing a Type II error (false negative) is higher for smaller sample

sizes. It should also be noted that  $p$ -values generated from chi-square tests are extremely sensitive to sample size, in general, producing small  $p$ -values even for small differences, given a large enough sample size. Chi-square has a number of benefits: it is easy to compute, it can handle contingency tables larger than  $2 \times 2$ , and it can be performed in most statistical software packages, including SPSS and R. Chi-square tests can also be performed using freely available online calculators (e.g., <http://graphpad.com/quickcalcs/chisquared1.cfm>; <http://langtest.jp/shiny/chi/>).

Another option for analyzing differences between frequency counts in categorical data is a likelihood ratio test. The family of likelihood ratio tests includes many formulas that have been proposed. One of the most commonly used is the G-test. Like chi-square, likelihood ratio tests can be computed relatively easily for contingency tables larger than  $2 \times 2$ . With reasonable sample sizes, likelihood ratio tests will give the same result as chi-square tests, and likelihood ratio tests have similar assumptions to chi-square. However, likelihood ratio tests are recommended over chi-squared when expected values are small or when sample sizes are very large. Keyword analysis, a corpus linguistic method of identifying words that occur much more in one corpus than another corpus, is typically performed using likelihood ratio tests. Sample Study 23.1 summarizes a corpus-based study that used keyword analysis, based on likelihood ratio tests. Likelihood ratio tests are supported by many statistical software packages, and they can also be computed using online calculators (e.g., <http://graphpad.com/quickcalcs/contingency1.cfm>; <http://ucrel.lancs.ac.uk/llwizard.html>).

### Sample Study 23.1

McEnery, T. (2016). Keywords. In P. Baker & J. Egbert (Eds.), *Triangulating methodological approaches in corpus-linguistic research*. New York: Routledge.

#### Research Aims

The objective of this study is to compare the words used in online question + answer (Q+A) forums between four world English varieties: the United States, the United Kingdom, India, and the Philippines. For each variety, the researcher generates a list of "keywords," or words that are statistically more frequent in that variety when compared with the other three.

#### Research Method

##### Data:

- Q+A corpus
- 265 texts, each including a question and a series of answers (from Yahoo! Answers)
- 406,407 words
- Balanced across four world English varieties (UK, US, IN, PH)

*Instruments/techniques*

- Keyword analysis, using WordSmith Tools
- Log-likelihood tests to compare the frequency of each word between:
  - The target corpus (one of the four varieties)
  - The reference corpus (the three other varieties, combined)
- Keywords:  $p < 0.001$

**Key Results***The UK*

- 46 keywords
- More politeness words (e.g., *sorry*), often preceding a disagreement
- Advice giving

*The US*

- 57 keywords
- Words related to entitlement and rights

*IN*

- 83 keywords
- Moral values of society and cultural
- Religious words
- Most acronyms (e.g., *r*, *u*, *ur*)

*PH*

- 83 keywords
- Conservative social values
- Religious words
- Most code-switching

**Comments**

This study demonstrates the potential usefulness of statistics designed to measure differences in categorical data. The use of log-likelihood in this study identified lists of keywords that are strongly associated with each of four world English varieties. Most of these keywords could be meaningfully interpreted by the author.

An alternative to chi-square and likelihood ratio tests that has become widely accepted in many disciplines is Fisher's exact test. Like chi-square, Fisher's exact test is used to analyze contingency tables of frequency data. Although it is most often suggested when sample sizes are small or when the frequencies in contingency table cells are very unbalanced, Fisher's exact test is a valid statistical test for samples of any size. In fact, some researchers suggest always using Fisher's exact test over chi-square and G-tests (McDonald, 2014, pp. 86–89). However, computation becomes much more difficult with larger sample sizes. One of the strengths of Fisher's exact test is that it allows for the computation of an exact  $p$ -value, rather than an approximation. This is possible because Fisher's exact test is a permutation test that accounts for all possible contingency tables of the sample data in order to measure the actual probability of seeing results as

extreme or more extreme than the difference found in the sample data. We discuss other examples of permutation tests in section [Permutation Test and Bootstrapping](#). One limitation of Fisher's exact test is that it requires a substantial amount of computational processing to calculate, and some statistical packages only support the computation of Fisher's exact on 2x2 contingency tables. Online calculators are available to compute Fisher's exact tests on 2x2 contingency tables with relatively small frequencies (e.g., <http://graphpad.com/quickcalcs/contingency1.cfm>).

In conjunction with these three statistics for measuring frequency differences, researchers can compute effect sizes which measure the strength of the association between two categorical variables. The Phi coefficient is perhaps the best effect size measure to use when the variables in an analysis only include two levels. When one of the variables contains more than two levels, the correct measure is Cramer's V. Both of these effect size measures are based on the chi-square statistic. Their values are bound between  $-1$  and  $1$ , where  $-1$  is a perfect negative association and  $1$  is a perfect positive association. In many cases, however, simple percentages can also be used as a straightforward index of the relationships being investigated.

## Other Statistics for Categorical Data

The main focus of this section has been the analysis of univariate categorical data, with a focus on measures of statistical difference and association. There are, however, many other statistical techniques that have been developed to answer a wide range of other questions that may be answered with categorical data. For example, it is possible to include categorical predictor variables in regression analyses (see Norouzian & Plonsky, Chap. 19; Plonsky & Oswald, 2017). Dichotomous categorical variables can be directly entered into a regression model. Categorical variables with three or more levels can also be included in regression analyses through dummy coding. Dummy coding is the process of creating a new independent variable for each level of a categorical variable, and coding each of the new variables as a dichotomous variable (for a demonstration, see Amoroso, Chap. 22). Regression is also possible with categorical dependent variables using logistic regression. There are also categorical analogs for many multivariate statistics, such as multiple correspondence analysis which is similar to principal component analysis (see Phakiti, Chap. 20), but for categorical data.

## Nonparametric Tests

Researchers are often interested in answering questions about the variability of parameters, the similarity of distributions, or the extent to which their data satisfy the assumptions for parametric tests. However, as mentioned in the [introduction](#), these questions cannot be answered using traditional parametric statistical techniques when dealing with data with extremely small sample sizes or severely non-normal distribution. In such cases, researchers can turn to nonparametric tests. This section discusses nonparametric analyses that are appropriate for non-normal data and small sample sizes. First, permutation tests and bootstrapping are discussed. This is followed by a discussion of statistical tests for nonparametric data and for comparisons of distributions.

### Permutation Test and Bootstrapping

It is common for researchers in AL to work with participant groups for which it is difficult to collect large sample sizes. In extreme cases, this makes hypothesis testing inappropriate because it cannot be assumed that important assumptions (random sampling, random assignment, the data come from a normal distribution) have been met (Davison & Hinkley, 1997). However, if the researchers are confident that their small samples are representative of the population that they intend to generalize to, bootstrapping and the permutation test (or randomization test) could provide important insights into the data beyond descriptive statistics and nonparametric tests. Both analyses use resampling methods to build a hypothetical population distribution. The hypothetical population distributions provide more robust information about the shape of the data and about the answer to the research question under investigation (e.g., mean differences, correlation, and regression) than might be gained by examining only the sampling (i.e., observed) distribution.

Permutation tests rely on resampling, and reassigning to groups, data points from the sample without replacement to make  $R$  reconfigured unique permutations of the data. With each resampled permutation of the data, the test statistic of interest is calculated. In other words, a permutation test calculates a test statistic for all of the possible arrangements of the data. Then the original test statistic is compared to the distribution of resampled test statistics. The original test statistic is considered *meaningful* if it is more extreme than

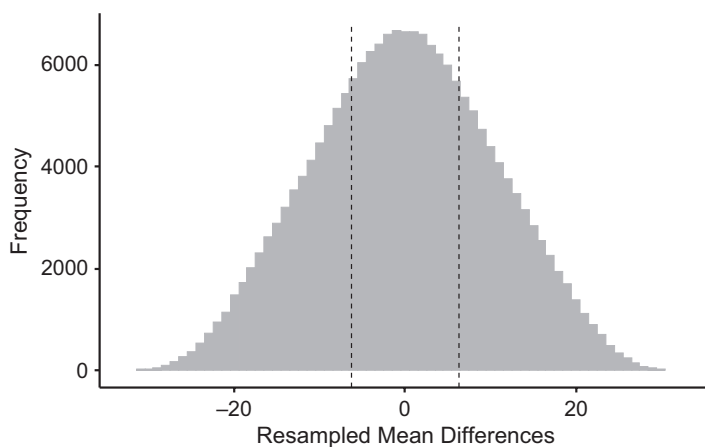


95% of the resampled statistics (Keuhl, 2000). Bootstrapping is similar to permutation tests except that the  $R$  resampled data sets may not be unique, and bootstrapping is typically used to estimate the variability in the bootstrap sample distribution by estimating confidence intervals around the parameter of interest.

For example, Donaldson (2011) compared adult near-native learners of French and native speakers of French on their usage of left dislocation in French—a phenomenon in which a well-formed utterance is preceded by a redundant constituent. This study used an independent-samples  $t$ -test to test the hypothesis that native speakers of French use more left dislocation than non-native speakers ( $d' = 6.3$ ). Donaldson's original  $t$ -test yielded the following result:  $t(18) = 0.546$ ,  $p = 0.598$ , which clearly did not reject the null hypothesis. While  $t$ -tests are robust estimates of mean differences, this design also would be appropriate for a permutation test or a bootstrap analysis because there are questions about the three assumptions that underpin parametric tests: (1) that the data are randomly sampled, (2) that the participants are randomly assigned to groups, and (3) that the data come from a normal distribution.

If a permutation test was used to address the three assumptions, the data from Donaldson (2011), which were provided in the article and which we use for the basis of this reanalysis, would be used to create  $R$  (e.g.,  $R = 10,000$ ) unique possible orderings of the data in the original data set. For each data set, the mean difference among the two groups is calculated. Finally, the original mean difference is compared to the other 10,000 mean differences. For a two-tailed hypothesis, if the original mean difference is more extreme than 95% of the resampled mean differences (i.e.,  $2.5 > d'$  or  $d' > 97.5$ ), then the null hypothesis is rejected. In conducting a permutation test, this can be determined in one of two ways. First, the distribution of the resampled means can be seen in Fig. 23.2. The dotted lines represent the original mean difference in both the lower and upper tails. Under the two-sided null hypothesis, native speaker status has no effect on the number of left dislocations used by a speaker ( $H_0 : d' = 0$ ). We fail to reject this hypothesis because 56% of the resampled mean differences are as extreme or more extreme than the observed mean difference. Second, it could be examined using the Fisher's exact test as discussed above.

A nonparametric bootstrap analysis of this same data set would differ from the permutation test because the  $R$  replications of the data set may not be unique orderings, and the bootstrap analysis could preserve the original



**Fig. 23.2** Resampled mean differences based on the data from Donaldson (2011)

grouping structure of the data during resampling if required (LaFlair, Egbert, & Plonsky, 2015). In other words, unlike the permutation test, which accounts for every possible unique ordering of the data exactly one time, bootstrapping resamples with replacement from the sample, allowing for the data points to be included more than once in the bootstrapped sample. Additionally, the bootstrap analysis could examine the null hypothesis by comparing the original mean to the bootstrap mean as was done in the permutation test above, or the hypothesis could be examined by creating bootstrap confidence intervals. There are a number of choices of nonparametric bootstrap confidence intervals (e.g., studentized, percentile, or BCa). The bias-corrected and accelerated (BCa) method is an adjusted percentile method which accounts for bias and skewness in the distribution of the bootstrap estimates (Efron, 1987; LaFlair et al., 2015), and it is generally recommended for use. A nonparametric bootstrap analysis of the Donaldson (2011) data revealed that 59% of the bootstrap mean differences were as extreme as, or more extreme than, the original mean difference. Additionally, the BCa 95% CI  $[-10.89, 31.68]$  clearly contains zero, thereby corroborating the permutation test and the original  $t$ -test. Sample Study 23.2 summarizes the incorporated bootstrap analyses to investigate the effects of task repetition on L2 learners' attention to the fluency, complexity, accuracy, and lexical variety of their oral responses to the tasks.

### Sample Study 23.2

Fukuta, J. (2016). Effects of task repetition on learners' attention orientation in L2 oral production. *Language Teaching Research*, 20(3), 321–340.

#### Research Aims

The purpose of this study is to examine the repetition of oral tasks that affects what aspects of language learners pay attention to: fluency, complexity, accuracy, lexical variety.

#### Research Method

##### *Participants, tasks, and procedures*

- 28 EFL college students at an upper-intermediate proficiency level.
- Random assignment to experimental and control groups.
- Experimental group responded to the same picture narration task twice (one week apart).
- Comparison group responded to two different picture narration tasks twice (one week apart).
- Groups and tasks were counterbalanced.
- Stimulated retrospective interviews were carried out after each production to triangulate participant attention to measures and their scores on the measures.

##### *Measures*

- Complexity: number of clauses per analysis of speech (AS) unit
- Accuracy: percentage of error-free AS units
- Fluency: number of words/minute excluding repairs and repetition
- Lexical variety: number of word types of the square root of tokens (Guiraud's Index)

##### *Analysis*

- Bootstrap means and confidence intervals were calculated to examine the extent to which they would support the results of the parametric tests of mean differences.

#### Key Results

The bootstrap analysis corroborated the results of the parametric tests and supported the inference that participants in the experimental group focused more on syntactic aspects of performance than on conceptual aspects when completing the task for the second time. This result was not seen in the comparison group because the two tasks they completed contained different pictures and thus each task required them to focus on the conceptual aspects equally (also supported by the bootstrap analysis).

#### Comment

This study illustrates a potential use of bootstrapping: a method for evaluating the results of parametric tests in experiments with small samples. The calculation of the bootstrap confidence intervals allowed the author insight into the accuracy of the estimated parameters (i.e., mean scores on the measures).

## Nonparametric Hypothesis Tests and Distribution Tests

Frequently, applied linguists are trying to answer a question about differences between the central tendencies of one or more groups on a measure administered at one or more times. This might be the case when researchers are comparing test scores between two or more groups. Researchers may also be interested in the extent to which the distributions of more than one set of data are similar as in studies focused on the linear relationship between two continuous variables (e.g., age and test scores). These types of questions can be answered by nonparametric hypothesis tests and distribution tests.

Nonparametric hypothesis tests are often explained as substitutes for parametric tests when the data do not satisfy parametric assumptions. The nonparametric counterparts examine differences in medians, signs, and ranks of the data from one or more groups. The assumptions for these tests are much fewer and much less stringent than their parametric counterparts. They all assume independent random samples and have fewer assumptions about the distribution of the data. For example, the Wilcoxon rank-sum test (or Mann-Whitney U), a nonparametric parallel to an independent-samples  $t$ -test, and the Kruskal-Wallis test, the substitute for ANOVAs, assume that the population distributions of the two groups are the same shape but differ in their location. The Wilcoxon signed-rank test, the corollary to a paired-samples  $t$ -test, assumes that the distribution of differences is symmetrical about the population median (Ott & Longnecker, 2010). In addition to less stringent assumptions, the null hypothesis for each of these tests is different from their parametric counterparts.

The Wilcoxon signed-rank test examines the null hypothesis that the difference in medians between two groups is zero. The Wilcoxon rank-sum test examines whether or not the populations are identical ( $\Delta > 0$ ) which contrasts with a test of mean differences in an independent-samples  $t$ -test in that the alternative hypotheses question the shift, and its magnitude, of one distribution in respect to the other. From this null hypothesis, it is clear that the Wilcoxon rank sum, although a corollary for the independent-samples  $t$ -test, does not directly examine the difference between the means of two groups. However, it does directly examine whether or not the distributions are the same. Furthermore, the rank-sum test can be applied to a broader range of two-group comparisons as it does not require the assumption of normality and thus may be better suited for small-sample data sets (Ott & Longnecker, 2010).

This is also true for the Kruskal-Wallis test when examining  $K$  samples instead of two samples. If the shapes of the distributions are the same, as shown by QQ plots, for example, then inferences about mean or median differences can be made (Thas, 2010). There are two other tests that can be employed in order to answer the question about the similarities of two or more continuous distributions. The Kolmogorov-Smirnov (KS) test and the Anderson-Darling (AD) test examine two distributions in regard to the location and the shape of their empirical distribution functions. These two statistics test the null hypothesis that the populations that the samples were drawn from are identical. The strength of these tests is that they include both location (e.g., mean) and shape (e.g., variation) in the statistical test without requiring the assumption of normality that is required by the  $t$ -test. The KS test may be useful for researchers who are interested in examining whether or not two groups are at the same proficiency level before an experiment begins. Traditionally, this has been addressed by a  $t$ -test where failure to reject the null hypothesis results in the conclusion that the two groups are the same. However, the  $t$ -test will only indicate if there is, or is not, a significant difference in mean scores. In other words, it is sensitive to differences between the two groups in location. The KS and AD tests may indicate that two distributions are different due to differences in location and/or differences in shape. For example, the two groups under comparison may have similar location (e.g., mean scores on a pretest), but they may be shaped differently (e.g., one group has larger variance, or is bimodal). The KS test would be sensitive to the difference in shape in addition to any differences in location.

## Conclusion

In this chapter, we have introduced statistical techniques for dealing with three types of nontraditional data: categorical data, data with small sample sizes, and non-normally distributed data. We began the chapter by establishing each of these three data types as common challenges in AL data. In section [Analyzing Categorical Data](#), we introduced three statistical techniques and two effect size measures for modeling categorical data. In section [Nonparametric Tests](#), we moved to permutation tests, bootstrapping, and nonparametric tests for analyzing measurement data with small sample sizes or non-normal distributions. Table 23.3 contains a summary of these measures and techniques, including assumptions, null hypothesis, when to use them, and which software to use.

Table 23.3 Summary of key information for the analyses included in this chapter

Analysis	Assumptions/criteria	Null hypotheses	When to use	Software
Chi-square	Categorical data Independent obs. Exp. freq. > 1	Expected result is equal to observed result	Samples are not too small or too large	R: chisq.test; SPSS; online
Likelihood ratio	Categorical data Independent obs.	Expected result is equal to observed result	Small exp. freq. Large samples	R: lrtest, online
Fisher's exact	Categorical data Independent obs.	Expected result is equal to observed result	Small samples Unbalanced freq.	R: fisher.test; online
Permutation test	Valid representation of the population, independent observations	The original test statistic is not more extreme than (1- $\alpha$ ) % of the permuted test statistics	Small sample sizes, non-normal data, lack of random sampling and assignment	R: boot, bootstrap, broom, car; SPSS
Bootstrapping	Valid representation of the population, independent observations	Depends on what statistic is being resampled	Small sample sizes, non-normal data, lack of random sampling and assignment	R: boot, bootstrap, broom; SPSS
Wilcoxon signed-rank test	Independent samples/paired data	$H_0: M_1 - M_2 = 0$ , where $M$ is the median	Small sample sizes, non-normal data	R: stats, coin; SPSS; online
Wilcoxon rank-sum test	Independent observations/ identical distributions	$H_0: F_1 = F_2$ , where $F$ is an unknown distribution	Small sample sizes, non-normal data	R: stats, coin; SPSS; online
Kruskal-Wallis	Independent observations/ identical distributions	$H_0: F_1 = F_2 = F_k$ , where $F$ is an unknown distribution	Small sample sizes, non-normal data	R: stats, coin; SPSS; online
Kolmogorov-Smirnov/ Anderson-Darling		$H_0: F_1 = F_2 = F_k$ , where $F$ is an unknown distribution	Non-normal data	R: stats, kSamples

In this chapter, we have also described several methods that can be used to analyze nontraditional data, including categorical data, as well as data that are non-normally distributed or distribution free. Table 23.3 also summarizes each of the statistical techniques covered in this chapter, in relation to four key pieces of information: underlying assumptions, null hypotheses that are tested, when to use, and available statistical software.

As discussed in the introduction, if a particular data set was collected using methodologically rigorous and appropriate sampling, instruments, and data entry and management practices, there is no immediate reason to assume that it contains bad data, regardless of its type, size, and distribution. These characteristics do not render such data *bad*, but they might render them *nontraditional*. As such, care should be taken by the researcher to identify and apply the most appropriate statistical measures to ensure that any conclusions and claims made about the data set are appropriate and accurate.

## References

- Biber, D., & Jones, J. (2009). Quantitative methods in corpus linguistics. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (Vol. 2, pp. 1286–1304). Berlin and New York: De Gruyter.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, England: Cambridge University Press.
- Donaldson, B. (2011). Left dislocation in near-native French. *Studies in Second Language Acquisition*, 33(3), 399–432.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82, 171–185.
- Fukuta, J. (2016). Effects of task repetition on learners' attention orientation in L2 oral production. *Language Teaching Research*, 20(3), 321–340.
- Ghanbar, H., & Plonsky, L. (under review). Multiple regression in L2 research: A methodological synthesis and meta-analysis of  $R^2$  values. Manuscript under review.
- Gries, S. T. (2013). Testing independent relationships. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 5817–5822). New York, NY: John Wiley & Sons.
- Keuhl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis* (2nd ed.). Belmont, CA: Cengage.
- LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46–77). New York, NY: Routledge.

- McDonald, J. H. (2014). *Handbook of biological statistics* (3rd ed.). Baltimore, MD: Sparky House Publishing.
- Ott, R. L., & Longnecker, M. (2010). *An introduction to statistical methods and data analysis*. Belmont, CA: Cengage.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687.
- Plonsky, L., Egbert, J., & LaFlair, G. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, *36*(5), 591–610.
- Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366.
- Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, *39*, 579–592.
- Thas, O. (2010). *Comparing distributions*. New York, NY: Springer.





# 24

## Reliability Analysis of Instruments and Data Coding

Kirby C. Grabowski and Saerhim Oh

### Introduction

Reliability, particularly as it relates to the field of applied linguistics (AL), is not a monolithic concept. There are perhaps almost as many conventions for approaching reliability analysis and its associated theories as there are research paradigms in the field. In this way, it is not a prescriptive enterprise, and attempts at maximizing and investigating reliability should be driven by both theoretical and practical considerations. Notwithstanding the need to allow for some flexibility within different methodologies, reliability analysis should still be rigorous and conducted from a place of statistical and theoretical literacy. Therefore, it is incumbent upon the researcher to understand the variables at play in any empirical study that may affect the reliability of the data collected, and how to investigate any effects on the consistency of measurement. An important distinction should also first be made between reliability as it relates to the external replicability of research findings and the reliability of instruments and data coding practices within a given study. While the former is certainly dependent on the latter since a study cannot really be replicated when research instruments or data coding practices are unreliable, the focus of this chapter is on reliability as it pertains to the consistency of observations or behaviors within a single study.

---

K. C. Grabowski (✉) • S. Oh  
Teachers College, Columbia University, New York, NY, USA  
e-mail: [kjc33@tc.columbia.edu](mailto:kjc33@tc.columbia.edu); [so2322@tc.columbia.edu](mailto:so2322@tc.columbia.edu)

Within quantitative research studies, reliability broadly refers to the extent to which results are consistent or stable and, by this very nature, is most often associated with the reliability of the data collected from various types of instruments<sup>1</sup> as well as the consistency seen in rating and data coding practices. By contrast, qualitative researchers tend to take a slightly different approach, where the emphasis is on maximizing reliability through study design and data collection strategies. Although both quantitative and qualitative considerations are obviously complementary, the purpose of this chapter is primarily to orient readers to reliability considerations with respect to instruments and data coding practices in quantitative AL research specifically, rather than maximizing reliability through qualitative strategies.

A body of recent research in the field of AL has demonstrated and critiqued how reliability estimates have been reported in quantitative studies, and most notably in second language acquisition (SLA) research (Derrick, 2015; Larson-Hall & Plonsky, 2015; Plonsky, 2013; Plonsky & Derrick, 2016). Meta-analyses and related studies like these serve to outline patterns seen in reporting practices in AL and highlight the importance of transparency in instrument reliability and data coding practices, which leads to a more meaningful and evidence-based interpretation of the results. As such, the aim of this chapter is to (1) raise awareness for AL researchers so that they will have a better understanding of some of the factors both affecting and affected by the reliability of their instruments and data coding practices, and (2) provide some guidelines and solutions for maximizing and investigating reliability within quantitative paradigms in the field.

It is important to note that although reliability is often considered subordinate to validity (and construct validity, in particular) within AL research, reliability should be seen as a necessary (though not sufficient) precondition for validity (Guilford, 1954). Therefore, the importance of reliability analysis should not be underestimated. Taken this way, ensuring the reliability of instruments and data coding practices is a preliminary hurdle—a prerequisite that needs to be addressed before additional analyses should continue. In certain areas of AL (e.g., language assessment), researchers will generally not proceed with analyses beyond an initial reliability analysis of an instrument if certain minimum criteria are not met. If a threshold is, in fact, not met, instruments may be overhauled and repiloted, rating procedures improved, data coding refined, and other issues that may affect the reliability are often individually investigated and improved before proceeding. Given all of the variables at play, uncovering issues of unreliability can be a very high-stakes enterprise, since this can often lead to costly detours in research projects, not to mention the uncomfortable situation in which researchers may find them-

selves having to disappoint stakeholders and go back to the drawing board. Nonetheless, ensuring reliability in this way is a necessary precondition. Otherwise, researchers are left with inconsistent data from which invalid findings and misleading interpretations and poorly-informed decisions may be made.

## Types of Reliability

A useful, if imperfect, distinction is often made between factors that are *external* to an instrument that affect reliability (i.e., external reliability) and those that are *internal* to an instrument that affect reliability (i.e., internal reliability). External reliability is often referred to as comparability between different uses of an instrument and variables, such as when there are multiple test forms in use or there is a need for the consistency in raters' assignment of scores on a writing test to be investigated. These are variables that can be thought of as being external to the individual items or tasks comprising an instrument itself. Variables associated with data coding practices, such as the extent to which independent coders agree when evaluating or categorizing instances or characteristics of a given unit of data, would also fall under the framework of external reliability. By contrast, internal reliability, as it relates to instruments, primarily relates to the consistency of results obtained from a collection of items or tasks within a single instrument, such as a 75-item grammar test, a speaking assessment with three prompts, or a motivation questionnaire. This type of internal reliability, termed *internal-consistency* reliability, is often what researchers are most interested in when investigating and reporting the reliability of their instruments. Although these two types of reliability are presented below as separate considerations within AL research, in many cases, a researcher will need to look at multiple variables affecting reliability, both external and internal to any instruments used, within the context of a single study.

## External Reliability Considerations

Results from instruments can vary on a number of different dimensions. Some of this variability is expected and meaningful given the purpose of the instrument. For example, variance in learners' test scores due to the differences in their reading ability is expected and meaningful on a test of reading ability, since the object of measurement matches the interpretations being made from test scores. However, factors that are unrelated to the purpose of

the instrument can also affect the results. For example, test scores can be affected by a difference in the types of test forms used, when the test was administered, the procedures of the test administration, and the consistency between raters/coders, among countless other factors. These factors, which are external to the instrument, may randomly and/or systematically affect the results and could be considered sources of measurement error. For an instrument to be reliable, the effects of these external factors should be minimized to the extent possible so that measurement error can also be minimized. Following is a discussion of the ways and indices used to estimate reliability related to four common factors (i.e., forms, occasions, raters, and coders) encountered in AL research, including methods for improving reliability. It goes without saying that there are other variables external to an instrument that can affect reliability, both in random and systematic ways, but it is our hope that the principles presented here can logically be applied to other instruments and contexts.

## Reliability Across Instrument Forms

One approach to estimating the reliability of an instrument is comparing scores or results across two different forms of an instrument given to the same group of participants. These different forms would ideally be *parallel* in nature. Parallel forms tap into the same underlying construct and use the same item or task types and formats even though the content of the items or tasks may be different. In order to ensure that the items and tasks on the parallel forms are equivalent, a detailed specification of the items and tasks is necessary in the design and development phase, as suggested in Bachman (1990) and Bachman and Palmer (1996, 2010), including characteristics of the task setting, the input, the expected response, and scoring procedures, among many others. Following such specifications when designing parallel forms of an instrument is a critical step that should not be undervalued as any unwanted variability between the two instruments will undoubtedly negatively impact the consistency of measurement between the two forms.

It should be said that creating parallel forms of an instrument simply to check the reliability of said instrument may not be a practical idea (Carr, 2011). However, there may be circumstances in which a researcher needs to ensure that two or more forms of an instrument are equivalent. In experimental research where different forms of an instrument for the pretest and posttest are used in succession or interchangeably, piloting the two forms prior to the actual data collection and verifying that the two forms can be used inter-

changeably is a crucial step needed to accurately and meaningfully interpret the results of the study. Similarly, in cases in which different forms of an instrument are used for the treatment group and the control group, again, the parallel nature of forms must first be investigated to confirm that the two forms can be used interchangeably. Without first establishing equivalence, the internal validity of the study is threatened. Systematically comparing forms is one way to uncover threats to reliability that may result from two different forms that actually may not be equivalent after all.<sup>2</sup>

The most common way of estimating the reliability of parallel forms of an instrument in most AL research contexts is by calculating the correlation coefficient between scores on parallel forms (see Norouzian & Plonsky, Chap. 19). A correlation procedure is typically a way of investigating the extent to which two sets of scores are related, and correlation coefficients indicate the magnitude and direction of any linear relationship. Correlation coefficients range from  $-1.0$  to  $1.0$ , with  $1.0$  indicating a perfect positive relationship (i.e., when values in one data set increase or decrease, the values in the other data set move in the same direction with the same magnitude). An example of a positive correlation would be hours reading outside of class and reading test scores. A correlation of  $-1.0$  indicates a perfect negative relationship (i.e., when values in one data set increase or decrease, the values in the other data set move in the opposite direction with the same magnitude; an example of a negative correlation would be days absent from class and course achievement). A correlation of  $0$  indicates no linear relationship between the two data sets (e.g., blood pressure and language learning motivation). In terms of interpreting the magnitude of correlation coefficients, it depends heavily on the context of measurement and the stakes of the decisions that are to be made based on the results. That said, convention says that a correlation of  $0.70$  or greater indicates a strong positive relationship, a correlation of  $0.50$  indicates a moderate positive relationship, and a correlation of  $0.30$  indicates a weak positive relationship (see further guidance in Plonsky & Oswald, 2014). A similar interpretation would also hold for the same values as negative correlation coefficients. The expectation for parallel-forms reliability is to have a correlation of at least  $0.75$  or higher, depending, of course, on the stakes and context of the study.

Two types of correlation procedures commonly used in AL research are the Pearson product-moment correlation and the Spearman rank-order correlation. The Pearson product-moment correlation coefficient is used when two sets of interval variables (e.g., scores from an instrument reported as the number of correct items out of the total number of items on the test) are correlated. In this way, each item contributes equally to the total score. For example, the Pearson product-moment correlation coefficient is calculated

when estimating parallel-forms reliability between participants' scores on two forms of a multiple-choice reading test. However, if the two data sets being correlated are ordinal variables (e.g., ratings from 0 to 5 given on a speaking test) or responses on a survey that include questions for which the responses are on a Likert-type scale (e.g., strongly agree to strongly disagree), the Spearman rank-order correlation coefficient is used. The variables in an ordinal scale have the property of being rank-ordered in terms of greater or lesser values, but unlike interval variables, the distance between the values on the ordinal scale may not, in reality, be equal. In other words, it may be more difficult for a test-taker to go from a rating of 4 to 5 on a speaking test than it is to go from a rating of 3 to 4. This artifact is taken into account when the Spearman rank-order correlation is used.

Another piece of information that needs to be taken into account when reporting either the Pearson product-moment correlation coefficient or the Spearman rank-order correlation coefficient is the significance of the correlation observed. In other words, it needs to be determined whether there is a probability that the correlation coefficient that is observed was due to chance or was systematic and, therefore, not due to chance. If the probability associated with the correlation coefficient is less than 0.05 ( $p < 0.05$ ), it can be said that the correlation coefficient is statistically significant at the 0.05 level, which indicates that there is a 95% chance that the correlation coefficient calculated is not due to chance. However, if the probability associated with the correlation coefficient is equal to or greater than 0.05 ( $p \geq 0.05$ ), it can be said that the correlation is not statistically significant, indicating that the correlation coefficient could be a chance phenomenon. It should be noted here that the correlation itself (and not the statistical significance) is what gives meaningful information about the magnitude and direction of the relationship between the two variables; statistical significance is simply an indication of the extent to which a correlation coefficient obtained is trustworthy and can be interpreted with confidence.

The correlation coefficient between two variables (e.g., scores from two forms of an instrument) includes measurement error; like the scores from any instruments, they are never perfectly reliable. Consequently, the correlations between test scores (or results from other instrument measures) are attenuated by the error in those measures, resulting in a lower correlation coefficient. To correct for this attenuation, Spearman (1904) proposed that a disattenuated correlation coefficient between two measures can be calculated by correcting the two measures for unreliability through an equation known as the double correction.<sup>3</sup> In cases in which only one of the measures needs to be corrected for, a single correction equation can be used. The results of such

a calculation often reveal a much higher correlation coefficient than the original, since the scores from the two forms of the instruments are assumed to have perfect reliability. Since the gap between the observed correlation coefficient and the disattenuated coefficient can be quite large, the results should be interpreted with caution (see Muchinsky, 1996, for the debates surrounding its use).

Another formula can be applied to correct for measurement unreliability when a correlation between two continuous variables is attenuated due to range restriction (i.e., when the entire range of population variability might not be represented in a sample). For example, a researcher may be trying to make inferences about the relationship between language proficiency test scores and job success, but there may only be data available on individuals in the upper end of the proficiency test score range. This can obviously impact the correlation coefficient and, thus, the inferences to be made, since higher proficiency learners may be the ones who have greater job success when compared to the (unrepresented) lower proficiency learners. Only relying on a necessarily restricted sample is fairly common in second language (L2) empirical research and meta-analyses, so researchers should be aware of the potential complications of it and also ways to correct for it (see Card, 2015, for further information).

## Instrument Reliability Across Occasions

Another approach to estimating the reliability is comparing results across different occasions, where an identical form of an instrument is given to the same group of participants. This approach entails several issues that need to be considered up front when estimating the reliability of an instrument. First, whenever an instrument is administered multiple times, participants may become familiar with its content. In other words, there may be gains in participants' scores on the second occasion due to practice effects. This is especially likely to occur when the time between the two occasions is relatively short. A second issue has to do with the changes, or maturation, in participants' ability in between the two occasions that has nothing to do with any treatment being delivered. On the one hand, they may have learned something related to the instrument content on their own between the two occasions (rather than the treatment or lack thereof) leading to receiving a higher score than expected on the second occasion of the test. On the other hand, they may receive a lower score than expected on the second occasion due to unexpected attrition between the two occasions or some unwanted interaction with the treatment.



This can be especially problematic if the time between the two occasions is relatively long. These unwanted sources of variance and measurement unreliability can be highly detrimental to the internal validity of a study. It should also be said that in the case of treatment groups in experimental research, the expectation is that there will, in fact, be a difference in the data collected at the pretest and posttest. In this case, it would be ideal to find that the two sets of data were *not* consistent.

Although testing contexts present some challenges when examining reliability across occasions, it may be beneficial to examine the reliability of surveys in this way, as familiarity and learning are not as major a concern for surveys as it may be for tests. It is important to note here that for any study involving an estimation of reliability across occasions, the expectation of measurement has to be that the results from the two occasions should be identical. This could also be true of a pretest/posttest design for the *control* group alone, when the expectation is for both sets of data to be consistent. Another example when we would expect scores to be identical or very close would be when measuring individual differences such as aptitude or personality. To estimate the reliability of a survey using this approach, the survey would be administered to the same group of respondents on two different occasions to confirm that the respondents' answers are the same across the occasions. The two occasions should be spread out in time to an extent so that the respondents do not remember the questions asked from the first occasion. However, the two occasions should be administered adjacent enough so that the respondents' opinions and perspectives do not considerably differ between the two occasions due to intervening variables or other factors.

One way of estimating the reliability of a survey instrument using the repeated approach would be by calculating the correlation coefficient between the responses from the two occasions. However, this method would be used only if all the questions in the survey were on a same scale (e.g., ordinal scale) and the responses to the survey questions add up to a meaningful number. For example, for a survey that consists of a set of statements asking the respondents to choose the degree to which they agree (e.g., on a scale of 1 – 5, from strongly disagree to strongly agree) with the given statement related to their reading habits, the reliability of each of the questions in the instrument can be estimated by using the Spearman rank-order correlation (as discussed above), as the two data sets being correlated are ordinal variables. The reliability of the entire instrument, then, can be estimated using the Pearson product-moment correlation procedure, since the composite value would be on an interval scale. However, for a survey that consists of a set of statements not related to each other and/or with responses on different types of scales, computing the correlation coefficient is not appropriate. For instance, if the survey consists of



some questions asking the participants to select the degree to which they agree with statements related to their learning habits in L2 (i.e., ordinal variables) and other questions asking about the types of materials they use to assist their learning (i.e., nominal/categorical variables), the type of scales used in the questions is not consistent, and the sum of the responses would not be meaningful. Needless to say, as described earlier, if the data are suitable to report a correlation coefficient, the statistical significance of the relationship should also be reported on.

Another common way of estimating the reliability of a survey instrument across two occasions is by comparing the respondents' responses between the two occasions and computing the agreement rate. The number of questions that all respondents gave the same responses across the two occasions is counted and then is divided by the total number of questions in the survey. For example, if a survey with 30 questions was administered to 50 participants, and if all 50 participants gave the same response to 19 questions, then the agreement rate would be 0.63:

$$\frac{(\text{Number of questions with the same responses from all respondents})}{(\text{Number of questions in the survey})} = \frac{19}{30} = 0.63$$

This indicates that 63% of the responses were in agreement between the two occasions, and this may be interpreted that the survey was 63% reliable.

The agreement rate for each question can also be computed to examine each question in the survey individually. The agreement rate can be calculated for each item by counting the number of respondents who gave the same response for the two occasions and dividing that number by the total number of respondents. For example, for questions answered by 50 participants in each occasion, if the number of respondents who answered exactly the same in both occasions was 25 for the first question and 45 for the second question, the agreement rate between the two occasions would be 0.50 and 0.90 for the first and second questions, respectively. This means that the second question elicited more consistent responses.

For each question to have a high agreement rate and eventually lead to a high agreement rate of the entire survey, each question in the survey should be designed carefully. Brown (2001) provides detailed guidelines for writing precise and clear survey questions. He discusses three aspects of survey questions that need to be considered: the form of the questions (e.g., length, clarity, negative and positive form, format), the meaning associated with the questions (e.g., leading information, bias content, embarrassing content), and the respondents who would be answering the questions (e.g., level of lan-

guage, relevance to the respondents). The repeated approach of calculating the correlation coefficient between responses from two different occasions is especially helpful for checking for these three aspects of survey questions. For example, if respondents have different answers on the second occasion as opposed to the first occasion, there is a possibility that the question was not written in an appropriate way in terms of its form, meaning, and appropriateness for the respondents.

## Reliability Within and Across Raters

Best practices in instrument scoring would indicate that it is best to have at least two raters wherever possible when performance-based responses are being elicited for measurement purposes, as is often the case in speaking and writing assessments. As such, these types of instruments usually involve human judgment, and, due to the subjectivity of their behavior, raters are a potential source of measurement error. For instance, though recommended, having two or more raters involved in the scoring process may result in systematic inconsistencies (e.g., one rater being more lenient and the other rater being more severe) and/or erratic behavior affecting the scores. Similarly, individual rater may be self-inconsistent during the rating process (e.g., being lenient in the beginning but severe toward the end). In both cases, these inconsistencies can lead to measurement error and consequently to low inter-rater reliability (in the case of two or more raters) or low intra-rater reliability (in the case of a single rater).

In order to improve consistency in raters' scoring, first and foremost, a clear scoring rubric is necessary. A rubric should be designed based on the construct underlying the instrument, and it should be thought of as an explicit representation of that construct. In some cases, researchers utilize scale descriptors that are adapted from existing rubrics or those that include general representations of knowledge, skills, and abilities at various levels of proficiency. At other times, researchers may opt for the interpretive benefits of empirically-derived descriptors or indigenous scoring criteria, even though more effort is necessarily needed during the design stage. A rubric can be *holistic* or *analytic* depending on how the construct is conceived for scoring: holistic for when a single score is assigned to the overall performance or analytic when a profile of scores is assigned, each representing a subcomponent of the ability being measured. To take a summary writing assessment rubric as an example, the construct being measured may include the subcomponents of Content Control, Organizational Control, and Language Control scored together (as in a holistic rubric) or separated into

individual subscales (as in an analytic rubric). In both cases, having transparent and unambiguous descriptors of the abilities being measured is a crucial first step for raters to be able to interpret the scales and be consistent.

In order to ensure that the rubric matches the variability of performances that may be elicited, a clear description of what each level or score means can also be accomplished through descriptors of what learners know or can do in each level or score band. Explicit descriptors help maximize accountability and transparency in scoring for all stakeholders. Finally, the language used in rubric descriptors should be parallel across the levels or scores. To continue with the example of the summary writing assessment rubric, if a score of 3 (the highest score) for Organizational Control is described as the following: “The summary is *well organized*, and the information is *coherently presented*. Relationship between ideas are *clear and coherent*,” then the descriptor of a score of 2 for Organization Control should be parallel to this description in terms of the language used and, thus, could read something like the following: “The summary is *generally well organized*, and the information is *generally coherently presented*. Relationships between ideas are *at times not immediately clear*.” This logical gradation of language not only makes it easier for raters to compare the level descriptors in reference to one another, but it also provides a concrete representation of the progression of development in the abilities being measured through the task.

Despite best efforts to include detailed and clear descriptors in a rubric, individual raters may still interpret the rubric in different ways. And without technology, as is the case in most small-scale research studies, it is difficult to detect discrepancies while the rating process is taking place. In order to help raters understand and interpret the rubric in the same way and help them be both self-consistent in their own ratings and consistent with other raters' scores, rater training and norming are crucial components of the scoring process (Elder, Knoch, Barkhuizen, & von Randow; Kim, 2015; Lim, 2011; Weigle, 1998). In the training and norming process, raters are introduced to the rubric and asked to rate sample responses (e.g., essays or recorded spoken responses) in order to discuss the performances in relation to the rubric descriptors and arrive at a common interpretation of the rubric. The specific samples used for norming purposes are typically selected to represent a range of abilities and to problematize issues that may appear during the rating process. One of the main goals of rater training and norming is to reduce the amount of random error in rater judgments and concurrently increase raters' self-consistency. In addition, these procedures have been shown to have the effect of reducing extreme differences (i.e., outliers in terms of harshness or leniency are brought back into conformity and consistency with the other

raters) (Weigle, 1994). The training and norming process also clarify the intended scoring rubrics and modify expectations by exemplifying the nature of the tasks and the characteristics of the learners (Huot, 1993; Weigle, 1994, 1998).

One typical way of carrying out the norming process is by having the rater trainer and the trainees (i.e., the raters) meet face-to-face as a group. Scoring of the samples is then carried out independently and subsequently discussed as a whole group. In addition to the group rater training, the norming process can be streamlined and even made ongoing by using online training (e.g., Elder, Knoch, Barkhuizen, & von Randow, 2005; Hamilton, Reddel, & Spratt, 2001) or by using various forms of individual feedback, such as judge performance reports (e.g., Lunz & Stahl, 1990; Wigglesworth, 1993). There are many approaches to rater norming and feedback mechanisms, and individual researchers need to decide for themselves what is the best approach to maximize rater reliability. For a discussion on rater feedback techniques, see Elder et al. (2005), Wigglesworth (1993), and O'Sullivan and Rignall (2007).

Similar to estimating the reliability related to parallel forms and different occasions, in order to estimate the reliability of the scores assigned by raters, one can first calculate the agreement percentage. In doing so, the exact agreement rate along with the adjacent agreement rate can be calculated. The exact agreement rate can be computed by counting the instances in which the two raters assigned the same score and dividing that number by the total number of ratings, similar to how it would be done for the agreement rate for occasions, as exemplified above. For example, if there were 50 test-takers, and two raters assigned the same score to 10 test-takers, then the exact agreement rate would be 0.20.

$$\frac{(\text{Number of scores with the same scores assigned by the two raters})}{(\text{Number of scores})} = \frac{10}{50} = 0.20$$

A difference of one point between the raters score also provides evidence of the two raters being consistent in their ratings to a certain extent. Thus, the *adjacent* agreement rate should also be calculated and reported on where raters are concerned. In computing the adjacent agreement rate, instead of counting the instances in which the two raters' scores were exactly the same, the instances where a difference of one point was found in the raters' scores are counted. This figure is then divided by the total number of scores available. Again, for instance, among the 50 test-takers' scores, if 25 test-takers had a one point difference between the two raters' scores, the adjacent agreement rate would be 0.50.

(Number of scores assigned by the two raters with a difference of 1 point) /

$$(\text{Number of scores}) = \frac{25}{50} = 0.50$$

In calculating the difference of scores between the two ratings, the direction of whether the first rater assigned a point higher or lower does not matter, as the absolute value of the difference of scores is used in computing the difference of scores. If an analytic rubric is used instead of a holistic rubric, the exact agreement rate and the adjacent agreement rate should be computed for each subscale of the rubric. For example, in the summary writing assessment, since the rubric includes three subscales (i.e., Content Control, Organizational Control, and Linguistic Control), then the exact agreement rate and the adjacent agreement rate should be calculated separately for each subscale.

In addition to the agreement rate, the most commonly used method of estimating the reliability of scores assigned by raters is by calculating inter-rater reliability in the form of a correlation coefficient. As mentioned above, a correlation coefficient of two data sets (i.e., two sets of scores each assigned by one rater) expresses the extent to which the two sets of scores are related. As described earlier, the Spearman rank-order correlation coefficient would be used for ordinal data (e.g., ratings from two raters based on a holistic rubric or a single subscale on an analytic rubric) and the Pearson product-moment correlation coefficient would be used with interval data (e.g., the average or total scores from the raters across the subscales on a scoring rubric). However, in cases where the score distribution of interval data is non-normal (i.e., skewness and/or kurtosis values are not within the 2 range; Bachman, 2004), the Spearman rank-order correlation coefficient should be used instead. Again, no matter which correlation procedure is employed, the significance level of the coefficient should also be calculated and reported in order to confirm that the observed correlation coefficient is not a chance phenomenon. A rough guideline for inter-rater reliability would be a minimum coefficient of 0.80; however, it goes without saying that this criterion can and should be adjusted (usually up, but sometimes down), as long as the level can be defended, depending on the purpose and stakes of the measurement context.

In cases where there is only one rater involved in the process of scoring, the same procedure can be used to calculate *intra-rater reliability*. The only difference would be that instead of having two raters, the one rater would score the performances twice with a time interval in between the two occasions. The rater would then blindly score (i.e., without looking at the scores given on the first occasion) a second time. The two sets of scores would then be correlated using an appropriate correlation coefficient, depending on the type of data that was obtained.

While a high inter-rater reliability indicates that the two raters were in alignment with each other, this calculation does not necessarily account for the severity or leniency of the raters, nor how self-consistent was their assignment of scores. One statistical model that can account for individual rater variation is many-facet Rasch measurement (MFRM) (Linacre, 1989; Linacre & Wright, 1992). The utility of MFRM is that it allows for an investigation into the behavior of individual raters. For instance, it can provide information about how systematically severe (or lenient) raters are, how they may be random or erratic in their assignment of scores, and also how they may show bias when interacting with individual candidates, items or tasks, or other facets of the measurement design. This information can be used for ongoing rater training or even score adjustment, all depending on the purpose and stakes of the measurement context.

## Reliability Within and Across Coders

In addition to quantitative data, there are an increasing number of studies in AL research where qualitative data may be the focus of analysis. These qualitative data may include data from interviews, think-aloud protocols, retrospections, written texts, and observations, and coding is a technique used to organize these data into subcategories and themes in order to make interpretations of the learners' knowledge, skills, and abilities (Ellis & Barkhuizen, 2005). In such cases, coders are involved in the process, and these coders could contribute to the variance (and error) in the results, and, thus, are an important factor to consider when discussing how reliable the codings are.

Just as having a clear scoring rubric is a necessity in improving the reliability of scores assigned by raters, having a clear coding scheme is also crucial in improving the reliability of qualitative codes. Designing a coding scheme prior to reviewing the data may be possible, but in most cases, a coding scheme emerges from the data by making sense of what is found in the data. Ellis and Barkhuizen (2005) describe the process of developing a coding scheme as the following: first, researchers need to decide on unit of analysis (i.e., what is going to be coded). This could include words, phrases, thinking, and behavior that appear in the data. Second, the actual coding is accomplished by assigning tags that represent ideas, performances, experiences, or other concepts identified in the data. These codes along with their definitions should be organized in the coding scheme. Ellis and Barkhuizen state that these codes and the coding scheme may change as the study continues.

Even if the coding scheme is clear, when there is only one coder involved in the coding process, it is difficult to assure that the analysis of the data is reliable since the notion of reliability entails consistency. Therefore, it is crucial to have at least two coders involved in the process, and similar to estimating the reliability of scoring procedures where raters are involved, the agreement rates of the coding between the two coders should be calculated. This can be accomplished as a single percentage agreement. However, when coding categorical variables (e.g., correct vs. incorrect; lexical vs. syntactic error), it is possible for two individuals to agree by chance. For this reason, Cohen's kappa  $\kappa$  (often simply called kappa) can be used to calculate the degree of agreement between individuals, accounting for statistical inflation due to chance agreement. Additionally, having a third coder analyze any data that include sources of disagreement could improve the reliability of the analysis. However, in situations where it is impossible to have more than two coders, the alternative would be to have the one coder involved in the coding process code the data twice, as is recommended with intra-rater reliability. The agreement rate between the two occasions would then be reported to provide evidence of reliability. It is important to note that reliability of qualitative data is not "a fit between what they [qualitative researchers] record as data and what actually occurs in the setting under study" but "the literal consistency across different observations" (Bogdan & Biklen, 2003, p. 36). Thus, reliability in the context of coding should still be seen in terms of consistency alone, rather than as something more closely related to the internal validity of the coding scheme.

Coders may be normed in much the same way as raters are, but unlike how raters are often normed with a predetermined rubric that is provided to them, coders may be more involved in the process of developing the coding scheme from the data itself (perhaps through grounded theory; see Hadley, 2017) and making changes in the coding scheme when unexpected findings are observed. Additionally, compared to the norming procedure for raters' scoring, where raters are being trained to be consistent in the scores they assign, the norming process for the coders may be more complex and adaptive and may not be a one-time occasion.

## Internal-Consistency Reliability Considerations

Instruments can be thought of as a collection of items or tasks that are intended to measure the same underlying construct. The hope, then, is to maximize the amount of homogeneity in the items within a single instrument. In other words, the goal is to compose an instrument in such a way that



each individual item or task is working together with all the other items or tasks to collectively measure the same underlying trait. Within this perspective, the concept of homogeneity doesn't so much refer to *what* an instrument is measuring—that's under the purview of (construct) validity—but it should give an indication of the extent to which the items or tasks are all measuring the same thing. So, while an internal-consistency reliability estimate can give an indication of how well items or tasks in an instrument work together to measure the same thing, they don't necessarily give any indication of what, exactly, that thing is. This is why having an instrument with high reliability does not necessarily guarantee validity (i.e., that it is measuring what it was designed to measure). Nonetheless, without high internal-consistency reliability, it is hard to argue that there is meaningful evidence for the construct validity of an instrument. In this way, reliability, then, can be seen as a fundamental precondition for validity.

It is also important to point out that instruments in and of themselves do not have a *de facto* reliability. This can potentially be confusing for researchers, especially when they are using already-published instruments with accompanying literature that may include a reliability statistic for the said instrument. Large-scale assessments also often have reliability estimates reported in research reports and other published materials, and these may be very stable over time, but conventional raw-score reliability estimates are sample dependent. In other words, a reliability estimate calculated for any given instrument administered on any given occasion will be unique to that sample of participants at that particular occasion. If the instrument were to be administered to a seemingly similar subset of a population of learners, say in another course section of an identically leveled language class in a school, it does not guarantee that the same estimate will present for that second sample. Therefore, it is of paramount importance for researchers to understand how to calculate and interpret estimates of internal-consistency reliability for each individual instrument and each individual occasion or administration that they may be including in their research.

While there are several procedures for estimating internal-consistency reliability based on a single administration of an instrument on one occasion (e.g., Spearman-Brown split-half and Guttman split-half), Plonsky and Derrick's (2016) synthesis has shown that coefficient alpha  $\alpha$  (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), frequently shortened simply to "alpha," is the index most frequently used to estimate the reliability of a single set of scores. Although alpha has well-documented limitations (Cortina, 1993; Hogan, Benjamin, & Brezinski, 2000; Lord & Novick, 1968; Yang & Green, 2011), it is often considered superior to other split-half estimates used in



social science research, since it accounts for all the random possible inter-item correlations on an instrument (i.e., it is an average of all possible split-half estimates, and not just one configuration of instrument halves). Alpha is also the most frequently reported measure of internal-consistency reliability in the social sciences, and it is easy to calculate and obtain since it is available in most commercial statistical packages (e.g., SPSS) and can even be calculated using Excel (Carr, 2011). Some researchers choose to use Kuder-Richardson Formula 20 (KR-20) estimates as a special case of alpha for dichotomous data (e.g., item scores of 0 and 1), but most researchers opt for alpha as it can accommodate dichotomous scores as well as continuous data (e.g., ratings from a rubric or composite scores on a test). Since alpha tends to be a lower-bound estimate of reliability, some researchers choose to use alternative estimates for different types of data (e.g., ordinal vs. interval), each estimate with its own strengths and limitations. Every estimate takes account of different restrictions on the data set in terms of assumptions of uni-/multidimensionality, treatment of uncorrelated errors, sensitivity to sample size, and assumptions of a non-skewed distribution, among others. While many of these issues are quite technical in nature and beyond the scope of this chapter, researchers should be mindful that these issues may be present. As such, they should approach reliability analysis accordingly, depending on the type of data they have and the context of measurement.

In terms of awareness building, one issue that is of particular interest to AL researchers is that of skewed distributions, since many of the instruments used in the field and decisions made about participants are criterion-referenced (CR) in nature. This means that scores are best interpreted in relation to descriptive criteria, such as standards, objectives, or scale descriptors, which reference the knowledge, skills, and/or abilities that candidates possess at a given score level or band. Typically, CR instruments are administered in small-scale settings (e.g., classroom achievement testing) and result in most participants' scores being clustered at the high (or low) end of the scale, as opposed to being normally distributed, as is more often the case in large-scale and/or norm-referenced assessments. In experimental research, pretests and posttests also often result in skewed distributions (i.e., positive skewness for the pretest and negative skewness for the treatment group's posttest). These types of instruments are extremely common in AL research. While it is true that many large-scale assessments in use today are labeled as CR, they nonetheless have many qualities of norm-referenced tests (e.g., large sample sizes, extended test length, normal distribution of scores, and context-independent constructs such as "language proficiency"), even if their score interpretations can be viewed with reference to descriptive criteria rather than in terms of the relative standing of

candidates. Although the estimates of internal-consistency reliability (and external reliability, for that matter), including alpha, are most closely related to norm-referenced (NR) test uses, most researchers in AL report alpha as an estimate of instrument reliability, whether their instrument and intended decisions would be classified as NR or CR. Nonetheless, by convention, researchers in all areas of AL tend to use alpha as a default reliability estimate. However, it should be noted that additional investigations into reliability can and should be made for instruments that fall on the extreme end of the CR/NR continuum (e.g., those with heavily skewed distributions, limited data sets, etc.). CR approaches to reliability analysis may even help researchers salvage instruments and data that may otherwise be deemed unreliable in the NR framework. (For a thorough discussion on CR testing and estimating reliability for criterion-referenced tests, refer to Brown & Hudson, 2002, and Bachman, 2004, respectively).

No discussion of alpha would be complete without a brief introduction to classical test theory (CTT), since an accurate and meaningful interpretation of alpha rests on a firm understanding of the theory of reliability within CTT. It should be noted that for this discussion, we will be using the word “test” as an example of an instrument and “test-takers” to exemplify candidates since the frame given is CTT, but the theory of reliability as it pertains to other types of instruments would be viewed in much the same way. When a test-taker receives a raw score on a test, that observed score is meant to be an indicator of some underlying ability with as little error included in the measure as possible. However, an observation of performance is necessarily removed from a person’s true ability given that the sample of underlying ability being elicited was done so indirectly through a measurement instrument. Within CTT, that observed score (e.g., a test-taker’s raw score on a test) can be seen as an approximation of a person’s true score (i.e., a test-taker’s average score if they were to take a test an infinite number of times) with the addition of some amount of random error, since a single observed score cannot ever match a true score with 100% accuracy. Very generally, then, alpha is a representation of the amount of meaningful variance in test scores that can be attributed to a test-taker’s true score variance and not error variance. The goal is to have as little error as possible. Thus, the greater the proportion of variance in the observed scores that can be accounted for by the variance in the test-takers’ true scores, the less error variance there is, and the more reliable the test will be. Conversely, the more error in the measure, the less reliable the scores will be as a reflection of the test-takers’ true scores.

As an example, if the alpha of a test were to be estimated at 0.78, we can say that 78% of the variance in observed test scores can be accounted for by

the variance in the test-takers' true scores (i.e., construct-relevant variance), with 22% of the variance accounted for by measurement error (i.e., construct-irrelevant variance). Similarly, if an alpha estimate comes out at 0.45, we can say that 45% of the variance is meaningful, or construct-relevant, variance, with 55% of the variance accounted for by error variance. Reliability also affects other estimates, such as the standard error of measurement (SEM), of which one constituent component in the calculation is instrument reliability. Very generally, SEM can be used, by extension, to give an indication of how confident we can be that a person's true score falls within a certain range of a single observed score, or that a given cutoff score (and not some lower or higher score) on an assessment can be seen as fair. The lower the reliability is, the wider this range would be, and the less sure a test-score user, such as a language program administrator or researcher, can be that any given observed score reflects a person's true score, or that a chosen cutoff score should not be modified up or down for fairness' sake. Low reliability can also affect the magnitude and statistical significance of correlations, *t*-tests, ANOVAs, and chi-square statistics, and effect sizes, among myriad other calculations that may transpire beyond a reliability analysis. Simply put, if there is low reliability due to a lack of systematicity of the patterns in the data itself, it will affect the meaningfulness and trustworthiness of nearly every other analysis and interpretation.

In terms of standards or guidelines for a minimum acceptable alpha estimate, it ultimately depends on many contextual factors, not the least of which are the stakes of the decisions that are to be made based on the instrument data and/or the study results. Each individual researcher needs to decide what a minimum level of acceptability is for their instrument. Ideally, reliability evidence should be established and buttressed within an accompanying validation framework, like Kane's (2006, 2013) argument-based approach to validity or Bachman's (2005) and Bachman and Palmer's (2010) assessment use argument, but specific standards for this practice will vary somewhat within different disciplines of AL. Very roughly, while an alpha of 0.75 may be acceptable for most classroom contexts (and other similarly low-stakes decisions), an alpha of 0.90 or higher would be imperative for large-scale contexts and high-stakes decision-making. That said, any guidelines for minimum acceptability for an alpha estimate are crude when considered independent of any context. Therefore, it is important to keep in mind that stakes are often in the eye of the beholder, and test results that are used to make decisions (about test-takers, curricula, teacher promotion, etc.) should meet a defensible standard for reliability in whatever context in which they are to be used.

While the APA suggests that instrument reliability should always be reported, along with the number of items that make up the subscale, we would further emphasize that, at a minimum, simply reporting alpha as evidence of an instrument's reliability is not enough; a reliability estimate should be interpreted for its meaningfulness and acceptability within the context of a given instrument's context of use. If a minimum level of acceptability cannot be adequately defined and defended, it is incumbent upon the researcher to remedy the situation through instrument revision, or even a complete redesign if that is warranted, given the many practical and ethical complexities inherent in empirical research. Although any instrument setbacks will inevitably delay the study, implications can only be as good as the data they are based on.

Given the, at times, high-stakes nature of instrument revision, researchers may be interested in the many safeguards available for protecting against low reliability at the outset. There are also many techniques for remedying an instrument after unfortunate results have come in. It should go without saying that, in the instrument design phase, piloting is a must. Securing grants or other funding, recruiting participants, getting IRB and other necessary approvals, and implementing a study's design and data collection phases often take time, dozens, if not hundreds, of people, and a tremendous amount of resources to complete. All too often, researchers are blind to the negative consequences of having an instrument that is not reliable and find that these are exactly the circumstances in which they find themselves after the initial data analysis phase. What are they to do? Start over? Many do not for a variety of reasons (e.g., lack of ongoing funding or other resources, lack of understanding or training, or simple ego), and, thus, gloss over their low reliability in their reporting, or they may ignore it altogether and not report it at all. Many unresolved or controversial issues in AL research can be traced back to deficient reliability analyses, including insufficient piloting of instruments and/or assumptions that reliability is an inherent quality of an instrument regardless of the sample or occasion. In fact, many published studies in the field could be considered pilot studies since the first time the instrument was administered and analyzed to any great degree was in the operational context of the study. This is not an irrevocable problem, however. Low reliability is a pitfall that can be assuaged through the adequate piloting of an instrument in advance of the operational context. Once instruments are piloted, researchers should familiarize themselves with relatively accessible procedures such as classical item analysis (CIA), which provide insight into instrument quality, including item difficulty and item discrimination, both of which give clues about the individual items' homogeneity within the scale of items that com-

prise the instrument. There are many complementary approaches to improving the quality of items, but CIA is recommended as a crucial first step in improving instrument reliability (see Carr, 2011, for information on conducting CIA). Based on empirical evidence (e.g., statistics from CIA), decisions can be made about whether items should be removed or retained in an instrument. While removing items that have less than ideal discrimination (i.e., a D-index below 0.3) will increase an alpha estimate, it is still important to evaluate practical and theoretical considerations for an item's inclusion or exclusion from an instrument (e.g., an item captures a unique aspect of the construct). While it is true that piloting procedures and instrument revision may take up a significant portion of the researcher's effort and resources, there is usually more than return on this initial investment in instrument quality, and therefore study quality, in the long run.

## Going Beyond Classical Test Theory Reliability

If AL researchers have multiple sources of variability in their research design (e.g., raters, tasks/prompts, occasions, nesting of participants within treatment groups, etc.), they may want to consider measurement models as alternatives to CTT, such as generalizability theory (G-theory) (Brennan, 2001; Shavelson & Webb, 1991) and many-facet Rasch measurement (MFRM) (Linacre, 1989), which make it possible to model and address the effect of multiple sources of variance (and their interactions) on reliability at a range of ability levels. These two complementary approaches provide information about the relative effects of test facets (e.g., raters and tasks/prompts) through G-theory, and rater or task/prompt specific effects (e.g., a rater who is systematically more lenient than the model predicts) through MFRM. In performance assessment contexts, an examination of test method facets purported to have the greatest impact on scores, such as raters and tasks/prompts (Bachman, 1990, 2004, 2005; McNamara, 1996), is crucial in an investigation into score reliability. These two multifaceted measurement models allow for an investigation into the effect of test method facets on scores, as well as test-takers' ability level differences, and systematic interaction effects between facets (e.g., whether certain rater was systematically more severe in scoring a certain task), not accounted for in CTT. Although well beyond the scope of this chapter, researchers may want to familiarize themselves with these models in an ongoing effort to approach reliability analysis from a place of statistical literacy, to gather more sophisticated evidence in support of their instrument's validity argument.

## Conclusion

Most AL researchers would agree that instruments and data coding practices should be reliable. Nevertheless, limitations with respect to reliability analysis and reporting are still apparent in L2 research. The purpose of this chapter was to discuss reliability in the context of AL research, particularly to raise awareness of the importance of analyzing reliability appropriately, interpreting estimates of reliability in an accurate and meaningful way, and improving the reliability of instrument scoring and data coding. In addition to the discussion of reliability within CTT, we outlined alternative measurement models that researchers can consider when multiple sources of variability are included in their research.

## Resources for Further Reading

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Chapters 5 and 6, respectively, provide a technical yet accessible discussion of reliability and reliability estimates for both norm-referenced and criterion-referenced instruments and decision-making. These chapters also include formulae and calculations for researchers wanting to know more about the foundational statistics behind various reliability estimates.

Carr, N. (2011). *Designing and analyzing language tests*. Oxford, UK: Oxford University Press.

Although primarily aimed at a language testing audience, this handbook is a practical guide for novice researchers developing and analyzing instruments of their own. It includes chapters not only on reliability analysis and correlations but also on a conceptual introduction to CIA, and procedures and digital tutorials for carrying out all analyses using Excel.

Phakiti, A. (2014). *Experimental research methods in language learning*. London: Bloomsbury.

Chapter 12, in addition to providing an introduction to reliability and reliability analysis for L2 research, includes screenshots and step-by-step instructions for calculating various reliability estimates using SPSS. There are examples for various instrument types, as well as interpretations of relevant statistics.

## Notes

1. In this chapter, we will be using the general term *instrument* to encompass a variety of tools that may be employed in empirical studies, such as tests, surveys, performance assessments, questionnaires, and others. We will also use more specific terms where relevant and when a more precise illustration is helpful or required.
2. It should also be noted that in some areas of AL research (e.g., large-scale language assessment), certain approaches to reliability analysis, such as parallel-forms reliability or test-retest reliability, are becoming more and more obsolete as item- and task-banking and computer-adaptive testing are replacing former assessment delivery methods, such as traditional paper-and-pencil tests and even some first-generation computer-based tests. Most of the data collected from these newer large-scale test-delivery systems have properties that make traditional approaches to reliability estimation inefficient, if not impossible. In most cases, psychometricians and other measurement professionals charged with analyzing the data often use item response theory (IRT)/Rasch in their analysis, as each test-taker may, in theory, receive a unique set of items or tasks on any given occasion. As access to this type of technology is quite rare in most AL research contexts, we have chosen to present approaches to reliability analysis here that are accessible to most, if not all, AL researchers.
3. Most statistics discussed in the chapter are easily obtained using statistical software, such as SPSS (SPSS, Inc.), and, thus, do not require hand calculations. However, for information on formulae related to different types of reliability, or how to calculate reliability statistics by hand, please see Resources for Further Reading at the end of the chapter.

## References

- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 20(1), 1–34.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.
- Bogdan, R. C., & Biklen, S. K. (2003). *Qualitative research in education*. Boston, MA: Allyn and Bacon.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.



- Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, UK: Cambridge University Press.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, UK: Cambridge University Press.
- Card, N. (2015). *Applied meta-analysis for social science research*. New York: The Guilford Press.
- Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford, UK: Oxford University Press.
- Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.
- Derrick, D. (2015). Instrument reporting practices in second language research. *TESOL Quarterly*, 50(1), 132–153.
- Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, 2(3), 175–196.
- Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language*. Oxford, UK: Oxford University Press.
- Guilford, J. P. (1954). *Psychometric methods*. Bombay, India: Tata-McGraw Hill.
- Hadley, G. (2017). *Grounded theory in applied linguistics research: A practical guide*. London: Routledge.
- Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perception of on-line rater training and monitoring. *System*, 29(4), 505–520.
- Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, 60, 523–561.
- Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206–232). Cresskill, NJ: Hampton Press.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: American Council of Education and Praeger Series on Higher Education.
- Kane, M. (2013). Validating and interpretation and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, 12(3), 239–261.
- Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(1), 127–159.
- Lim, G. (2011). The development and maintenance of rater quality in performance writing assessment: a longitudinal study of new and experienced raters. *Language Testing*, 28(4), 543–560.



- Linacre, J., & Wright, B. (1992). *A user's guide to FACETS: Rasch measurement computer program*. Chicago, IL: MESA Press.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley Publishing Company.
- Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13(14), 425–444.
- McNamara, T. (1996). *Measuring second language performance*. London: Longman.
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56(1), 63–75.
- O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers. Research in speaking and writing performance* (pp. 446–478). Cambridge, UK: Cambridge University Press.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35(4), 655–687.
- Plonsky, L., & Derrick, D. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, 100(2), 538–553.
- Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912.
- Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1), 72–101.
- Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11(2), 197–223.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–287.
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305–335.
- Yang, Y., & Green, S. B. (2011). Coefficient Alpha: a reliability coefficient for the 21st Century? *Journal of Psychoeducational Assessment*, 29, 377–392.



# 25

## Analyzing Spoken and Written Discourse: A Role for Natural Language Processing Tools

Scott A. Crossley and Kristopher Kyle

### Introduction

Beyond archeological remains, history began with the written word. The written word has allowed us to track human thought, human emotion, and human development. Before the written word, histories, myths, and stories were passed on orally from generation to generation. These were fleeting, dynamic, and inconsistent because they were spoken and, as such, provided no record beyond the memory of those that spoke and heard them. More recently, modern technology has allowed us to capture stories electronically and store them as mostly permanent records in much the same way as paper retained earlier records of written words. Over time, as these electronic collections of written and spoken words have grown in size, we have at our disposal a vast repository of language available for examination at both the written and spoken levels. This repository is a treasure of information that can afford insight into a number of human phenomena of interest to scientists, historians, and the layperson. Unfortunately, the body of recorded speech and writings that comprises

---

S. A. Crossley (✉)

Department of Applied Linguistics & ESL, Georgia State University,  
Atlanta, GA, USA

e-mail: [sacrossley@gmail.com](mailto:sacrossley@gmail.com)

K. Kyle

Department of Second Language Studies, University of Hawai'i at Manoa,  
Honolulu, HI, USA

e-mail: [kristopherkyle1@gmail.com](mailto:kristopherkyle1@gmail.com)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_25](https://doi.org/10.1057/978-1-137-59900-1_25)

human history is colossal and overwhelming. It is beyond the comprehension of a single person working individually or even a team of scientists. However, the data contained within this body of language is incredibly rich and can provide evidence for a number of important singularities and spectacles unique to human cognition and social contexts. The question, of course, is how to access the language in this data? And more importantly, how to analyze the data and find meaningful patterns within the language?

The sheer number of texts and words would make the task impossible or at least extremely time-consuming and resource intensive using traditional linguistic methods such as qualitative discourse analysis (Fairclough, 2013; Schiffrin, 1994) or even corpus linguistics approaches (see Paquot, Chap. 17). But, of course, as technology offers the opportunity to access a greater number of texts, it also offers new language techniques with which to analyze these texts. Principally, these new techniques fall under the field of computational linguistics, which uses computational models to investigate language features. More specifically, computational techniques can be realized as applications of natural language processing (NLP; Chowdhury, 2003), which focuses on automating the analysis of linguistic features and structures to describe how language is learned, represented, generated, and understood.

The purpose of this chapter is to provide an overview of how spoken and written discourse has been quantitatively analyzed in the past and to contrast these approaches with more contemporary NLP techniques that are currently used. Specifically, this chapter will provide an overview of how NLP tools (i.e., NLP applications that can be used by nonspecialists to conduct language analyses) that measure lexical, syntactic, rhetorical, and cohesive features of text can be, and have been, used to examine spoken and written discourse. The chapter will also provide an introduction to freely available NLP tools, an overview of the output produced by these tools, and statistical methods used to analyze and interpret the output produced from these tools.

## Spoken and Written Discourse Analysis

Spoken and written discourse has been, and continues to be, analyzed at a variety of scales (e.g., from a few hundred words to a billion or more words), for a variety of purposes (e.g., descriptively outlining a target language, investigating developmental trajectories, examining ideological stances), and from a variety of theoretical and empirical perspectives (e.g., qualitative and quantitative; social and cognitive; idiosyncratic and society-wide). Below we focus on the trajectory of but one subset of data analysis, namely, the analysis of large datasets.

Prior to the advent and popularization of personal computers, spoken and written discourse had to be analyzed by hand. Such practices stretch back at least to the middle ages when monks would cross-reference the use of words across biblical passages (McEnery & Hardie, 2011). In more modern times, researchers have gathered relatively large collections of texts to analyze the nuances of a particular language, either for theoretical or for pedagogical purposes. Thorndike and Lorge (1944), for example, manually counted word frequencies in order to help teachers prioritize particular vocabulary items at appropriate grade levels. For example, the 1000 most frequent words were suggested for students in first and second grades and the 2000 most frequent words for those in third grade. Such analyses were relatively rare due to the time and cost involved in manually deriving counts. In the late 1970s, as computers began to become more readily available to researchers, corpora were computerized, allowing for analyses such as frequency counts and concordancing to be completed with relative ease (provided one had access to a mainframe computer). Concordancing (i.e., examining the contexts in which words appear) gave rise to new ways of analyzing language systems. Instead of looking at frequencies of isolated words, one could analyze probabilities of word co-occurrences (i.e., collocation) and see larger patterns. Such analyses led to corpus-based accounts of English such as the *COBUILD* dictionary (Sinclair, 1987) and lexical approaches to grammar such as *Pattern Grammar* (Hunston & Francis, 2000). The rise of personal computers in the late 1980s made relatively simple computerized analyses such as frequency counts and concordancing even easier—one could then create frequency lists or generate concordance lines from the comfort of one's own home or office.

The development of more sophisticated computers and software packages allowed for more complex linguistic analyses that went beyond strings of words. Some corpora, such as the Brown corpus (Kucera & Francis, 1967), were semiautomatically annotated with part-of-speech (POS) and syntactic constituency tags. This allowed for more fine-grained analysis at the lexical level (i.e., the verb *run* could be disambiguated from the noun *run*) and also allowed for syntactic analysis. Such annotation made fine-grained linguistic analyses, such as Römer's (2005) investigations of the use of progressives to reference the future time. Römer's research, which used concordance lines from POS annotated texts, demonstrated that the use of progressives to reference the future is common and may be useful in English language pedagogy, challenging traditionally held accounts of the future aspect. While POS annotation added to the flexibility of computerized analyses of texts, its uses were initially limited by time and resources—automatic tagging error rates were around 35% (Marcus, Marcinkiewicz, & Santorini, 1993) and required a great deal of costly manual annotation.

With each technological advancement has come more flexibility and intellectual independence. A single concordance analysis of a relatively small corpus such as Lorge's corpus of magazine articles (Thorndike & Lorge, 1944) may have required an entire team of researchers 60 years ago and access to a mainframe computer 50 years ago. Today much more powerful analyses can be conducted by anyone with access to an inexpensive computer using NLP tools developed specifically for the analysis of electronic texts.

## Natural Language Processing Tools

An alternative to hand-coding data and qualitative analyses using frequency lists and concordance data is the use of NLP tools to explore spoken and written discourse. Such tools are now in common use and their merits are such that their popularity will continue to grow. These merits include their speed of analysis, their cost benefits, and their accuracy in measuring language features within a text (Burstein, 2003; Higgins, Xi, Zechner, & Williamson, 2011; Myers, 2003). While NLP tools are not new, their use has been affected by developments in corpus linguistics, developments in computational linguistics, improved computational power, and the availability of both digitized texts and tools to researchers. Below we provide an overview of a number of NLP tools used in the 1990s and early 2000s and their applications. Many of these tools are freely available, some are available for a small cost, and some are only available to a small core of researchers.

### The Biber Tagger

The Biber Tagger was introduced in Biber (1988). This version tagged texts for 67 different linguistic features. A more recent version of the tagger includes tags for 131 linguistic features (Biber et al., 2004), including a number of semantic features.

To date, the tagger has been used in a number of studies that have generally employed multidimensional analyses to examine register differences (Friginal, 2013). The Biber Tagger has also been used to examine spoken and written discourse in terms of second language learning and acquisition. For instance, Grant and Ginther (2000) used the Biber Tagger to analyze different proficiency levels of second language (L2) writing. Using features related to lexical proficiency (type/token ratio and word length), lexical features (e.g., conjuncts, hedges, amplifiers, and emphatics), grammatical features (e.g., nouns, verbs, nominalizations, and modals), and clause level features (e.g., subordi-

nations, complementation, and passives), they found that as the level of the learner increased, so did the type/token ratio, average word length, conjuncts, amplifiers, emphatics, demonstratives, and downtoners. In addition, Biber, Gray, and Staples (2014) used the Biber Tagger to investigate spoken and written discourse in the language produced by TOEFL iBT test-takers. Biber et al. reported that test-takers demonstrated variation in the linguistic styles across both spoken and written discourse and independent and integrated testing tasks regardless of proficiency level. As an example, Biber et al. found that test-takers used more colloquial elements of language in speech and used more complex grammatical features in writing (i.e., long words, passives, and nominalizations). However, Biber et al. found few differences in the linguistic features produced by test-takers with different scores.

A major limitation of the Biber Tagger is accessibility. It was created and has been maintained by Doug Biber and is not publically available. Rather, the Biber Tagger is almost uniquely applied by Biber and/or his students and colleagues. Biber is willing to process texts for interested parties (Friginal & Weigle, 2014), but one must allow time for his research team to do so.

## Coh-Metrix

Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) was a computational tool developed in the early 2000s to investigate text cohesion and text difficulty at various levels of language, discourse, and conceptual analysis. It was expanded to include elements of lexical sophistication and syntactic complexity. At the time of its development, advances in various disciplines had made it possible to computationally investigate various measures of text and language comprehension that went beyond surface components of language (e.g., individual words) and allowed for the exploration of deeper, more global attributes of language than earlier tools. This advancement resulted in broader, more accurate, and more detailed analyses of language.

Coh-Metrix integrates lexicons, pattern classifiers, POS taggers, syntactic parsers, shallow semantic interpreters, and other components that had been developed in the field of computational linguistics (Jurafsky & Martin, 2008). Utilizing these resources, Coh-Metrix was able to analyze texts on several dimensions of cohesion including co-referential cohesion, causal cohesion, density of connectives, latent semantic analysis metrics, and syntactic similarity. It also includes several lexical metrics such as word frequency, concreteness, polysemy, word meaningfulness, hypernymy, word age-of-acquisition scores, word imageability, and word familiarity measures (Graesser et al., 2004) and syntactic features (e.g., POS tagging and sentence complexity measures). Unlike the Biber

Tagger, a version of Coh-Metrix was made freely available to researchers through an online interface. Due to its use of cutting-edge technology and availability, Coh-Metrix has helped to revolutionize NLP analyses in a number of fields beyond computational linguistics including composition studies, cognitive science, education, and applied linguistics.

An early adoption of Coh-Metrix to applied linguistic studies involved using the indices to examine differences between simplified and authentic texts (Crossley, Allen, & McNamara, 2012; Crossley, Louwse, McCarthy, & McNamara, 2007; Crossley & McNamara, 2008). The research found that simplified texts were characterized by higher levels of semantic and referential cohesion and lower lexical sophistication, providing evidence for the use of simplified text in the language classroom. Coh-Metrix has also been used to examine L2 writing proficiency. In an early study, Crossley and McNamara (2012) used Coh-Metrix indices to predict writing quality in essays written by high school students in Hong Kong. Crossley and McNamara found the primary predictors of essay quality were related to lexical sophistication and that cohesion indices were negatively related to essay quality. Follow-up studies (Crossley & McNamara, 2014; Guo, Crossley, & McNamara, 2013) reported similar findings, promoting a better understanding of successful L2 writing.

A number of studies have also used Coh-Metrix indices to investigate longitudinal development in the spoken utterances of L2 learners. In a series of studies, Crossley and his colleagues demonstrated how lexical features developed over time in L2 learners. For instance, their research demonstrated that L2 learners begin to produce more abstract words (Crossley, Salsbury, & McNamara, 2009), greater semantic relations (Crossley, Salsbury, & McNamara, 2010), and more imageable, familiar, and meaningful words (Salsbury, Crossley, & McNamara, 2011) as a function of time studying English.

While Coh-Metrix was made freely available to the public, its use was somewhat limited outside of the tool's developers for two reasons. First, the online tool did not allow for batch processing (i.e., each text had to be individually uploaded to the system), which effectively limited the scale of analysis. Second, the online version of Coh-Metrix was a pared-down version of an internal version available to a core number of Coh-Metrix researchers. Those researchers had access to an internal desktop package that allowed batch processing of texts (drastically decreasing text processing time) and that reported on hundreds of additional linguistic features that went beyond the features available in the online tool. Despite these limitations, Coh-Metrix has had an important and wide-reaching impact on NLP analyses in the social sciences in general and educational psychology and applied linguistics in particular.



## L2 Syntactic Complexity Analyzer

The L2 Syntactic Complexity Analyzer (SCA; Lu, 2010, 2011) focuses on the syntactic indices outlined in Wolfe-Quintero, Inagaki, and Kim's (1998) review of complexity indices employed in L2 development studies. SCA uses the Stanford Parser (Klein & Manning, 2003) and Tregex (Levy & Andrew, 2006) to count instances of eight structures (e.g., clauses, dependent clauses, verb phrases). These counts, along with text word counts, are then used to produce 14 indices of syntactic complexity that SCA calculates (e.g., mean length of clause, mean length of t-unit). SCA is freely available and works on any system equipped with the Python programming language and Java, allowing for near universal use (with a relatively small amount of programming knowledge). SCA can also process files in batches, allowing for large datasets to be analyzed in a relatively short period of time.

SCA has been used in a number of studies related to syntactic complexity. Lu (2011) used SCA indices to distinguish between intermediate and upper intermediate L2 writing samples and found that more advanced writing showed increased syntactic complexity (e.g., increased clause per t-unit). Ai and Lu (2013) used SCA to compare the syntactic complexity of native and non-native university student writing and found that native speaker writing at the university level tended to be more complex than non-native speaker writing. Yang, Lu, and Weigle (2015) used SCA to explore the effects of writing topic on syntactic complexity. They found that global measures of syntactic complexity were positively related to writing quality scores and were not affected by topic. Local measures of syntactic complexity (i.e., the use of coordinate phrases, complex nominals, and subordination), however, were affected by essay topic.

## Linguistic Inquiry and Word Count

Unlike the Biber Tagger, Coh-Metrix, and SCA, Linguistic Inquiry and Word Count (LIWC) is a sentiment analysis tool. LIWC was designed to capture conscious and unconscious psychological phenomena related to cognition, affect, and personal concerns. As a result, it has been widely used by sociologists, psychologists, computer scientists, and linguistics in a number of education and social media domains (Hutto & Gilbert, 2014; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007; Pennebaker, Francis, & Booth, 2001). LIWC is available for a small fee (\$90 at the time of writing) and, once purchased, is housed on the user's hard drive, allowing for secure data processing



in the absence of an internet connection. The LIWC dictionary is proprietary and contains about 4500 words. The word lists that comprise the LIWC dictionary include words compiled from previous lists, thesauri, and dictionaries and confirmed as construct-relevant by three to six independent judges. The initial word lists were refined through corpus analysis and new lists were added (Pennebaker et al., 2007; Tausczik & Pennebaker, 2010). LIWC has been used in hundreds of studies to investigate constructs such as social status (Sexton & Helmreich, 2000), deception (Newman, Pennebaker, Berry, & Richards, 2003), and personality differences (Mehl, Gosling, & Pennebaker, 2006). LIWC is not sensitive to POS tags and does not make use of valence markers such as negations. The LIWC software provides information on the percentage of words per text that are covered by its internal dictionary and the percentage of words per text in each of the 80 categories on which it reports.

LIWC does not have a history of use similar to the Biber Tagger, Coh-Metrix, or the SCA in applied linguistics. In a rare use, Jung, Crossley, and McNamara (2015) used LIWC along with Coh-Metrix indices to examine links between language features and human scores of essay quality in the MELAB writing test. Jung et al. reported that LIWC indices related to spatiality showed a positive correlation with essay quality ( $r = 0.252$ ) while simple verbs reported a negative correlation ( $r = -0.136$ ).

## The Suite of Automatic Linguistic Analysis Tools

There are a number of limitations to the tools discussed above. For instance, many of these tools are not publically available (e.g., the Biber Tagger or the desktop version of Coh-Metrix) or, if available to the public, limited in the use by a number of practical concerns, chief among them the number of indices available, the speed at which texts could be processed, the tool's usability, and the cost of the tools. In the last few years, though, a suite of tools known as the Suite of Automatic Linguistic Analysis Tools (SALAT; Crossley, Kyle, & McNamara, 2016a, b; Kyle, 2016; Kyle & Crossley, 2015) that address these limitations has become available. We discuss the tools that comprise the suite below.

Crossley and Kyle developed SALAT to address a number of the limitations associated with the tools discussed above (i.e., usability, portability, accessibility, and linguistic coverage). Four tools included in the suite provide updated measures of lexical sophistication (TAALES), (Kyle & Crossley, 2015; Kyle, Crossley, & Berger, *in press*), cohesion (TAACO) (Crossley, Kyle, & McNamara, 2016b), syntax (TAASSC) (Kyle, 2016), and sentiment analysis (SEANCE) (Crossley,

Kyle, & McNamara, 2016a). The tools are freely available, easy to use, are housed on user hard drives (freeing the systems from dependence on the internet), work on most operating systems (Windows, Mac, Linux), and allow for batch processing of text files. These tools are discussed below.

## TAACO

TAACO (Crossley, Kyle, & McNamara, 2016b) incorporates over 150 classic and recently developed indices related to text cohesion. For a number of indices, the tool incorporates a POS tagger from the Natural Language Tool Kit (Bird, Klein, & Loper, 2009) and synonym sets from the WordNet lexical database (Miller, 1995). The POS tagger affords the opportunity to look at content words (i.e., nouns, verbs, adjectives, and adverbs) as well as function words (i.e., determiners, propositions). TAACO provides linguistic counts for both sentence and paragraph markers of cohesion and incorporates WordNet synonym sets. Specifically, TAACO calculates type/token ratio (TTR) indices (for all words, content words, function words, and n-grams); sentence overlap indices that assess local cohesion for all words, content words, function words, POS tags, and synonyms; paragraph overlap indices that assess global cohesion for all words, content words, function words, POS tags, and synonyms; and a variety of connective indices such as logical connectives (e.g., *moreover*, *nevertheless*), conjuncts (*however*, *furthermore*), and the incidence of *and*.

TAACO was initially tested on a corpus of L1 writing samples that had been scored by expert raters in terms of text coherence and overall essay quality (Crossley, Kyle, & McNamara, 2016b). The findings of this study provided evidence that expert judgments of text coherence were either not predicted by local cohesion indices or negatively correlated local cohesion indices, supporting previous research (Crossley & McNamara, 2010, 2011). In contrast, expert ratings of coherence were positively predicted by and correlated with global indices of cohesion (i.e., indices based on paragraph level links such as noun overlap between adjacent paragraphs) supporting previous research (Crossley & McNamara, 2011). In terms of expert ratings of overall essay quality, the study reported that overall text cohesion devices (i.e., TTT indices) were negatively predicted by and negatively correlated with writing proficiency. In contrast, global cohesion indices positively predict essay quality. An analysis of student completion in massive open online classes (MOOCs) (Crossley, Paquette, Dasalu, McNamara, & Baker, 2016) demonstrated that language produced by students in MOOC forums was indicative of course completion

such that students whose posts contained greater global cohesion indices (as reported by TAACO) were more likely to complete the course.

## TAALES

TAALES 2.0 (Kyle et al., [in press](#); Kyle & Crossley, [2015](#)) incorporates over 200 indices related to basic lexical information (i.e., the number of words and n-grams, the number of word and n-gram types), lexical frequency (i.e., how many times an item occurs in a reference corpus), lexical range (i.e., how many documents in which a reference corpus an item occurs), psycholinguistic word information (e.g., concreteness, familiarity, meaningfulness), academic language (i.e., items that occur more frequently in an academic corpus than in a general use corpus) for both single words and multi-word units (n-grams such as bigrams and trigrams), age of exposure, semantic lexical relations (hypernymy and polysemy), strength of words associations, contextual distinctiveness of words, word neighbor information, and lexical decision times. TAALES draws on a number of databases to assess lexical sophistication including the British National Corpus (BNC) (BNC Consortium, [2007](#)), the SUBTLEXus corpus of subtitles (Brysbaert & New, [2009](#)), the Medical Research Council (MRC) psycholinguistic database (Coltheart, [1981](#)), the Academic Word List (AWL) (Coxhead, [2000](#)), the Academic Formulas List (Simpson-Vlach & Ellis, [2010](#)), and the English Lexicon Project (ELP), a large publicly available psycholinguistic dataset (Balota et al., [2007](#)). The ELP indices include word recognition norms and word neighborhood information.

TAALES has been used to model lexical sophistication in a number of studies related to second language assessment (e.g., Kyle & Crossley, [2015](#); Kyle, Crossley, & McNamara, [2016](#)) and second language development (Crossley, Kyle, & Salsbury, [2016](#)). These studies have reported that indices related to n-gram frequency, word range, word familiarity scores and word meaningfulness scores explained approximately 52% of the variance in lexical proficiency ratings and that indices related to trigram frequency, word range, academic words, word familiarity, and word frequency explained approximately 52% of the variance in speaking proficiency scores (Kyle & Crossley, [2015](#)). In addition, independent TOEFL speaking responses could be accurately distinguished from integrated responses in that responses to integrated were more lexically sophisticated than responses to independent tasks (Kyle et al., [2016](#)). Crossley, Kyle, and Salsbury ([2016](#)) used lexical property measures taken for TAALES to examine links between second language (L2) lexical input and output in terms of word information properties (i.e., lexical

salience) and found that similar linear trends were reported in L1 input and L2 output over time such that words with lower concreteness, lower familiarity, and lower meaningfulness were produced over the course of the study.

## TAASSC

TAASSC (Kyle, 2016) is a recently developed tool that measures large and fine-grained clausal and phrasal indices of syntactic complexity and usage-based frequency/contingency indices of syntactic sophistication. TAASSC includes 14 indices measured by Lu's (2010, 2011) SCA, 31 fine-grained indices related to clausal complexity, 132 fine-grained indices related to phrasal complexity, and 190 usage-based indices of syntactic sophistication. The fine-grained clausal indices calculate the average number of particular structures per clause and dependents per clause. The fine-grained phrasal indices measure seven noun phrase types and ten phrasal dependent types. The syntactic sophistication indices are grounded in usage-based theories of language acquisition (Ellis, 2002; Goldberg, 1995; Langacker, 1987) and measure the frequency, type/token ratio, attested items, and association strengths for verb-argument constructions (VACs) in a text.

TAASSC has been used to measure syntactic features in areas such as L2 writing and L2 development. Kyle (2016) found that L2 writing quality was best explained by indices related to phrasal complexity (e.g., dependents per nominal) and syntactic sophistication (e.g., verb-VAC strength of association). The findings indicate that essays that include more complex noun phrases and use more strongly associated verb-VAC combinations tend to earn higher writing proficiency scores. Kyle (2016) also reported that mean length of t-unit increased as a function of time in both an EFL and ESL longitudinal corpus, supporting previous related research (e.g., Lu, 2011; Ortega, 2003). The results also indicated that verb-VAC frequency decreased as a function of time across both corpora, suggesting that as individuals spend time learning English, they learn (and use) fewer frequent verb-VAC combinations. This latter finding supports usage-based perspectives on language learning (e.g., Ellis, 2002).

## SEANCE

SEANCE (Crossley, Kyle, & McNamara, 2016a), the last tool we describe in the SALAT suite, is a sentiment analysis tool that relies on a number of

preexisting sentiment, social positioning, and cognition dictionaries. SEANCE contains a number of predeveloped word lists developed to measure sentiment, cognition, and social order. These lists are taken from freely available source databases such as SenticNet (Cambria, Havasi, & Hussain, 2012; Cambria, Speer, Havasi, & Hussain, 2010) and EmoLex (Mohammad & Turney, 2010, 2013). For many of these lists, SEANCE also provides a negation feature (i.e., a contextual valence shifter; Polanyi & Zaenen, 2006) that ignores positive terms that are negated. The negation feature, which is based on Hutto and Gilbert (2014), checks for negation words in the three words preceding a target word. For example, because *not* occurs within three words of *happy* in “He is not happy,” the occurrence of *happy* would not influence positive valence scores. SEANCE also includes the Stanford POS tagger (Toutanova, Klein, Manning, & Singer, 2003) as implemented in Stanford CoreNLP (Manning et al., 2014). The POS tagger allows for POS-tagged specific indices for nouns, verbs, and adjectives.

SEANCE was initially tested on two corpora of positive and negative consumer reviews, and the accuracy of SEANCE was compared to that reported by LIWC indices. In the first task (classifying positive and negative movie reviews), SEANCE indices performed on par with domain-specific classifiers and outperformed LIWC indices. In the second analysis (classifying multiple domains of [Amazon.com](http://Amazon.com) reviews), SEANCE indices performed slightly better than domain-specific classifiers and outperformed LIWC indices for the corpus as a whole and for the sub-corpora (i.e., book, DVD, electronics, and kitchen appliance reviews). Follow-up studies indicated that SEANCE indices were important predictors of boredom during writing (Allen et al., 2016), sentiment in game reviews (Secui et al., 2016), and student completion rates in MOOCs (Crossley, Paquette, Dascalu, et al., 2016).

## Sample Analysis

To illuminate how the NLP tools discussed above can be used to answer theoretical questions of interest to applied linguists, we present here a sample analysis that predicts human ratings of speaking proficiency in a large corpus using the lexical indices reported by TAALES. For this analysis, we use the NICT Japanese Learner English (JLE) Corpus, which is freely available at [https://alaginrc.nict.go.jp/nict\\_jle/index\\_E.html](https://alaginrc.nict.go.jp/nict_jle/index_E.html). The JLE corpus comprises 1281 transcribed speaking samples from an oral proficiency interview (OPI) test. Each sample also includes a proficiency level score based on the ACTFL-ALC standard speaking test guidelines. Speaking proficiency scores range

from 0 to 9, providing a wide coverage of proficiency levels. We demonstrate how to use TAALES, organize the data, control for statistical assumptions, and run a simple machine learning algorithm (in this case a linear regression model). This is only one potential analysis of the JLE corpus and one that does not analyze a number of fixed effects (e.g., gender and tasks). However, the analysis is suitable to provide an example of how NLP tools can be used in learning analytics research.

Prior to having TAALES process the speaking sample, the JLE texts need to be separated by speaker and cleaned of metadata so that only the text of interest is available. In this case, we want only the words spoken by the test-taker. There are a number of ways to do this, but we chose to use a Python script relying on regular expressions that removed all metadata followed up by hand-cleaning of the data. This script is available on the IRIS database ([iris-database.org](http://iris-database.org)). Once the text files are cleaned, they can be run through TAALES 1.4. Instructions for running texts through TAALES are available in the TAALES user files available at a number of websites. Fig. 25.1 shows the TAALES inter-

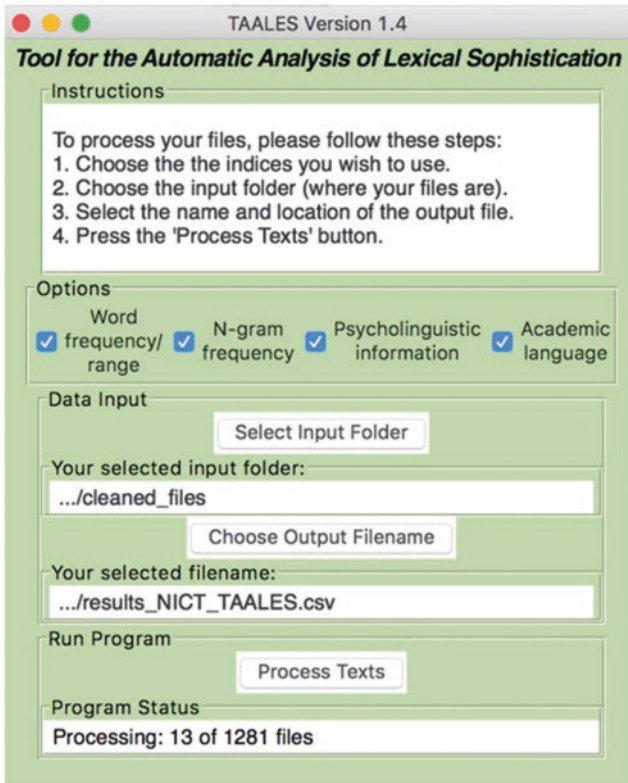


Fig. 25.1 TAALES GUI



	A	B	C	D	E	F	G	H
1	filename	SST_Level	nwords	KF Freq AW	KF Freq AW I	KF Ncats AW	KF Nsamp AV	TL Freq AW
2	file00001.txt	4	1470	5739.28865	2.91063953	13.546875	281.347039	28220.7744
3	file00002.txt	5	1146	5569.85547	2.90932577	13.4792774	281.343252	27963.036
4	file00003.txt	5	1164	7406.81061	2.95226808	13.7325702	276.408949	33006.6395
5	file00004.txt	4	1392	6884.2069	2.89716613	13.5037847	269.811606	30045.3316
6	file00005.txt	4	1432	5325.03641	2.85971957	13.4085258	275.360568	24007.4507
7	file00006.txt	6	1575	6097.50503	3.01800078	14.058908	291.615661	28789.8157
8	file00007.txt	6	1475	7197.71032	3.03473135	14.0261905	293.039683	31891.2308
9	file00008.txt	6	1625	7362.13173	3.01029521	13.9695467	289.550283	32515.1881
10	file00009.txt	6	2157	7051.50079	3.01984221	13.8546603	295.879937	31423.158
11	file00010.txt	4	1383	6673.20125	2.94629569	13.6696349	284.51024	30026.9448
12	file00011.txt	5	2359	6680.86688	2.83477187	13.1551363	274.240566	26537.6668
13	file00012.txt	5	1361	6022.65779	2.97163889	13.7893422	283.62448	28417.1315
14	file00013.txt	4	1409	8046.90878	2.95720193	13.722031	279.890706	36037.5792
15	file00014.txt	5	1274	6291.3497	2.94543477	13.828596	279.173127	29468.0725

Fig. 25.2 TAALES .csv file for analysis

face while running the JLE files. In short, the user selects the folder that contains the cleaned JLE files. The user then selects the location and the name of the output file produced by TAALES.

TAALES places the results into a .csv file that can be opened in a spreadsheet such as Excel or LibreOffice. Once the spreadsheet is open, the variables of interest (in this case the ACTFL-ALC standard speaking test scores) should be matched to the file names from the TAALES output (see Fig. 25.2).

Some TAALES indices occur rarely (e.g., the AWL sublist) and thus report a majority of zeros. Much of this data should be removed to avoid extremely non-normally distributed data which can influence cross-validation techniques, especially on small datasets. There are a number of ways to check for parametric data trends including statistical tests and skewness and kurtosis reports. Another way to check for gross violations of normality is to visualize the data via a histogram. This can be accomplished in a number of statistical and data visualization packages (see Egbert & LaFlair, Chap. 23, for a discussion on working with nonparametric and nontraditional data). For this example, we will use the machine learning software Waikato Environment for Knowledge Analysis (WEKA) (Hall et al., 2009). WEKA is a flexible and powerful software workbench that is freely available at <http://www.cs.waikato.ac.nz/ml/index.html>. WEKA can be used for data visualization and the creation of predictor models via a wide range of statistical and machine learning algorithms. To check the data for normality, we will first click on the “Explorer” button (see Fig. 25.3). We will then open the .csv file produced by TAALES (see Fig. 25.4). Each index analyzed by TAALES should now be listed (beginning with “filename”). To see each index visualized as a histo-



Fig. 25.3 WEKA explorer

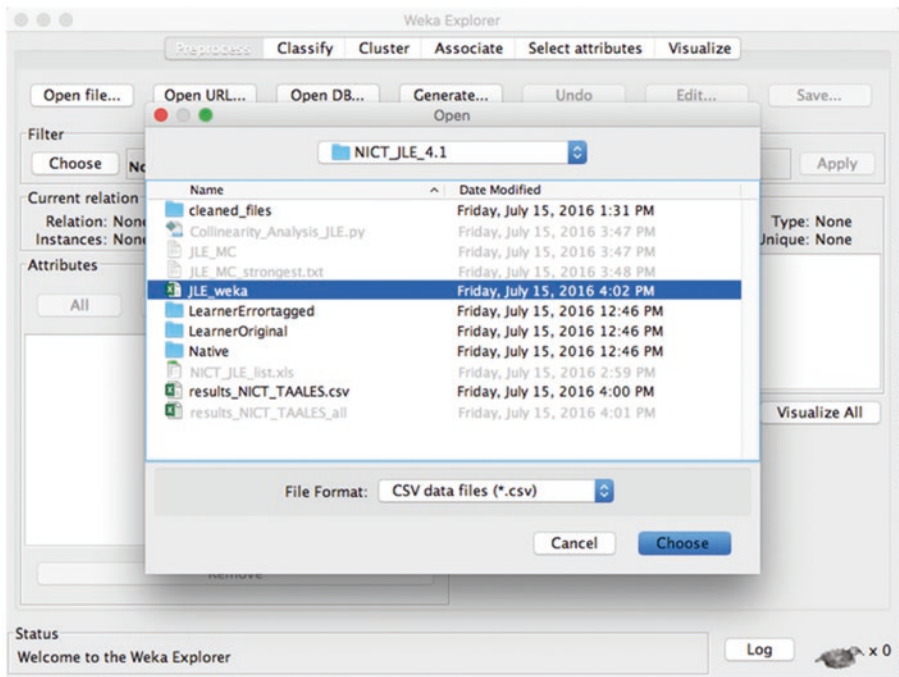


Fig. 25.4 File selection in WEKA



gram, click on the index name and look in the bottom right hand corner of the Explorer window. Normally distributed data will look roughly like a bell curve (see Fig. 25.5). Any indices that grossly violate normality, such as *AWL Sublist 1 Normed* (see Fig. 25.6) should be removed.

Next, we want to determine whether there are any meaningful relationships between our independent variables (i.e., our indices of lexical sophistication). To determine if there are meaningful relationships, we conduct Pearson correlations between our independent variables and our dependent variable (proficiency speaking score). This can be done in spreadsheet software such as Excel using the function `=correl(array1, array2)` where array1 is the first set of data to be correlated to the second set of data (array2). We set a criterion of  $r$  (absolute value)  $> 0.100$ , which represents the threshold for a small effect size (Cohen, 1988).

A problem with TAALES and a number of other NLP tools is that they calculate a number of features that may measure the same construct (e.g., frequency). For this reason, the indices need to be checked for multicollinear-

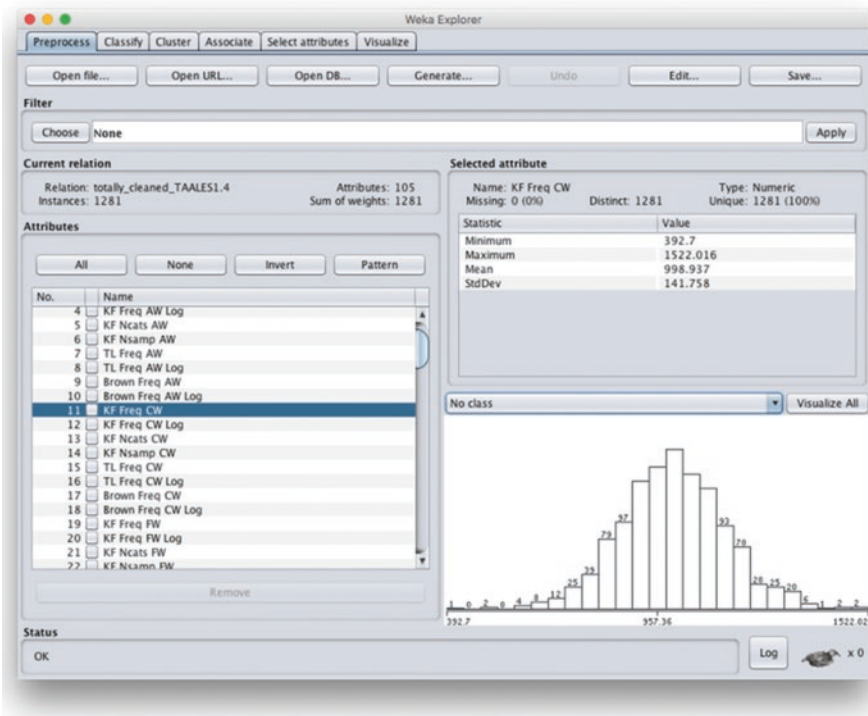


Fig. 25.5 Histogram for normally distributed TAALES index

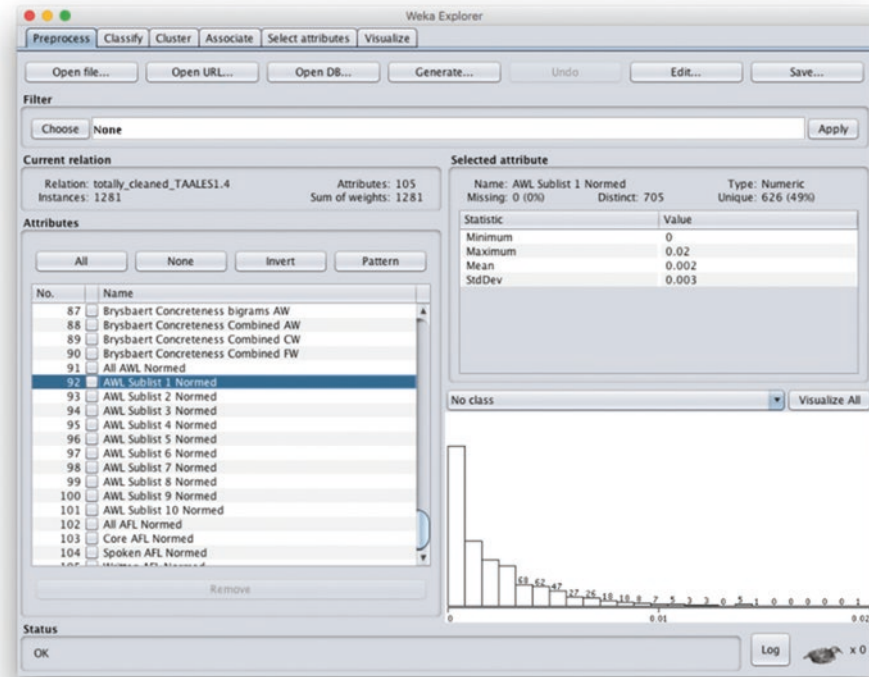


Fig. 25.6 Histogram for non-normally distributed TAALES data

ity to ensure they are not measuring similar features (Tabachnick & Fidell, 2014). This is usually done by examining a correlation matrix and checking to see which variables show a strong correlation (usually between  $r \geq 0.700$  and  $r \geq 0.900$ ). For this analysis, we used a conservative cutoff of 0.700. Correlation matrices can be created using most statistical packages, including SPSS, SAS, R, and even Microsoft Excel or LibreOffice Calc. If variables show strong multicollinearity, the variable that most strongly correlates with the criterion variable (the SST scores) will be retained while the other variable(s) is removed.

In this study, we found that 14 indices of lexical sophistication grossly violated normality. Of the 89 variables that were normally distributed, 15 failed to demonstrate a meaningful relationship with SST scores (i.e.,  $r \geq 0.100$ ). After checking for multicollinearity, 21 unique indices remained in our analysis (see Table 25.1). At this point, a researcher will want to control for overfitting the model by ensuring there is a ratio of 15 items (texts) to predictors (TAALES indices). This is to avoid random data reporting strong effects (i.e.,

**Table 25.1** Correlation between SST scores and variables entered into regression

Index name	<i>r</i>
BNC Spoken Trigram Proportion	0.697
MRC Imageability AW	-0.689
KF Number of Categories AW	0.519
Brysaert Concreteness Combined CW	-0.503
BNC Written Bigram Frequency Normed Log	0.483
MRC Concreteness FW	-0.465
BNC Spoken Range CW	0.454
BNC Written Frequency CW Log	0.450
MRC Meaningfulness CW	-0.445
Kuperman Age of Acquisition CW	0.425
BNC Spoken Bigram Normed	0.409
BNC Written Frequency AW	0.392
Brysaert Concreteness Combined FW	-0.342
SUBTLEXus Freq FW	-0.279
BNC Written Range FW	0.258
Kuperman Age of Acquisition FW	0.242
Brown Frequency AW	0.167
MRC Familiarity CW	-0.156
All Academic Formula List Normed	0.131
TL Frequency CW	0.129
BNC Spoken Frequency CW	0.113

AW, all words; CW, content words; FW, function words

noise in the data). With 1281 texts, we could use at least 85 TAALES indices, so 21 is well within the threshold.

Next we create a .csv spreadsheet that contains the 21 TAALES indices followed by the criterion variable. From the WEKA graphical user interface (GUI), select “Explorer” and then choose the file format .csv and select the file with the TAALES indices and the criterion. The user then clicks on the classify tab and for the type of classifier selects linear regression from the functions folder (see Fig. 25.7). It is best to use tenfold cross-validation set, which is the default (see Fig. 25.8). In a tenfold cross-validation multiple regression, the dataset is randomly divided into ten sections (called “folds”). A stepwise multiple regression is conducted using nine of the ten folds to train a statistical model, which is then tested on the remaining fold. This procedure is repeated nine more times until all of the folds have served as the test set and each of the ten models is averaged (Witten & Frank, 2005). Lastly, ensure that SST level is selected as the criterion variable (see Fig. 25.8) and press start. WEKA will automatically develop a regression model based on a tenfold cross-validation.

The initial model produced by WEKA includes 13 indices and explains 65% of the variance ( $r = 0.806$ ,  $R^2 = 0.649$ ) in SST scores (see Fig. 25.9). However, when we look at the model weights, we find that some indices (e.g.,

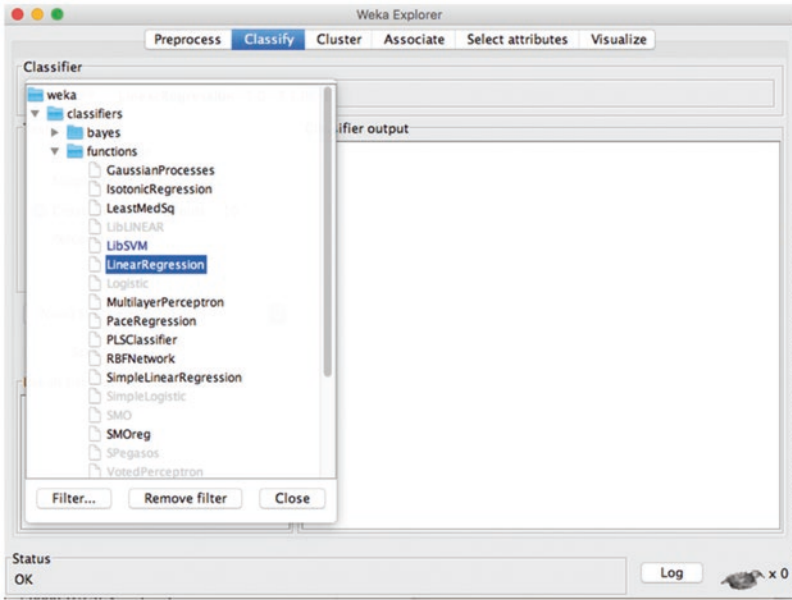


Fig. 25.7 Selection of model in WEKA

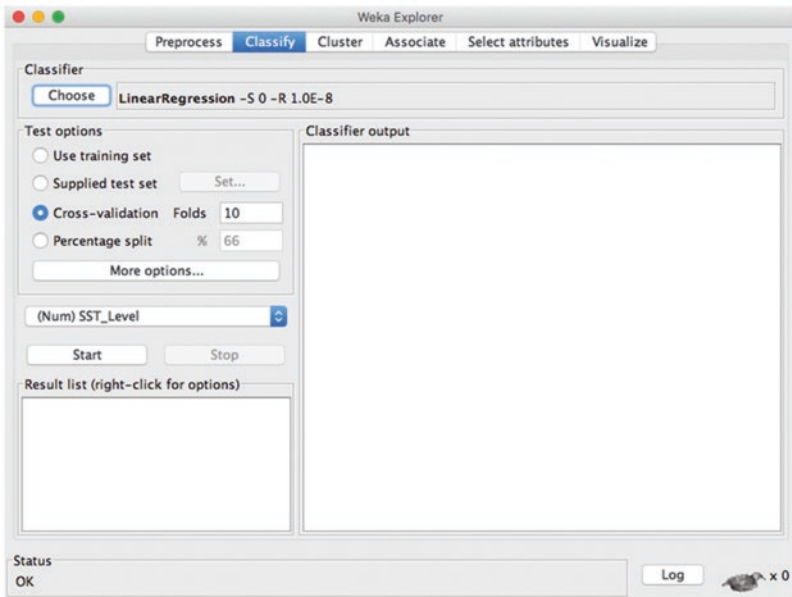


Fig. 25.8 Selection of cross-validation type in WEKA

```

=== Classifier model (full training set) ===

Linear Regression Model

score =

 18.2946 * BNC Spoken Trigram Proportion +
-0.0203 * MRC Imageability AW +
 0.3285 * KF Ncats AW +
-1.1703 * BNC Written Bigram Freq Normed Log +
-0.0091 * MRC Concreteness FW +
 0.068 * BNC Spoken Range CW +
 0.8876 * BNC Written Freq CW Log +
 0.7365 * Kuperman AoA CW +
 4.1092 * BNC Spoken Bigram Normed +
-0.0014 * Brown Freq AW +
-0.0824 * MRC Familiarity CW +
-14.5299 * All AFL Normed +
-0.8522 * BNC Spoken Freq CW +
 50.0903

Time taken to build model: 0.23 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient           0.8055
Mean absolute error              0.7283
Root mean squared error         0.9323
Relative absolute error         58.7374 %
Root relative squared error     59.2076 %
Total Number of Instances      1281

```

Fig. 25.9 Initial linear regression model reported in WEKA with suppression effects

Brown Freq AW) have been subjected to suppression (i.e., their model weight reports the opposite polarity of their correlation with SST score). In order to ensure that our predictor model reflects the relationship each index has with holistic score, suppressed variables need to be removed. Thus, we remove the suppressed indices and rerun the regression analysis. In this case, we ran the analysis four more times and removed a total of ten suppressed indices. The fifth and final model included nine variables (see Fig. 25.10) and explained 62% of the variance in SST scores ( $r = 0.790$ ,  $R^2 = 0.624$ ).

The final regression analysis as reported in Fig. 25.10 indicates that more proficient L2 speakers as judged by SST scores use a greater number of trigrams proportional to speakers in the BNC spoken section. In addition, more proficient speakers use words that are found in more categories in the Kucera Francis corpus and in a greater number of texts in BNC. Lastly, more proficient speakers use more sophisticated words. For instance, more proficient

```

=== Classifier model (full training set) ===

Linear Regression Model

score =

18.0483 * BNC Spoken Trigram Proportion +
 0.5601 * KF Ncats AW +
-0.7572 * Brysbaert Concreteness Combined CW +
-0.0171 * MRC Concreteness FW +
 0.0313 * BNC Spoken Range CW +
 1.3555 * BNC Written Freq CW Log +
 1.1051 * Kuperman AoA CW +
 0.5171 * Kuperman AoA FW +
-0.0865 * MRC Familiarity CW +
36.7508

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient           0.79
Mean absolute error              0.7499
Root mean squared error         0.9645
Relative absolute error         60.4841 %
Root relative squared error     61.2521 %
Total Number of Instances      1281

```

Fig. 25.10 Final linear regression model

speakers use fewer concrete words, words that are acquired later by children, and fewer familiar words. These findings support previous research using NLP tools which reported links between human judgments of lexical sophistication and speaking proficiency (Crossley & McNamara, 2014). However, the findings indicate that newer indices reported by TAALES have the propensity to increase our understanding of speaking proficiency, especially in terms of lexical items.

## Discussion and Conclusion

This chapter has provided an overview of how written and spoken discourse has been analyzed in the past for linguistic signatures of registers, domains, tasks, learner development, and other categorizations and criterion of interest to applied linguists. The majority of this chapter has focused on how newer approaches for examining written and spoken discourse, specifically through the use of NLP tools, can be used to address questions relevant to a variety of substantive domains of applied linguistics. Specifically, we have provided a

synopsis of available tools used in discourse research and, using a single tool, provided an example of how these tools can be used in conjunction with corpus linguistics, inferential statistics, and machine learning algorithms to better understand second language speaking proficiency. These tools have the capacity to greatly inform our understanding of discourse at multiple linguistic levels and provide researchers with a number of options for conducting analytic research in both second language research and in larger research areas of interest to applied linguists.

In second language learning specifically, the tools discussed in this chapter can be used to greatly increase our understanding of learner development on a number of linguistic metrics. For instance, as NLP tools become more fine-grained (see, e.g., TAASSC), language researchers will be able to examine linguistic features at levels that go beyond the text base (i.e., the information available at the surface level of text) and focus on underlying representations of knowledge such as verb-argument constructions, word association metrics, and psycholinguistic response data (e.g., reaction times; see Grey and Tagarelli, Chap. 14). Beyond the word and sentence level, NLP tools can also provide information at the discourse level and give researchers the opportunity to examine links between text structures that help develop coherent discourse.

The application of these tools can be used to enhance our understanding of language well beyond the scope of second language learning. The tools have the potential to revolutionize text analytics in a number of fields for which applied linguistics is of interest. With the advent of the worldwide web and digital communications such as e-mails, text messages, and Twitter, we are awash with text that can be used to explain human patterns of decision making, communication, and learning. Text analytics of this data will rely on NLP tools to seek meaningful knowledge arrays to explain human activity. This may include better understanding communication between physicians and patients in order to increase medical outcomes, expand online learning by providing students with better opportunities for deliberate practice aligned with formative feedback, increase threat detection to improve local and global security, better cater to consumer needs through targeted advertising and search engines, and develop more accurate artificial intelligence models to serve a variety of functions related to automated speech recognition and production.

Much of this will depend on linguistic research that investigates spoken and written discourse in order to provide a better understanding of how language comes to define types of discourse and proficiencies in discourse tasks. From these investigations we can better understand domain, register, and disciplinary differences. We can also gain important insights into how



language is organized, stored, and retrieved cognitively. Importantly, we can begin to understand how language influences peoples' responses and reactions to the variety of language available. On its own, this information will be important, but language in isolation is less meaningful than when it is combined with factors such as pragmatics, context, purpose, individual differences, and behavioral measurements. Thus, NLP approaches are only as successful as the diversity of approaches with which they are combined. Successful use of NLP approaches to understand discourse must therefore be mixed with interventions, surveys, affect detection, and environmental factors to understand the role of language when other fixed effects are taken into consideration. Such approaches will demand multidisciplinary collaboration between applied linguists, educational specialists, cognitive scientists, industry, and government. Collaborations between specialists where linguists are but one part of the equation will provide opportunities to increase the optics through which discourse is viewed such that larger appreciations for discourse in context and the effects that such discourse has on processing and action can be developed.

## References

- Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 249–264). Amsterdam: John Benjamins Publishing Company.
- Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., & McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: The essay, the writer, and keystrokes. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 114–123). Edinburgh: ACM.
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445–459. <https://doi.org/10.3758/BF03193014>
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., ... Urzua, A. (2004). Representing language use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. TOEFL Monograph Series. Retrieved from <http://www.ets.org/Media/Research/pdf/RM-04-03.pdf>
- Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, amu059. <https://doi.org/10.1093/applin/amu059>



- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol: O'Reilly Media, Inc.
- BNC Consortium. (2007). The British National Corpus, version 3. BNC Consortium. Retrieved from <http://www.natcorp.ox.ac.uk/>
- Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. <https://doi.org/10.3758/BRM.41.4.977>
- Burstein, J. (2003). The E-rater<sup>®</sup> scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In G. M. Youngblood & P. M. McCarthy (Eds.), *FLAIRS conference* (pp. 202–207). Palo Alto: Association for the Advancement of Artificial.
- Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). SenticNet: A Publicly Available Semantic Resource for Opinion Mining. In C. Havasi, D. Lenat, & B. Van Durme (Eds.), *AAAI fall symposium: commonsense knowledge* (Vol. 10).
- Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), 51–89.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4), 497–505.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.
- Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16(1), 89–108.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, 1–19.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of Text Cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227–1237.
- Crossley, S. A., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *Modern Language Journal*, 100(3), 702–715.
- Crossley, S. A., Louwse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, 91(1), 15–30.
- Crossley, S. A., & McNamara, D. S. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwse, McCarthy & McNamara (2007). *Language Teaching*, 41(3), 409–429.

- Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the Cognitive Science Society* (pp. 984–989). Austin, TX: Cognitive Science Society.
- Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1236–1241). Austin, TX: Cognitive Science Society.
- Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115–135.
- Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66–79. <https://doi.org/10.1016/j.jslw.2014.09.006>
- Crossley, S. A., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 6–14). Edinburgh: ACM.
- Crossley, S. A., Salisbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, 59(2), 307–334.
- Crossley, S. A., Salisbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60(3), 573–605. <https://doi.org/10.1111/j.1467-9922.2010.00568.x>
- Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, 24(2), 143–188. Retrieved from <https://doi.org/10.1017/S0272263102002024>
- Fairclough, N. (2013). *Critical discourse analysis: The critical study of language*. New York, NY: Routledge.
- Friginal, E. (2013). Twenty-five years of Biber's Multi-Dimensional Analysis: introduction to the special issue and an interview with Douglas Biber. *Corpora*, 8(2), 137–152.
- Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing*, 26, 80–95. <https://doi.org/10.1016/j.jslw.2014.09.007>
- Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193–202. <https://doi.org/10.3758/BF03195564>
- Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2), 123–145.

- Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18(3), 218–238.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1), 10. <https://doi.org/10.1145/1656274.1656278>
- Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25(2), 282–306. <https://doi.org/10.1016/j.csl.2010.06.001>
- Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
- Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth International AAAI Conference on Weblogs and Social Media.
- Jung, Y., Crossley, S. A., & McNamara, D. S. (2015). Linguistic features in MELAB writing performances.
- Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Englewood Cliffs, NJ: Prentice-Hall.
- Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics—ACL '03 (Vol. 1, pp. 423–430). <https://doi.org/10.3115/1075096.1075150>
- Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Kyle, K. (2016). Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Georgia State University. Retrieved from [http://scholarworks.gsu.edu/alesl\\_diss/35/](http://scholarworks.gsu.edu/alesl_diss/35/)
- Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786. <https://doi.org/10.1002/tesq.194>
- Kyle, K., Crossley, S. A., & Berger, C. (in press). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. Behavior Research Methods.
- Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, 33(3), 319–340. <https://doi.org/10.1177/0265532215587391>
- Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford: Stanford University Press.
- Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In 5th International Conference on Language Resources and Evaluation (LREC 2006).
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. <https://doi.org/10.1075/ijcl.15.4.02lu>

- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36–62. Retrieved from <http://www.jstor.org/stable/41307615>
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)* (pp. 55–60).
- Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: the penn treebank. *Computational Linguistics*, 19(2), 313–330. Retrieved from <http://dl.acm.org/citation.cfm?id=972470.972475>
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5), 862.
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39–41.
- Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26–34). Association for Computational Linguistics.
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436–465.
- Myers, M. (2003). What can computers and AES contribute to a K–12 writing program. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 3–20). Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.
- Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5), 665–675.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518.
- Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007: [LIWC.net](http://LIWC.net).
- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahwah: Lawrence Erlbaum Associates, 71.
- Polanyi, L., & Zaenen, A. (2006). *Contextual valence shifters*. In *Computing attitude and affect in text: Theory and applications* (pp. 1–10). Netherlands: Springer.
- Römer, U. (2005). Shifting foci in language description and instruction: Towards a lexical grammar of progressives. *Arbeiten Aus Anglistik Und Amerikanistik*, 30(1), 145–160.

- Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, 27(3), 343–360.
- Schiffrin, D. (1994). *Approaches to discourse*. Oxford, UK: Blackwell.
- Secui, A., Sirbu, M.-D., Dascalu, M., Crossley, S., Ruseti, S., & Trausan-Matu, S. (2016). Expressing Sentiments in Game Reviews. In *In International Conference on Artificial Intelligence: Methodology, Systems, and Applications* (pp. 352–355). Varna, Bulgaria: Springer.
- Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: the links between language, performance, error, and workload. *Journal of Human Performance in Extreme Environments*, 5(1), 6.
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512.
- Sinclair, J. M. (1987). *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. London: Collins ELT.
- Tabachnick, B. G., & Fidell, L. S. (2014). *Using Multivariate Statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.
- Thorndike, E. L., & Lorge, I. (1944). *The teacher's wordbook of 30,000 words*. New York: Columbia University, Teachers College. Bureau of Publications.
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—NAACL '03 (Vol. 1, pp. 173–180). Morristown, NJ, USA: Association for Computational Linguistics. <https://doi.org/10.3115/1073445.1073478>
- Witten, I. H., & Frank, E. (2005). *Data mining practical machine learning tools and techniques*. Amsterdam; Boston, MA: Morgan Kaufman. Retrieved from <http://public.eblib.com/choice/publicfullrecord.aspx?p=234978>
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & Complexity*. Honolulu, HI: University of Hawaii Press.
- Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 53–67. <https://doi.org/10.1016/j.jslw.2015.02.002>



# 26

## Narrative Analysis

Phil Benson

### Narrative Inquiry in Applied Linguistics Research

Narrative inquiry refers to a variety of approaches to research in the humanities and social sciences that make use of stories of personal experience. While there are many approaches to data collection and data analysis in narrative inquiry, studies tend to share, as Pomerantz and Kearney (2012, p. 224) put it, “a similar philosophical foundation by appealing to the primordial meaning-making functions of narrative.” When we tell stories, we cast light on the meanings of our experiences. Storytelling has been described as a “universal human activity” (Riessman, 1993, p. 3) and as “the primary form by which human experience is made meaningful” (Polkinghorne, 1988, p. 1). In the postmodern era, narrative is increasingly seen as a means for individuals to make sense of their identities in contexts of mobility, displacement, and rapid social change (Brockmeier & Carbaugh, 2001; Giddens, 1991). In contexts of global mobility, language learning and teaching increasingly involve narrative identity work (Benson, 2004).

The first applied linguistics studies to refer explicitly to narrative methods appeared in the early years of the twenty-first century (Bell, 2002; Benson & Nunan, 2002; Pavlenko, 2002), but narrative inquiry has illustrious ancestors in the fields of psychology and sociology. Sigmund Freud, the founder of psychoanalysis and a reader of Sherlock Holmes mystery stories (Brooks, 1979),

---

P. Benson (✉)

Department of Linguistics, Macquarie University, North Ryde, NSW, Australia  
e-mail: [philip.benson@mq.edu.au](mailto:philip.benson@mq.edu.au)



used dream narratives as research data and often presented his work in the form of narrative case studies (e.g., Rieff, 1996). The foundational work of the Chicago School of Sociology also attached great value to individual biographies. Thomas and Znaniecki's (1919) work on the migration of Polish peasants to America included an individual biographical case study of more than 300 pages. Describing "life records" as the "perfect type" of sociological data, Thomas and Znaniecki argued that the use of other types of data was "a defect, not an advantage, of our present sociological method" (pp. 6–7).

Appreciation of the value of storytelling in psychology and sociology suffered under the pressure of "scientific" experimental and survey methods during much of the twentieth century, but stories returned to the fore with the launch of the journal *Narrative Inquiry* in 1990 and the publication of a number of important texts addressing narrative methods (e.g., Atkinson, 1998; Chamberlayne, Bornat, & Wengraf, 2000; Clandinin & Connelly, 2000; Josselson & Lieblich, 1993; Lieblich, Tuval-Mashiach, & Zilber, 1998; Polkinghorne, 1995; Riessman, 1993). Holstein and Gubrium (2012) is a recent overview of the current status of narrative inquiry in the social sciences. The impact of this narrative turn on applied linguistics was somewhat delayed, as has been observed with other analytical techniques in applied linguistics (e.g., regression models; see Plonsky & Oswald, 2017), although narrative has arguably played an important role in the field since the 1970s in studies that have used diary, longitudinal, ethnographic, and case study methods (e.g., Schmidt & Frota, 1986; Schumann & Schumann, 1977).

Narrative approaches have gained ground in applied linguistics for two main reasons. First, researchers have questioned the emphasis in applied linguistics language acquisition research on the search for universal cognitive acquisition processes, arguing instead for examination of language learning in its social contexts and environments. Although socially oriented approaches to applied linguistics are often critical of the "individualism" of cognitive accounts of language learning, individual case studies have emerged as their research method of choice (Benson & Cooker, 2013). The focus on lived experience in narrative inquiry appears to be especially suited to approaches such as language learning ecology, where the goal is to explicate how language learning emerges from interactions between the learner and the learning environment (Menezes, 2011; Palfreyman, 2014; van Lier, 2004).

Second, narrative inquiry offers an approach to the presentation of research findings that is both accessible and convincing to many readers of applied linguistics research. In this respect, it is important to note that language teachers represent one of the primary audiences of applied linguistics studies, which often struggle to make their findings relevant to teachers' lives and work.

Narrative studies of teachers' lives and work in the field of education (Clandinin & Connelly, 2000; Goodson & Sikes, 2001) were an important stimulus for narrative studies of language teaching. Applied linguistics has made its own contribution to narrative inquiry in education through studies of the experiences of language *learners* (Benson & Nunan, 2004; Kalaja, Menezes, & Barcelos, 2008). Narrative inquiry appeals not only to researchers who have "become weary of variables and the quantification of the positivistic approach" (Josselson & Lieblich, 1993, p. xv) but also to broader audiences for research who are more likely to be convinced by a well-told story than they are by statistical outcomes.

Since their formal introduction into applied linguistics, narrative inquiry methods have been used in numerous data-based studies, and there is now a rich literature on methodological issues specific to applied linguistics (Barkhuizen, 2011, 2013, 2014; Barkhuizen, Benson, & Chik, 2013; Bell, 2002; Benson, 2004, 2014; Murray, 2009; Pavlenko, 2002, 2007). Barkhuizen (2014) provides a timeline of the development of narrative research on language teaching and learning. Barkhuizen et al. (2013) is an entry-level introduction to the use of narrative methods based on applied linguistics studies, and Benson (2014) reviews more recent methodological developments. This chapter adds to this earlier work by focusing on narrative *analysis* or the use of narrative writing as a means of analyzing research data and reporting research findings.

## Three Approaches to Narrative Research

Narrative inquiry is a broad term that can best be defined as any approach to research that makes use of stories or storytelling. Narrative research can be defined similarly, and in this sense, both are catchall terms that elude precise definition (Barkhuizen, 2014). Polkinghorne (1995) identified two broad approaches to narrative research, which he called *analysis of narratives* and *narrative analysis*. Here, I add a second distinction in order to identify three main approaches in current applied linguistics studies: "*content analysis of narratives*," "*discourse analysis of narratives*," and "*narrative analysis*." I will describe the first two approaches briefly in order to clarify what I mean by narrative analysis, which is the main focus of this chapter.

Many narrative studies take stories as their object of inquiry or source of data and then subject them to some form of content or thematic analysis. These are studies for which I use the term *content analysis of narratives*. Takeuchi (2003) systematically analyzed references to learning strategies in 67



books written by successful foreign language learners. This is an example of content analysis of published narratives or language learning “memoirs.” There are also many studies in the applied linguistics literature based on original narrative data, elicited in the form of written language learning histories (e.g., Menezes, 2011; Murphey & Carpenter, 2008), narrative frames (e.g., Barkhuizen & Wette, 2008; Xu, 2014), interviews (e.g., Kouritzin, 2000; Menard-Warwick, 2004), and visual materials (e.g., Chik, 2014; Nikula & Pitkänen-Huhta, 2008). These studies typically use thematic analysis methods from qualitative research and differ from qualitative research only in their use of narratives as data. Moodie (2016, p. 32) uses the term “grounded narrative inquiry” to describe his use of grounded theory (Charmaz, 2014) to analyze themes in Korean English language teachers narratives of their prior learning experiences.

Discourse analysis of narratives is, in part, a response to perceived problems in qualitative studies that rely on content analysis of narrative data. A recent reflexive turn in qualitative applied linguistics research has called into question an apparently naïve reliance on research participants’ accounts of experience as objective accounts of events, processes, or psychological states (Kasper & Prior, 2015; Mann, 2011; Pavlenko, 2007). This has led some researchers to focus more on the discourse of narratives and, in particular, their meanings in local contexts of interaction (De Fina & Georgakopoulou, 2012). In such studies, discourse and interaction analysis methods (see Crossley & Kyle, Chap. 25, on analysis of written and spoken language, and Miller, Chap. 27, on interactional analysis) have been used to gain insight into issues of language teaching and learning. The challenge in this approach is to avoid turning a study of language teaching or learning into one that is purely concerned with narrative in interaction. Barkhuizen (2010) and Simpson (2011) are good examples of studies that have overcome this difficulty and used discourse analysis of interview data to produce insights into the positioning of migrant teachers and learners in TESOL.

Analysis of narrative studies have made important contributions to applied linguistics research, but because they use data analysis methods that are discussed elsewhere in this volume, they are not discussed in detail in this chapter. Narrative analysis refers here to a distinct mode of narrative inquiry in which storytelling plays a significant role in the analysis of data and reporting of findings. Narrative analysis largely involves identifying narratives within collected data and representing findings as narratives in published work. While Polkinghorne (1995) assumes that narrative analysis begins with non-narrative data, the source data may also have a narrative character. Clandinin and Connelly (2000) use the term “restorying” for an approach in which storied

data are reanalyzed and reconstructed using storytelling as an analytical device. Barkhuizen (2013, p. 4) uses “narrative knowledging” to denote a multilevel approach to narrative inquiry that involves “making sense of an experience through narrating, analyzing narratives, reporting narrative research, and consuming research findings.”

Outcomes of narrative analysis in applied linguistics include first-person autobiographical accounts and third-person biographical case studies. In both cases, narrative writing is the key to narrative analysis as a research method. A recent search for articles involving narrative inquiry published between January 2014 and June 2016 in four language teaching journals (*ELT Journal*, *Language Teaching Research*, *The Modern Language Journal*, and *System*) produced 23 articles: 12 used content analysis of narratives, 2 used discourse analysis of narratives, and 9 used narrative analysis. In earlier reviews, I have observed that narrative analysis is infrequently used in applied linguistics (Benson, 2013, 2014), but the number of studies published in recent journals suggests that narrative analysis approaches may now be gaining ground.

## Autobiographical and Biographical Studies

A published narrative analysis study is always someone’s story. In applied linguistics research, this “someone” may be the author of the study (viewed as language teacher or learner) or a teacher or learner from whom the author has elicited data. In general, research narratives are not fictions, but interpretive accounts of lived experiences. For Polkinghorne (1995, p. 5), the outcome of narrative analysis might be “a historical account, a case study, a life story, or a storied episode of a person’s life.” Although the stories that appear in applied linguistics studies can cover relatively short episodes, one of the strengths of narrative research is its capacity to provide access to long-term experiences and act as a means of representing the coherence of such experiences through narrative writing. The narrative turn in the social sciences has also been called the *biographical turn* (Chamberlayne et al., 2000) and in order to emphasize the focus on language teaching and learning as lived experience, Benson (2004) used the term *(auto)biographical* research.

The word *(auto)biographical* covers two types of studies. In first-person autobiographical studies, researchers study and write about their own language teaching or learning experiences. Casanave (2012), for example, used diaries and retrospection to research and write about her own experiences of “dabbling” as an expatriate teacher learning Japanese in Japan. In third-person biographical studies, researchers study and write about the language teaching

or learning experiences of others. Liu and Xu (2011), for example, used interviews, a reflective journal, and written reflection reports to construct the story of one teacher's experience of educational reform in China. There are also examples of dialogic and multivoiced studies in which researchers and participants are listed as co-authors. In Murray and Kojima (2007), the second author (a highly autonomous adult language learner) tells the story of how she learned English and German, and this is followed by the first author's analysis of the strategies she used.

## **Memoirs, Autoethnography, and Autobiographical Studies**

In autobiographical research, the researchers study and write about their own experiences. Studies of this kind are also called memoirs and autoethnographies. A language learning memoir is an informally written account of language learning experiences based on recollection and reflection. Hoffman's (1989) account of her loss of Polish and acquisition of English as she migrated from Poland to America as a child is one of the best-known nonacademic language memoirs, which have stimulated applied linguists to follow suit in collections of personal accounts of language learning and teaching experiences (e.g., Curtis & Romney, 2006; Harbon & Moloney, 2013; Johnson & Golombek, 2002; Lee & Sze, 2015; Nunan & Choi, 2010). As they are based largely on recollection and reflection, rather than formal collection and analysis of data, the research credentials of memoirs can be questioned. The counter-argument is that memoirs are research writing if they provide informed insight into issues of importance to the applied linguistics community. Autoethnography (Ellis & Bochner, 2000) represents a more formal approach to autobiographical research writing, and has been adopted by Canagarajah (2012) and Choi (2012) in papers that are based largely on recollection and reflection.

Autobiographical work can also be data based. Casanave (2012), for example, has recently published a narrative of her learning of Japanese while resident in Japan, based on a systematic analysis of extensive diary entries. Her paper shows that there is much unlocked potential for narrative research into applied linguists' efforts to learn additional languages. Autobiographical narrative has also been used in a number of recent papers analyzing and presenting findings related to experiences of teaching and professional development (e.g., López-Gopar, 2014; O'Móchain, 2006; Pinner, 2016; Wyatt & Márquez, 2016). These papers are often based on

reflective journals and observational data gathered in the course of action research or exploratory practice projects. Pinner's (2016) paper was published under the heading of "Practitioner research" in the journal *Language Teaching Research*. It is a good example of an autobiographical or autoethnographic study, in which the researcher analyses and writes about his own experiences.

### Sample Study 26.1

Pinner, R. (2016). Trouble in paradise: Self-assessment and the Tao. *Language Teaching Research*, 20(2), 181–195.

Pinner describes this paper as a narrative study based on an exploratory practice inquiry that centered on his introduction of a self-assessment system for class participation in a Japanese university English-speaking skills course. The autobiographical narrative takes up around two-thirds of the paper and focuses on Pinner's interactions with a particular "problematic" student. Various sources of data were collected during the exploratory practice project, but the narrative is based mainly on the author's journal entries and field notes. Pinner describes narrative as both a form of data collection and the primary means of analysis and write-up. He notes that much of the analysis took place when preparing the manuscript for publication and in conversations with colleagues to whom he explained the story as it emerged. Pinner observes that "[w]riting this narrative has been an act of self-reflection for me and part of the learning came from the act of writing about it" (p. 93). The main outcome of the study was an affirmation of value of self-assessment and, in particular, the importance of issues of trust and taking responsibility that had led Pinner to introduce self-assessment into his teaching.

Griffiths et al. (2014) and Oxford et al. (2014) (published in a special issue of *System* on learning strategies) exemplify a new multiauthored approach in this genre, in which they present a series of short narratives that are then analyzed thematically. Griffiths et al. (2014), for example, is concerned with language learning strategy use in Asian settings. The paper is co-authored by two North American and six Asian researchers. The Asian researchers provide four personal essays on their involvement in strategy research, which occupy about half the length of the paper. These essays are then analyzed using a thematic coding approach. A growing acceptance of personal experience writing as research may reflect recognition that as researchers and practitioners in a globalized profession, we often have valuable insights to offer based on our personal experiences, which cannot be communicated through more conventional, and depersonalized, forms of research writing.

## Biographical Case Studies

In biographical research, researchers' construct narratives from the autobiographical experiences of others. While most autobiographical studies adopt a narrative analysis approach, the reverse is the case for biographical studies, which more often use content or discourse analysis methods. Examples of third-person narrative accounts include case studies of language learners (e.g., Chik & Benson, 2008; Kinginger, 2004; Umino & Benson, 2016) and teachers (e.g., Hayes, 2010; Liu & Xu, 2011; Tsui, 2007). Benson, Barkhuizen, Bodycott, and Brown's (2013) book on second language identity development in study abroad includes 10 narrative case studies of students' study abroad experiences, selected from 40 that were written for the project on which it was based. Narrative has also recently been adopted as an approach to organizing the findings of qualitative research papers, which are analyzed and reported as participant narratives, rather than under thematic headings (Allen & Katayama, 2016; Ferris, Liu, Sinha, & Senna, 2013; Tanghe & Park, 2016; Yu & Lee, 2015; Zheng & Borg, 2014).

### Sample Study 26.2

Hayes, D. (2010). Duty and service: Life and career of a Tamil teacher English in Sri Lanka. *TESOL Quarterly*, 44(1), 58–83.

Hayes's (2010) paper, published in *TESOL Quarterly*, discussed the life and career of Krishnan, a Tamil teacher in a war-torn area of northern Sri Lanka. The paper includes a literature review on TESOL in areas of conflict, focusing on Sri Lanka, and a discussion of the methodology of the study. Approximately half of the paper is devoted to Krishnan's story, which is told by Hayes in the third person and interspersed with direct quotes from an interview with Krishnan. The data for the study was gathered through a single life history interview. Hayes does not describe how the interview was transformed into the narrative that appears in the paper. However, he does note that Krishnan's story was "co-constructed by two individuals" and was influenced by Hayes's own background and beliefs (p. 66). This raises important ethical questions about the need for self-reflexivity on the author's role in telling the stories of others in biographical research (see Sterling and De Costa, Chap. 8, on ethical applied linguistics research). Hayes also notes that the length of the article did not allow for a full telling of Krishnan's story and that he chose extracts from it to support his personal understanding of important themes in Krishnan's learning and teaching experience. Hayes contends that Krishnan's story sheds light on how language teachers find motivation to work in situations of extreme personal danger and how they position themselves within their communities. He also suggests that the paper extends the knowledge base of TESOL by "providing space for a voice from a peripheral community to be heard" (p. 58).

In autobiographical studies, the author-participant's narrative tends to occupy much of the published outcome. In biographical studies, however, the form and length of participant narratives vary a great deal. Kinginger (2004) is, essentially, a narrative reconstruction of the participant's experiences of learning and using French in the United States, Canada, and France. Tsui (2007) adopts a more formal separation of researcher and participant voices in a study of a Chinese EFL teacher's professional identity formation. The paper largely consists of Tsui's retelling of the participant's story (interspersed with direct quotations from his communications with her), which is followed by a short interpretation of its significance in terms of communities of practice theory. Biographical studies usually include some interpretive comment or further thematic analysis. In Benson et al. (2013), this was formalized in a structure in which each chapter began with two 1500–2000-word participant narratives, which were followed by an extended interpretation of them in terms of second language identity theory. In other work, the researcher's interpretation can be little more than a concluding paragraph that sums up the main point of the narrative from the researcher's point of view. The basis for deciding on an appropriate approach is partly a question of whether the main findings of the study are to be found in the researcher's interpretation of the narrative or in the narrative itself (Benson, 2013).

## Narrative Analysis as Method

Narrative analysis refers, here, specifically to the use of narrative writing as a data analysis tool (Ely, 2007; Richardson, 1994). It is through the act of narrative writing that researchers come to understand the meanings within their data and communicate them to readers. In this context, narrative writing is more than simply writing or retelling a story. In his autobiographical account of an exploratory practice project on self-assessment, Pinner (2016, p. 193), for example, explains how writing the narrative was “an act of self-reflection for me and part of the learning came from the act of writing about it.” Storytelling is often an intuitive act, but in the context of research reflection on the process of storytelling is often called for. It is fair to say that narrative researchers are often less than explicit about how they write research narratives, but the question of how we go about narrative writing *as a method of research* is at the heart of narrative analysis. In the remainder of this chapter, I

will look at this question from two perspectives—data analysis and writing for publication—although these aspects of the question are, in fact, much more closely intertwined in narrative analysis than they are in other approaches to research.

## Analyzing Data

Polkinghorne (1995, p. 5) writes that the researcher's task in narrative analysis is "to configure the data elements into a story that unites and gives meaning to the data as contributors to a goal or purpose." He goes on to argue that the *analytic* task is "to develop or discover a plot that displays the linkage among the data elements as parts of an unfolding temporal development culminating in the denouement." This emphasis on temporal development and plot marks the key difference between narrative analysis and thematic analysis. In thematic analysis, codes are added to segments of the data, which are reassembled to represent emergent themes that often cut across individual cases. Coding is part of a process of theoretical abstraction in which the coherence of individual cases is often sacrificed. Narrative analysis aims to maintain the integrity of the individual case by drawing together segments of the data in order to create a temporally coherent plot with a meaning that has some bearing on a research issue.

Thematic coding is, in fact, often used in narrative analysis in order to establish potential themes that can be developed through the storyline of the narrative. However, narrative analysis usually requires another method of data analysis that helps the researcher reduce the data to manageable proportions, while maintaining the integrity of the storyline to be developed. In a narrative study that aimed to explain why some students continued foreign language learning beyond high school, while others did not, Shedivy (2004) used a four-step *phenomenological* approach to analyzing interview transcripts, which aimed to capture the essence of the participants' experiences (Willis, 1991). The approach involved (1) reading the transcripts in their entirety; (2) extracting significant statements from each transcript; (3) formulating statements into meanings, which were clustered into themes; and (4) integrating the themes into narratives. The difference between this approach and a thematic coding approach lies mainly in Step (2), which is also a key step in an approach that Kvale and Brinkmann (2009, p. 205) call "meaning condensation," a procedure in which "long statements are compressed into briefer statements in which the main sense of what is said is rephrased in a few words." Whereas thematic codes often represent the researcher's interpretation of data



segments, the annotations used in meaning condensation aim to articulate the meaning of the data for the participants in a condensed form.

In my own research on language learning history narratives, I have also found it helpful to structure the data collection chronologically. For example, a language learning history interview might prompt the participant to explain how they learned a language from the earliest stages up to the present. In Benson et al. (2013), a pre-study abroad interview briefly covered prior life-long learning experiences and expectations for study abroad, data was collected from participants' blogs or emails while they were studying overseas, and a post-study abroad interview prompted participants to recount their study abroad experiences and reflect upon them. The data collection was designed, in other words, to help the participants construct their own study abroad stories. Because the data already had a rudimentary narrative structure, the researchers were better able to focus their attention on the meaning of experiences to the participants within the rewritten and condensed third-person narratives.

Future research may also reveal underlying narrative structures in learners' accounts of their long-term language experiences. Based on the analysis of language learning history interviews with Hong Kong university students, Benson (2011) developed the metaphor of a *language learning career* to refer to the narratives that many of these adult learners appeared to have developed to account for their lifelong English language learning experiences in interviews. These interviews suggested that language learning careers tend to be divided into relatively long *phases* that are characterized by repetitive learning *processes* as well as steady progress, stasis, or regression. These phases were often interrupted by particular episodes or events (such as an experience with a particular teacher or a short period of study abroad), which could be thought of as *critical incidents* if they marked a transition between phases. In subsequent research, I have found this to be a useful framework for the sequential analysis of data and the construction of narratives of language learning. Critical incidents can be an especially important focus for narrative analysis, because they help to explain change. Much applied linguistics research is concerned with causality, or what causes a change from one state of language learning and use to another. Narrative analysis of critical incidents can help us to understand the complex interactions of the learner's agency and psychological state with situated experiences of language in its social contexts that are involved in moments of change.

Whatever the approach to data analysis is, it is important for authors to give a full account of the process that leads from data collection to the production of findings. Editors are often skeptical of the lack of detail provided on



data analysis in narrative analysis papers. While some of the papers cited in this chapter do include transparent descriptions of data analysis (e.g., Liu & Xu, 2011; Shedivy, 2004), others provide little or no information. Below, I cite two extracts from my own work that were written in response to reviewers' requests for more information on data analysis.

The first extract comes from a four-year case study of a Hong Kong student (Ally) and her experience of studying for a university degree in the United Kingdom (Chik & Benson, 2008). The study was based on three interviews conducted before, during, and after Ally's study abroad.

After transcription, the first two interviews were coded thematically, with Ally's language background and her expectations and evaluations of overseas study emerging as important themes, and the clash between her expectations prior to departure and evaluations after two years overseas emerging as a key to the narrative structure of her experience. Using short narratives of critical incidents within the data as core elements, Ally's experiences were written up in story form. Ally then read this story and commented on it during the third interview. This interview was also coded, with the themes of re-evaluation and awareness of identity change emerging strongly. A summary of this interview was used as a concluding section to the narrative. Although the second author of this chapter contributed to the analysis of the data, he has not met Ally and the story that follows is narrated by the first author, who interviewed her and knows her well. (Chik & Benson, 2008, pp. 158–159)

The second extract is taken from a paper in which I presented and discussed a study abroad narrative of a student, Selina, who participated in the project on which Benson et al. (2013) was based. Part of the purpose of this paper was to advocate narrative analysis as an approach within narrative inquiry. The detailed description of the methodology used to write Selina's narrative was intended to support this purpose.

First, I read the data several times, in the order that it was collected, to gain an overall sense of who Selina was and what she was trying to say. Through repeated readings, I tried to work myself into her way of thinking. I then eliminated data that was obviously irrelevant to the research issues in which we were interested and highlighted sections where she spoke about, or hinted at, development or change (at this stage keeping an open mind about whether or not these developments were related to second language identity). As I gradually reduced the quantity of data that would remain in the narrative, I also began to cluster data extracts together into what would eventually appear as thematically-structured paragraphs. The data already had a rudimentary 'beginning-middle-end'

structure, which contributed to the sequential ordering of paragraphs, but at this stage, data extracts were also moved around in an attempt to create greater coherence. Lastly, I began connecting extracts together into coherent paragraphs, a process that involved weaving my own words into Selina's, while trying as far as possible to retain her wording where it mattered (Benson, 2013, pp. 257–258).

While these extracts are not intended to serve as models for narrative analysis, they are intended to suggest that narrative analysis does typically involve systematic analysis procedures and that these procedures can be described.

## Narrative Writing for Publication

In most narrative analysis studies, narrative is part and parcel of the data analysis process as drafts are written and rewritten, discussed with participants and within the research team, and then rewritten again. Narrative writing for publication, therefore, is a multilayered process from which published narratives are emergent outcomes. This section considers issues that tend to arise as narratives approach the stage where they are to be read by a wider public, by exploring some of the qualities that make narrative analysis studies worth reading.

Bell (2002, p. 207) introduced narrative inquiry to the field of applied linguistics with the suggestion that it was “more than just telling stories.” By this she meant that the narratives in narrative research must have bearing on some question or problem that is of interest to a community of applied linguistics researchers and practitioners. In general, published research should meet the basic criterion of providing data-based evidence to support an argument that addresses one or more research questions. Earlier in this chapter, I raised the question of whether this is the job of the researcher's interpretation of a narrative or of the narrative itself. In analysis of narrative studies, the responsibility clearly lies with the interpretation, because narrative plays the role of data to be analyzed and interpreted. In the case of narrative analysis studies, however, it seems reasonable to place the burden of addressing research questions on the narrative, especially where there is a claim that the narrative *is* a product of analysis and it occupies a considerable proportion of the space in the published work. This criterion might be applied by both author as well as manuscript reviewers and editors seeking to make a recommendation or decision regarding the value of a particular study.

Benson (2013) used the narrative described in the extract cited above to address the question of whether a narrative could perform a similar function to the more concise answers to research questions that are typically expected of research studies. I argued that Selina's narrative did, in fact, address an important question in language learning research: "What is actually involved in second language identity development in study abroad?" The narrative was a *research* narrative, because it had been written in order to address this question.

In practice, however, narratives are rarely left to stand for themselves as research findings. They are usually accompanied by some interpretation on the part of the researcher, often in the form of a succinct statement of how the narrative addresses the research questions at hand. There are at least three potential problems with this approach. First, the interpretation of the narrative may throw a veil over the analytical work that has already gone into the writing of it. By explaining *how* Selina's narrative answers the research question, I may be guilty of concealing how I wrote the narrative *in order to* answer to that question. Second, pointing out how a narrative answers a research question will often involve making an unjustifiable generalization from a single case. While generalizations can be made from observation of shared features in multiple narratives (Josselson, 2007; McAdams, 2012), this is the terrain of analysis of narratives, not narrative analysis. Last, it is, perhaps, one of the functions of narrative analysis to show how research questions in applied linguistics often defy succinct answers by offering more complex, fine-grained, situated answers to them in the form of narrative accounts of language teaching and learning.

## Conclusion: The Challenge of Narrative Analysis

By focusing on storytelling as a method of data analysis, this chapter has put the spotlight on an approach to narrative inquiry in applied linguistics research that offers a particular challenge to established approaches to research. Narrative is of interest, not only as a relatively new and evolving approach in the field but also as an approach that questions the nature of academic knowledge and discourse. Bruner (1986, p. 11) articulated this challenge through an opposition between narrative and paradigmatic "ways of knowing": the difference between "a good story" and "a well-formed argument." Both are convincing in their own ways, Bruner argued. There is also some evidence that applied linguists are beginning to accept that both can be accepted as outcomes of research.

## References

- Allen, D., & Katayama, A. (2016). Relative second language proficiency and the giving and receiving of written peer feedback. *System*, 56, 96–106.
- Atkinson, R. (1998). *The life story interview*. Thousand Oaks, CA: SAGE.
- Barkhuizen, G. (2010). An extended positioning analysis of a pre-service teacher's better life small story. *Applied Linguistics*, 31(2), 282–300.
- Barkhuizen, G. (2011). Narrative knowledging in TESOL. *TESOL Quarterly*, 45(3), 391–414.
- Barkhuizen, G. (2013). Introduction: Narrative research in applied linguistics. In G. Barkhuizen (Ed.), *Narrative research in applied linguistics* (pp. 1–16). Cambridge: Cambridge University Press.
- Barkhuizen, G. (2014). Research timeline: Narrative research in language teaching and learning. *Language Teaching*, 47(4), 450–466.
- Barkhuizen, G., Benson, P., & Chik, A. (2013). *Narrative inquiry in language teaching and learning research*. London: Routledge.
- Barkhuizen, G., & Wette, R. (2008). Narrative frames for investigating the experiences of language teachers. *System*, 36(3), 372–387.
- Bell, J. S. (2002). Narrative inquiry: More than just telling stories. *TESOL Quarterly*, 36(2), 207–218.
- Benson, P. (2004). (Autobiography) and learner diversity. In P. Benson & D. Nunan (Eds.), *Learners' stories: Difference and diversity in language learning* (pp. 2–21). Cambridge: Cambridge University Press.
- Benson, P. (2011). Language learning careers as a unit of analysis in narrative research. *TESOL Quarterly*, 45(3), 545–553.
- Benson, P. (2013). Narrative writing as method: Second language identity development in study abroad. In G. Barkhuizen (Ed.), *Narrative research in applied linguistics* (pp. 244–263). Cambridge: Cambridge University Press.
- Benson, P. (2014). Narrative inquiry in applied linguistics research. *Annual Review of Applied Linguistics*, 34, 154–170.
- Benson, P., Barkhuizen, G., Bodycott, P., & Brown, J. (2013). *Narratives of second language identity in study abroad*. Basingstoke: Palgrave Macmillan.
- Benson, P., & Cooker, L. (Eds.). (2013). *The applied linguistic individual: Social approaches to identity agency and autonomy*. London: Equinox.
- Benson, P., & Nunan, D. (Eds.). (2002). The experience of language learning. Special issue of the *Hong Kong Journal of Applied Linguistics*, 7(2).
- Benson, P., & Nunan, D. (Eds.). (2004). *Learners' stories: Difference and diversity in language learning*. Cambridge: Cambridge University Press.
- Brockmeier, J., & Carbaugh, D. (Eds.). (2001). *Narrative and identity: Studies in autobiography, self and culture*. Amsterdam: John Benjamins.
- Brooks, P. (1979). Fictions of the Wolfman: Freud and narrative understanding. *Diacritics*, 9(1), 71–81.
- Bruner, J. (1986). *Actual minds: Possible worlds*. Cambridge, MA: Harvard University Press.

- Canagarajah, A. S. (2012). Teacher development in a global profession: An autoethnography. *TESOL Quarterly*, 46(2), 258–279.
- Casanave, C. P. (2012). Diary of a dabbler: Ecological influences on an EFL teacher's efforts to study Japanese informally. *TESOL Quarterly*, 46(4), 642–670.
- Chamberlayne, P., Bornat, J., & Wengraf, T. (Eds.). (2000). *The turn to biographical methods in social science: Comparative issues and examples*. London: Routledge.
- Charmaz, K. (2014). *Constructing grounded theory*. Thousand Oaks, CA: SAGE.
- Chik, A. (2014). Constructing German learner identities in online and offline environments. In D. Abendroth-Timmer & E. M. Henning (Eds.), *Plurilingualism and multiliteracies: International research on identity construction in language education* (pp. 161–176). Berlin: Peter Lang.
- Chik, A., & Benson, P. (2008). Frequent flyer: A narrative of overseas study in English. In P. Kalaja, V. Menezes, & A. M. F. Barcelos (Eds.), *Narratives of learning and teaching EFL* (pp. 155–168). Basingstoke: Palgrave Macmillan.
- Choi, J. (2012). Multivocal post-diasporic selves: Entangled in Korean dramas. *Journal of Language, Identity and Education*, 11(2), 109–123.
- Clandinin, D. J., & Connelly, F. M. (2000). *Narrative inquiry: experience and story in qualitative research*. San Francisco, CA: Jossey-Bass.
- Curtis, A., & Romney, M. (Eds.). (2006). *Color, race, and English language teaching: Shades of meaning*. Mahwah, NJ: Lawrence Erlbaum.
- De Fina, A., & Georgakopoulou, A. (2012). *Analyzing narrative: Discourse and sociolinguistics perspectives*. Cambridge: Cambridge University Press.
- Ellis, C., & Bochner, A. P. (2000). Autoethnography, personal narrative, reflexivity. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 733–768). Thousand Oaks, CA: SAGE.
- Ely, M. (2007). In-forming re-presentations. In J. Clandinin (Ed.), *Handbook of narrative inquiry: Mapping a methodology* (pp. 567–598). Thousand Oaks, CA: SAGE.
- Ferris, D. R., Liu, H., Sinha, A., & Senna, M. (2013). Written corrective feedback for individual L2 writers. *Journal of Second Language Writing*, 22, 307–329.
- Giddens, A. (1991). *Modernity and self-identity: Self and society in the late modern age*. Stanford, CA: Stanford University Press.
- Goodson, I., & Sikes, P. (2001). *Life history research in educational settings: Learning from lives*. Buckingham, UK: Open University Press.
- Griffiths, C., Oxford, R. L., Kawai, Y., Kawai, C., Park, Y. Y., Ma, X., ... Yang, N.-D. (2014). Focus on context: Narratives from East Asia. *System*, 43, 50–63.
- Harbon, L., & Moloney, R. (2013). *Language teachers' narratives of practice*. Newcastle-upon-Tyne: Cambridge Scholars Press.
- Hayes, D. (2010). Duty and service: Life and career of a Tamil teacher English in Sri Lanka. *TESOL Quarterly*, 44(1), 58–83.
- Hoffman, E. (1989). *Lost in translation: A life in a new language*. London: Penguin.
- Holstein, J. A., & Gubrium, J. F. (Eds.). (2012). *Varieties of narrative analysis*. Thousand Oaks, CA: SAGE.

- Johnson, K. E., & Golombek, P. R. (Eds.). (2002). *Teachers' narrative inquiry as professional development*. Cambridge: Cambridge University Press.
- Josselson, R. (2007). Narrative research and the challenge of accumulating knowledge. In M. Bamberg (Ed.), *Narrative: State of the art* (pp. 7–16). Amsterdam: John Benjamins.
- Josselson, R., & Lieblich, A. (Eds.). (1993). *The narrative study of lives*. Thousand Oaks, CA: SAGE.
- Kalaja, P., Menezes, V., & Barcelos, A. M. F. (Eds.). (2008). *Narratives of learning and teaching EFL*. Basingstoke: Palgrave Macmillan.
- Kasper, G., & Prior, M. T. (2015). Analyzing storytelling in TESOL interview research. *TESOL Quarterly*, 49(2), 226–255.
- Kinginger, C. (2004). Alice doesn't live here anymore: Foreign language learning and identity reconstruction. In A. Pavlenko & A. Blackledge (Eds.), *Negotiation of identities in multilingual contexts* (pp. 219–242). Clevedon: Multilingual Matters LTD.
- Kouritzin, S. (2000). Immigration mothers redefine access to ESL classes: Contradiction and ambivalence. *Journal of Multilingual and Multicultural Development*, 21(1), 14–19.
- Kvale, S., & Brinkmann, S. (2009). *InterViews: Learning the craft of qualitative research interviewing* (2nd ed.). Los Angeles, LA: Sage Publications.
- Lee, I., & Sze, P. (Eds.). (2015). *Voices from the frontline: Narratives of nonnative English speaking teachers*. Hong Kong: The Chinese University Press.
- Lieblich, A., Tuval-Mashiach, R., & Zilber, T. (1998). *Narrative research: Reading, analysis, and interpretation*. Thousand Oaks, CA: SAGE.
- Liu, Y., & Xu, Y. (2011). Inclusion or exclusion? A narrative inquiry of a language teacher's identity experience in the 'new work order' of competing pedagogies. *Teaching and Teacher Education*, 27(3), 589–597.
- López-Gopar, M. E. (2014). Teaching English critically to Mexican children. *ELT Journal*, 68(3), 310–320.
- Mann, S. (2011). A critical review of qualitative interviews in Applied Linguistics. *Applied Linguistics*, 32(1), 6–24.
- McAdams, D. P. (2012). Exploring psychological themes through life-narrative accounts. In J. A. Holstein & J. F. Gubrium (Eds.), *Varieties of narrative analysis* (pp. 15–32). Thousand Oaks, CA: SAGE.
- Menard-Warwick, J. (2004). "I always had the desire to progress a little": Gendered narratives of immigrant language learners. *Journal of Language, Identity, and Education*, 3(4), 295–311.
- Menezes, V. (2011). Affordances for language learning beyond the classroom. In P. Benson & H. Reinders (Eds.), *Beyond the language classroom* (pp. 59–71). Basingstoke: Palgrave Macmillan.
- Moodie, I. (2016). The anti-apprenticeship of observation: How negative prior language learning experience influences English language teachers' beliefs and practices. *System*, 60, 29–41.



- Murphey, T., & Carpenter, C. (2008). The seeds of agency in language learning histories. In P. Kalaja, V. Menezes, & A. M. F. Barcelos (Eds.), *Narratives of learning and teaching EFL* (pp. 17–34). Basingstoke: Palgrave Macmillan.
- Murray, G. (2009). Narrative inquiry. In J. Heigham & R. Croker (Eds.), *Qualitative research in applied linguistics* (pp. 45–65). Basingstoke: Palgrave Macmillan.
- Murray, G., & Kojima, M. (2007). Out-of-class learning: One learner's story. In P. Benson (Ed.), *Learner autonomy 8: Insider perspectives on autonomy in language teaching and learning* (pp. 25–40). Dublin: Authentik.
- Nikula, T., & Pitkänen-Huhta, A. (2008). Using photographs to access stories of learning English. In P. Kalaja, V. Menezes, & A. M. F. Barcelos (Eds.), *Narratives of learning and teaching ELF* (pp. 171–185). Basingstoke, UK: Palgrave Macmillan.
- Nunan, D., & Choi, J. (Eds.). (2010). *Language and culture: Reflective narratives and the emergence of identity*. London: Routledge.
- O'Móchain, R. (2006). Discussing gender and sexuality in a context-appropriate way: Queer narratives in an EFL college classroom in Japan. *Journal of Language, Identity, and Education*, 5(1), 51–66.
- Oxford, R. L., Rubin, J., Chamot, A. U., Schramm, K., Lavine, R., Gunning, P., & Nel, C. (2014). The learning strategy prism: Perspectives of learning strategy experts. *System*, 43, 30–49.
- Palfreyman, D. M. (2014). The ecology of learner autonomy. In G. Murray (Ed.), *Social dimensions of autonomy in language learning* (pp. 175–191). Basingstoke: Palgrave Macmillan.
- Pavlenko, A. (2002). Narrative study: Whose story is it, anyway? *TESOL Quarterly*, 36(2), 213–218.
- Pavlenko, A. (2007). Autobiographic narratives as data in applied linguistics. *Applied Linguistics*, 28(2), 163–188.
- Pinner, R. (2016). Trouble in paradise: Self-assessment and the Tao. *Language Teaching Research*, 20(2), 181–195.
- Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, 39, 579–592.
- Polkinghorne, D. E. (1988). *Narrative knowing and the human sciences*. Albany, NY: State University of New York Press.
- Polkinghorne, D. E. (1995). Narrative configuration in qualitative analysis. *Qualitative Studies in Education*, 8(1), 5–23.
- Pomerantz, A., & Kearney, E. (2012). Beyond 'write-talk-revise-(repeat)': Using narrative to understand one multilingual student's interactions around writing. *Journal of Second Language Writing*, 21(3), 221–238.
- Richardson, L. (1994). Writing: A method of Inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 516–529). Thousand Oaks, CA: SAGE.
- Rieff, P. (1996). The collected papers of Sigmund Freud. Volume 7: Three case histories. New York, NY: Touchstone. (First published by Macmillan 1963)
- Riessman, C. K. (1993). *Narrative analysis*. Newbury Park, CA: SAGE.

- Schmidt, R. W., & Frota, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237–326). Rowley, MA: Newbury House.
- Schumann, F. M., & Schumann, J. H. (1977). Diary of a language learner: An introspective study of second language learning. In H. D. Brown, C. A. Yorio, & R. Crymes (Eds.), *On TESOL '77 teaching and learning English as a second language: Trends in research and practice* (pp. 241–249). Washington, DC: TESOL.
- Shedivy, S. L. (2004). Factors that lead some students to continue the study of foreign language past the usual 2 years in high school. *System*, 32(1), 103–119.
- Simpson, J. (2011). Telling tales: Discursive space and narratives in ESOL classrooms. *Linguistics and Education*, 22(1), 10–22.
- Takeuchi, O. (2003). What can we learn from good foreign language learners? A qualitative study in the Japanese foreign language context. *System*, 31(3), 385–392.
- Tanghe, S., & Park, G. (2016). “Build[ing] something which alone we could not have done”: International collaborative teaching and learning in language teacher education. *System*, 57, 1–13.
- Thomas, W. I., & Znaniecki, F. (1919). *The Polish peasant in Europe and America: Monograph of an immigrant group* (Vol. 3). Boston, MA: Richard G. Badger.
- Tsui, A. B. M. (2007). The complexities of identity formation: A narrative inquiry of an EFL teacher. *TESOL Quarterly*, 41(4), 657–680.
- Umino, T., & Benson, P. (2016). Communities of practice in study abroad: A four-year study of an Indonesian student’s experience in Japan. *Modern Language Journal*, 100(4), 1–18.
- Van Lier, L. (2004). *The ecology and semiotics of language learning: A sociocultural perspective*. Boston, MA: Kluwer Academic Publishers.
- Willis, G. (1991). Phenomenological inquiry: Life-world perceptions. In E. C. Short (Ed.), *Forms of curriculum inquiry* (pp. 173–186). Albany, NY: State University of New York Press.
- Wyatt, M., & Márquez, C. P. (2016). Helping first-year undergraduates engage in language research. *Language Teaching Research*, 20(2), 146–164.
- Xu, Y. (2014). Becoming researchers: A narrative study of Chinese university EFL teachers’ research practice and their professional identity. *Language Teaching Research*, 18(2), 242–259.
- Yu, S., & Lee, I. (2015). Understanding EFL students’ participation in group peer feedback of L2 writing: A case study from an activity theory perspective. *Language Teaching Research*, 19(5), 572–593.
- Zheng, X., & Borg, S. (2014). Task-based learning and teaching in China: Secondary school teachers’ beliefs and practices. *Language Teaching Research*, 18(2), 205–221.





# 27

## Interaction Analysis

Elizabeth R. Miller

### Introduction

Interaction Analysis can be understood as an umbrella term for numerous approaches that pursue “an empirical investigation of the interaction of human beings with each other” (Jordan & Henderson, 1995, p. 39). Influences on contemporary applications of Interaction Analysis are often traced back to what Streeck (2010) refers to as the “interactionist canon” (p. 1), including work dating to the 1930s (i.e., Bakhtin, Mead, and Vygotsky) as well as to more commonly cited research from the 1960s and 1970s, that is, research in ethnography of communication (Hymes, 1964), ethnomethodology (Garfinkel, 1967), Goffman’s (1974) work on interaction order, interactional sociolinguistics (Gumperz, 1982), Conversation Analysis (CA) (Sacks, Schegloff, & Jefferson, 1974), and microethnography (Erickson, 1982). More recent instantiations of Interaction Analysis from the last few decades can be found in discursive psychology (Edwards & Potter, 1992), positioning theory (Davies & Harré, 1990) and/or positioning analysis (Bamberg, 1997), interactional linguistics (Ford & Wagner, 1996; Selting & Couper-Kuhlen, 2001), and linguistic ethnography (Rampton et al., 2004). Given such a long list of influences and offshoots or strands of what might legitimately be identified as Interaction Analysis, with

---

E. R. Miller (✉)

University of North Carolina Charlotte, Charlotte, NC, USA

e-mail: [ermiller@uncc.edu](mailto:ermiller@uncc.edu)

links to disciplinary traditions in sociology, psychology, anthropology, linguistics, philosophy, and education, among others, it becomes clear that selectivity for this chapter is a necessity. But more importantly, the diversity and sheer number of influences also point to the broad appeal and utility of Interaction Analysis for investigations into many facets of social reality and meaning-making, such as investigations into the role of particular social identities in interaction (i.e., gender, race and ethnicity, among others) or features of institutional interactions such as in classrooms or between doctors and patients.

In this chapter, I will focus primarily on the use of Interaction Analysis in research on language learning. It is important to clarify that second language research that uses Interaction Analysis does not share the foundational premises of the Interaction Hypothesis (Long, 1996). The Interaction Hypothesis postulates language learning to be primarily a cognitive process that occurs in an individual mind when language learners receive comprehensible input and interactional feedback, often from individuals identified as “native speakers” and who can engage in negotiation for meaning in order to arrive at target-like forms. Gass (2004) provides a clear demonstration of how interactions involving language learners are interpreted differently by scholars drawing on the Interaction Hypothesis compared to those who draw on Interaction Analysis (CA in this case). The following section introduces the theoretical foundations of Interaction Analysis.

## What Interaction Analysis Is About

Interaction Analysis is based on the premise that as we interact with other people, in particular situations and spaces, and at particular moments in time, we simultaneously create, maintain, and/or transform our social worlds, our understanding of social order, our identities, along with various social norms and ideologies. As Jacoby and Ochs (1995) pointed out some time ago, not only is interaction always coconstructed, the “things allegedly in people’s heads—such as cognition and attitudes, linguistic competence, or pragmatic and cultural knowledge—are made relevant to communication through social interaction” (p. 175). More significantly, they note, these interactive processes are formative to “the constitution, management and negotiation of social reality and social relationships” (p. 175). Given this kind of theoretical orientation to the constitutive power of human interaction, the potential scope for research topics approached through Interaction Analysis is incredibly broad.

Importantly for this chapter, adopting the view that human interaction is enormously consequential to the formation of our social realities has “methodological consequences,” as Jordan and Henderson (1995, p. 41) put it. Interaction Analysis focuses on what is observable in human interaction, though microethnography analysts would certainly not discount the insights that introspective accounts such as from interviews or research participants’ reflective journals can lend to their analysis (Trognon & Batt, 2010). Furthermore, interaction is understood to be orderly rather than chaotic (Goodwin, 1981; Kerbrat-Orecchioni, 2010), and analysts can thus gain insight into its normative effects by exploring how participants enact that orderliness and communicate its situated meanings to each other (Jordan & Henderson, 1995).

For language learning research, Interaction Analysis is used to explore how knowledge is distributed across interlocutors and artifacts and how language learning is practiced, afforded, and constrained in specific situations. Thus, instead of using research protocols that attempt to elicit the language competence stored in the minds of people (e.g., grammaticality tests), this approach attempts to understand how participants demonstrate learning through their changing participation in particular discursive practices, as coconstructed activity (Hellermann & Harris, 2015; Young & Miller, 2004). Developing competence in a language is thus more appropriately conceived as developing “interactional competence” rather than linguistic competence. Young (2008) defines interactional competence as “a relationship between the participants’ employment of linguistic and interactional resources and the contexts in which they are employed. It is not an individual phenomenon, but is coconstructed by all participants in a particular discursive practice” (p. 101). Likewise, Markee (2008) contends that the *development* of interactional competence among language learners is observable as they deploy various “inter-subjective resources [such as turn-taking, repair, eye gaze or gestures] to co-construct with their interlocutors locally enacted, progressively more accurate, fluent, and complex interactional repertoires” (p. 406).

These views of learning and interactional competence presuppose that analysts will explore what language learners *do* when participating in discursive or interactive practices and, ideally, how they change their participation over time, thus requiring longitudinal research (Hellermann, 2008; Hellermann & Harris, 2015; Young & Miller, 2004). Some scholars who use a particular form of Interaction Analysis, namely, Conversation Analysis (CA), have also conducted longitudinal research that focuses on learners’ developing capacity to use more accurate renderings of selected linguistic features such as particular vocabulary terms (Markee, 2008), negation (Eskildsen, 2012), and motion

constructions (Li, Eskildsen, & Cadierno, 2014) in interaction, rather than focus on changes in participation.

As already noted, this chapter cannot do justice to all of the analytic approaches that fall under the umbrella of Interaction Analysis, and for that reason it will not provide an in-depth introduction to CA as applied to language learning research (see Markee, 2000; Seedhouse, 2004; Sert & Seedhouse, 2011). However, it is important to acknowledge the enormous influence that CA has had on Interaction Analysis more broadly. CA's rigorous attention to the details of interaction, its use of standardized transcription conventions first introduced by Gail Jefferson (1984), its focus on naturally occurring interactions, and its insistence on grounding one's interpretive claims in what the participants themselves demonstrate to be relevant to the interaction are some of the methodological characteristics that have been adopted, in various ways, by many interaction analysts.

Given that my own analytic practice aligns more closely with microethnography—though still heavily influenced by CA transcription practices—I will discuss this approach more fully in this chapter. Microethnography has been described as “the study of how human realities are produced, activities are conducted, and sense is made, by inspecting video recordings of actual events frame by frame” (Streck & Mehus, 2005, p. 382). This description points to the strong methodological reliance on audio-visual/video data and the attention given to the details of the sequential unfolding of interaction (i.e., the *micro*-level frame-by-frame analysis). Its label, *microethnography*, suggests an acceptance of using additional forms of ethnographic data to inform the interaction analysis (e.g., fieldnotes, artifacts, interviews, etc.). Given its origins in school-based research (Erickson & Mohatt, 1982; Mehan, 1981), it is not surprising that microethnography has been adopted for research exploring aspects of classroom culture, school identities, and learning and teaching practices (e.g., Bloome, Carter, Christian, Otto, & Shuart-Faris, 2005; Miller & Zuenger, 2011).

## Current Core Issues

A long-held issue for interaction analysts is the need to discern and document the relationship between micro-level interactions and their macro-level social concerns or effects. While interaction has been theorized as the micro-level site for creating the macro-level social and ideological effects, showing how the micro and macro are simultaneously produced requires great care.

However, this dichotomous micro-macro view of human interactions has been strongly challenged in recent scholarship. Lin (2010), for example, advocates for replacing the dualistic macro-micro view of human interactions with an integrated perspective in which social structures are understood to be “enacted, maintained, reproduced, taking shape or being contested, being transformed ... etc., through and through in the micro interactional event” (p. 131), such that there is no separation between the micro and macro. Indeed, many contemporary scholars regard “linguistic, discursive and cultural products [as phenomena] that cannot be detached from the specific local and social conditions that are responsible for them coming into being” (Pérez-Milans, 2016, p. 86), a perspective which foregrounds the inextricable interconnectedness of the macro-micro domains and which casts the alleged macro phenomena as always “coming into being,” much like the turn-by-turn production of interaction.

For language learning researchers, a core issue is that of providing evidence of learning, made more challenging given that the learning process is “entwined in the progress of interaction” (Sert & Seedhouse, 2011, p. 5). But an even greater challenge has emerged with the “multilingual turn” (May, 2014) in applied linguistics more generally and language learning research more particularly. Ortega (2005, 2014) has long challenged the monolingual bias in second language research, which, she argues, confers a deficit competence to non-monolinguals and simultaneously promotes an ideological view of language as relatively fixed and bounded, a symbolic system that can be “owned” by “native speakers” of a language (Ortega, 2014, p. 36). In calling for an “epistemic reorientation” (p. 38) that aligns with the bi/multilingual turn, Ortega promotes usage-based theories. I do not aim to endorse a particular learning theory in this chapter, but I do want to note that such a theoretical shift also entails a methodological shift to examining “usage events [and] ... emergent phenomena and thus a process rather than object” (Ortega, 2014, p. 40). Interaction Analysis can certainly accommodate such theoretical and methodological perspectives (e.g., Eskildsen, 2012; Li et al., 2014).

At the same time that the multilingual turn has been gaining ground among language learning researchers, a growing number of language researchers has advocated abandoning the nomenclature of multiplicity or plurality altogether as suggested in terms such as *multilingualism*, *plurilingualism*, and/or *multivocality*. They have promoted new ways of naming and conceiving of language and language knowledge such as “translanguaging,” “translingual practice,” and/or a “complexity perspective” to avoid re-entrenching long-held ideological views in which languages are commonsensically treated as

autonomous systems that can be demarcated, counted, sometimes mixed (as in code-switching), and sometimes pluralized (as in multilingual practices) (Makoni & Pennycook, 2007). These perspectives instead promote concepts of language such as *soft assemblage* (García & Flores, 2014), *mobile semiotic resources* (Canagarajah, 2014), *code-meshing* (Michael-Luna & Canagarajah, 2008), and *complexity* (Blommaert, 2010).

This attention to complexity, movement, and hybridity of language requires an analytic approach that is sensitive to such concepts. Rampton (2015) has argued that:

interactional perspectives are geared up for indeterminacy, contestation and ambivalence in the links between linguistic forms and macro-social categories, and they are analytically equipped to describe situations where such categorizations either do not matter to the participants there-and-then or alternatively, when they do matter, where their indexical associations are themselves open to local negotiation and dispute. (p. 41)

Rampton's description positions Interaction Analysis as a uniquely apt methodological approach for exploring language (and learning) as both locally emergent but also indexical of other socially familiar practices across temporal and spatial scales.

## Kinds of Data

Because of analysts' focus on the moment-to-moment sequential development of interaction, video-recorded data are regarded as essential. An important advantage of video-recorded data is that analysts can replay selected excerpts numerous times. Garcez (1997) notes that such repeated viewing of video excerpts enables "deep analysis of phenomena which may be impossible to perceive in real-time observation and which may be too heavily laden with common-sense perceptions for participant observers to see through them" (p. 193) such as when one relies primarily on note-taking. For this reason, it is important that researchers seek to obtain the highest quality recorded data possible. Good sound and image quality greatly enhance the ability to conduct in-depth analysis and make the process far more enjoyable. As Rymes (2016) wryly noted, watching a "jostling video later, repeatedly, is not a pleasant experience" (p. 60). While obtaining good recordings enables more careful analysis, it is important for analysts to never forget that working with

video data is still not the same as capturing or extracting the event itself. Rymes (2016) also points out that “a recording of anything subtracts from the multidimensionality of lived experience” (p. 54). Even so, working with video remains the best medium available for rendering fleeting interactional moments into concrete data that can be scrutinized repeatedly, shared with other analysts, and, through sharing insights, gain credibility for data interpretations and confirmation of their explanatory power. The research benefits of this kind of data-sharing effort are demonstrated in an edited volume in which the separate chapter authors examine the same video-recorded classroom-based interactional event but use different interaction-based analytic and theoretical approaches (Cole & Zuengler, 2008). Interwoven among the separate chapters are brief “conversations” among the researchers regarding the decisions they made during their analytic processes and how those decisions affected their research outcomes.

After recording the interactions, the video files need to be transcribed into written representations for in-depth analysis to proceed. Though researchers should always work from both the video data and written transcriptions when conducting analysis, the transcribed renderings of the interaction allow one to pinpoint subtle behaviors and compare them to other moments in the interaction such as to identify a language learner’s participation patterns and any changes in those patterns over time. Transcription is never simply a mechanical methodological practice. The transcription process is not entirely separate from the analysis process in that as researchers give careful, focused attention to the unfolding actions of interaction, while transcribing, they often begin to identify salient patterns in the talk (see Ford & Couper-Kuhlen, 2004). Jordan and Henderson (1995) advise researchers to consider carefully the kind of analysis they intend to do “before launching into full-scale transcription” (p. 48) because the transcription format and its level of granularity inevitably shape the analysis. Ochs (1979) long ago reminded us that all transcription choices (such as whether to use standard orthography or phonetic representations of talk) are never neutral; transcription is always, she argued, “unavoidably theory” (Ochs, 1979, p. 43).

Conversation analysts have drawn on a relatively standardized set of transcription conventions (Jefferson, 1984), many of which have been adopted by other interaction analysts. These symbols are used to represent many of the nonverbal behaviors that are consequential to meaning-making in talk such as voice pitch and prosody, pausing, and overlapping talk. Scholars who are interested in gesture have created innovative methods of representing these embodied movements as integrated into the verbal

interaction, often through including supporting images (e.g., Mori & Hayashi, 2006). A particularly innovative transcription, though one that is incredibly time-consuming to produce, is the use of musical notation to capture rhythm patterns and pitch contours in interaction (e.g., Erickson, 2004). A growing number of researchers are working with digitally mediated interactions involving language learners (see Thorne, Sauro, & Smith, 2015 for a recent overview) and are working to create transcription methods that allow them to incorporate the many channels, modalities, and affordances that contribute to online interaction and language learning, such as in game play and social media. For example, Zheng, Newgarden, and Young (2012) explore how language learning can occur in *World of Warcraft* game play. In conducting their analysis, they incorporated screen grabs of game play, transcribed the gestures and actions of gamers' avatars, as well as gamers' verbal interactions (mediated through Skype) into their transcriptions. Other recent efforts to create transcription conventions for digitally mediated interaction analysis include efforts by Meredith and Potter (2013), Sindoni (2014), and Giles, Stommel, Paulus, Lester, and Reed (2015).

Given that the process of transcribing interaction can lead to “substantive analytic insights” (Jordan & Henderson, 1995, p. 47), many researchers prefer to do their own transcription. But one needs to be mindful of how incredibly time- and labor-intensive the transcription process is. An hour of classroom video-recorded data may take from 10 to 20 hours to transcribe, and for this reason transcribing all of one's recordings before any analysis begins may be prohibitively time-consuming. It is often preferable for analysts to view the video-recorded data multiple times, taking notes on what appear to be salient, important, or confusing moments in order to select key moments for detailed transcription and further analysis.

## What to Look for

A shared characteristic of all approaches to Interaction Analysis is a focus on the sequential moment-by-moment development of interaction, that is, a focus on how participants engage in turn-taking. Conversation analysts regard sequence organization (including adjacency pairs and preference structures), turn-taking, and repair to be fundamental foci for all analyses. Working from these fundamental interactional practices, Markee's (2008) CA approach to language learning scholarship has led him to look at the “particular behaviors



[that] actually promote language learning,” which he notes “are massively achieved as repair sequences” (p. 408). Markee has identified the kinds of interactional behaviors that are embedded in these language learning repair sequences that seem to signal language learning activity. They include:

statements of non-comprehension, and/or emphatic assertions of understanding (these verbal behaviors are often accompanied by smiling, clapping, and embodied actions such as thinking gestures and pointing to information in written texts); changes of epistemic state, including the use of tokens such as *oh* ... ; participants independently volunteering new information that connects the learning object to practices or knowledge that are already part of their interactional repertoires; and translation from language to another. (Markee, 2008, pp. 408–409)

Ethnographically oriented interaction analyses of classroom sites often focus on how turn-taking is managed by considering who talks, who listens, who asks the questions and who answers, and what kinds of questions are being asked (Rymes, 2016). Erickson (2004) proposed that analysts pay attention to topic initiation and how authority over topic definition is displayed, how attentiveness between speakers and hearers is displayed, and how turn authorization is accomplished.

Informed by practice-based theories of language learning, Young and Miller (2004) advocated that analysts consider the following aspects of a practice: (a) the boundaries of a discursive practice, (b) the selection of speech acts and their sequential organization in a practice, (c) the turn-taking system and how speakers manage their turn-taking activity, (d) how participants construct a participation framework involving particular roles (i.e., tutor/tutee, peer/peer, etc.), (e) the register of the practice produced through characteristic vocabulary and syntactic structures, and (f) the ways in which meaning is coconstructed in a practice (see p. 520). Evidence for learning is traceable, they contend, as interlocutors change how they enact, construct, and participate in these features of a practice.

Though the type of interaction analysis one undertakes may guide what the analyst looks for (as in CA), there is no a priori set of features of interaction other than turn-taking that analysts always account for. Standardization of transcription methods is desirable given that it creates a shared resource for analysis and interpretation, but existing methods will inevitably need to be stretched to accommodate new practices or communicative affordances (such as in digitally mediated interaction).

## Computer Programs and Online Training for Transcription

The minimum equipment needed for a beginning researcher include a video recorder, a computer and a transcription foot pedal. A foot pedal is an inexpensive investment that will be amply rewarded in terms of faster and more pleasant transcription work. If a researcher is striving to create highly detailed annotations of gesture and audio features, one can advance to using annotation software such *ELAN* <https://tla.mpi.nl/tools/tla-tools/elan/>, a free resource, or *Transana* <http://www.transana.org/index.htm>. Though the latter must be purchased, for a modest fee, it provides an excellent interface between the video recordings and the transcription platform. Both of these programs offer exciting options for annotating and transcribing interactional behaviors, but they do require a substantial time commitment for learning how to operate them. For information on other potentially useful programs, see *TalkBank's* software list: <https://talkbank.org/software/>. In addition, if one is completely new to transcription work, one can undergo basic training in CA transcription at an online site created by Emanuel Schegloff

<http://www.sscnet.ucla.edu/soc/faculty/schegloff/TranscriptionProject/index.html> or one by Charles Antaki <http://homepages.lboro.ac.uk/~ssca1/sitemenu.htm>.

## Example Studies Using Interaction Analysis

In this section, I introduce two studies that demonstrate related applications of Interaction Analysis for language learning research in quite different learning sites. The first study (Hellermann & Harris, 2015) showcases an excellent example of how Interaction Analysis can be applied in a traditional English language classroom in which students and teachers share the same physical space. The second study (Messina Dahlberg & Bagga-Gupta, 2014) nicely demonstrates how Interaction Analysis can be used to understand how language learning interactions emerge in complex multilingual, multimodal, digitally mediated spaces.

### Example 1

Hellermann and Harris's (2015) longitudinal study (duration 9 months) examines the language learning trajectory of Li (a pseudonym), an immigrant to the US, as she participated in an ESL class for adult learners. Li's

language class was part of a classroom-based research project (The National Labsite for Adult ESOL), and thus all of her ESL classes were held in a room that was equipped with six video cameras, two of which focused on student pairs during the language practice sessions. These student pairs also wore wireless microphones to guarantee the collection of clear audio data. As such, this study is able to work with multiple-angled, high-quality video/audio data. While a large-scale, grant-funded study such as this benefits from numerous technical and administrative resources, a beginning or solo analyst can follow similar methodological steps, even if on a more modest scale and still produce insightful and informative research. The researchers watched all of the videos of the class sessions in which Li was a participant (an astonishing total of 156 hours), and while viewing the videos, they kept a record of each case in which Li participated in classroom interactions. The video data were transcribed according to CA conventions, and the authors revised and refined the transcription segments after repeated subsequent viewings.

Although this published chapter focuses on only ten brief excerpts of talk that displayed Li's "marked increase" (Hellermann & Harris, 2015, p. 58) in participation in the classroom interactions over nine months, it should be clear that these selections follow from a staggeringly large number of hours invested in preparing the classroom space, managing the video data, transcribing and retranscribing it, and analyzing it. While it is true that the selected excerpts give us entry into only a few, brief moments of Li's learning trajectory, the authors' methodological process, as described in the chapter, demonstrates care and integrity. Moreover, they supply a hyperlink to their video data, enabling other scholars to view and analyze the data for themselves (and to challenge the interpretations, if warranted): <http://www.labschool.pdx.edu/Viewer/viewer.php?LiExcerpts>.

I will discuss only the first and last of the ten excerpts which come from Li's 3rd and 25th week in the class. In Excerpt 1, the participants are working through a task on daily routines (i.e., *What time do you get up?*). Li and Shen have been assigned to work together with the teacher so that she can assist them. The bolded and underlined text in the transcript provides an English gloss of what Li and Shen say to each other in Taiwanese. It is also important to know that Li had never studied in a classroom of any kind before joining this ESL class; thus, not only is she learning English, she is also learning how to be a student and how to participate in classroom practices. Please refer to the Transcription Key located below Excerpt 2.

### Excerpt 1 Li's Third Week (Adapted from Hellermann & Harris, 2015, pp. 58–59)

- 1) Tchr: It's okay, okay. what ti::me (.) do you get up.  
 2) time. ((taps watch)) What time is it (now)  
 3) Li: What time? What time get up? She said what time you  
 4) get up in the morning  
 5) Shen: What is that? 'w' is what time?  
 6) Li: What time  
 7) Tchr: Wha:t [time  
 8) Li: [that is what time. She is saying what time you  
 9) get up ( [ )  
 10) Tchr: [what time  
 11) Shen: Oh she said what time.  
 12) Tchr: What ti:me do you get up. So you're sleepi:ng,  
 13) ((puts head on hands)) and then suddenly you go ↑BRLINNG::  
 14) and I have to wake up. what time ((points to Li))  
 15) Li: Seven (ty)  
 16) Tchr: seven?  
 17) Li: mhm ((nods head))  
 18) Tchr: okay. seven.  
 19) Li: yeah

Hellermann and Harris indicate that Li seems to understand the content of the question the teacher wants her to answer (lines 1–2) given her ability to translate the teacher's question into Taiwanese (3–4, 6, 8–9), but the absence of a response to the teacher's question over the course of several speaker changes and the teacher's repetition of the question five times suggest that Li does not understand *why* the teacher is asking the question. Li and Shen appear to be trying to figure out the purpose of the teacher's question (3–6, 8–9, 11). After the teacher mimes the action of sleeping and mimics the sound of an alarm clock, Li finally indicates that she gets up at seven o'clock (15) and the teacher produces a confirmation check, "seven?" (16, 18).

Nearly 22 weeks later, Li again participated in a practice task focusing on daily routines with a classmate, but this time the teacher is not a participant in the interaction. On this occasion, Li displays no confusion about the task, and, in fact, seems to take the lead in performing this practice task.

### Excerpt 2, Li's 25th Week (Adapted from Hellermann & Harris, 2015, pp. 68–69)

- 1) Li: ((holds hand over Sergio's workbook, looks at board))  
 2) What time (.) do you get up. ((looks at Sergio,  
 3) touches his arm, points at board))  
 4) huh ((points to an image in Sergio's book))  
 5) Sergio: s:::even (.) o [clock  
 6) Li: [se- seven o'clock. What time do you  
 7) okuh. ((nonstandard pronunciation of "work"))  
 8) ((gazes at board))

Eight turns later

- 17) Li: What time do you lunche.  
 18) ((points at image in workbook))  
 19) Sergio: (eat) lunche eh mm::  
 20) (2.3)  
 21) Sergio: I eat lunch ( fore) twelve.  
 22) Li: What time do you:: (2) lunche  
 23) ((points at image in workbook))  
 24) (1.5)  
 25) Li: What time lunche.  
 26) ((points at image in workbook))  
 27) (1.5)

- 28) Sergio: what time do you eat lu:nch, three.  
 29) Li: euhyeah. What time do you watcheh tv.

Transcription Key

(.) or (3)	Micropause or pause in seconds
((points))	Transcriber's descriptions
(ly)	Uncertainty on transcriber's part
you::	Drawn-out sound
[	Point of overlap onset
.	Falling or final intonation contour
,	Continuing intonation

In Excerpt 2, Li directs the “what time” questions to her classmate Sergio (lines 2, 6, 17, 22, 25, 29). She repeats the first “what time” question (6) for him after he displays some tentativeness in his answer, marked by a sustained “s” sound and a micropause between words in “s:::even (.) o'clock” (5). We can see that Sergio repeats part of Li's second question along with some hesitation markers “eh mm:::” and then allows a 2.3-second pause to elapse before producing a response (20–21). Li repeats this question twice more (22, 25) and points to an image in Sergio's textbook both times. When Sergio supplies an answer (28), she ratifies it as correct “euhyeah” (29) and then moves on to the next practice question with no delay (29). The researchers note that on this occasion, “as evidenced in the timing of her questions, Li goes through the task confidently and her demeanor suggests that she is not only engaged in the task but even excited about leading her peer through the task” (Hellermann & Harris, 2015, p. 70).

In taking a theoretical orientation to learning as participation (Lave & Wenger, 1991), Hellermann and Harris can show that Li's enhanced expertise is not limited to gains in her individual linguistic competence but is coconstructed with her interlocutors in the interaction and distributed across material

features such as a textbook and the whiteboard. Her enhanced expertise is also performed in her bodily actions (pointing, touching, eye gaze) and verbal behaviors. Because the authors transcribed the fine-grained verbal and nonverbal details of the interaction, their interpretation is grounded in the striking differences between Li's performed tentativeness in the first excerpt and her performed confidence and proficiency in the second. It is not merely the content of what she is able to comprehend and produce (*what time do you ...*) but how adeptly she can manage the turn-taking and recognize *how* to participate within a given interactional configuration.

## Example 2

Messina Dahlberg and Bagga-Gupta (2014) explore how learning and interaction are performed in an online Italian language classroom. As the authors note, this kind of learning space, which allows for participation by language learners from any location, also affords “the emergence of hybrid, translingual and transmodal forms of communication” (p. 469), and they regard interaction in such sites as the “doing of language rather than in terms of competencies in a so-called first, second or additional language” (p. 470). As such, they show their alignment with language scholars who advocate for a complexity perspective to language studies (e.g., Blommaert, 2010, among others noted earlier).

A primary concern in this study is to better understand how online spaces enable interaction and produce particular learner identities, but the authors are also concerned with how to interpret and represent the “transmodality” of these online learning spaces (p. 472), pointing to methodological challenges. For instance, they needed to create a transcription method that allowed them to capture written (typed) and spoken (audio-recorded) as well as visual (video-recorded) components of interaction (such as interlocutors' facial expressions and gestures) produced by multiple interlocutors who often used several different languages and who were “meeting” through their separate computer monitors. The researchers are physically located in Sweden and thus Swedish, along with Italian (and English and to a lesser degree Spanish) are part of the meaning-making activity in this learning space. The students in the course were physically located in Sweden as well as in the UK, Madagascar, Denmark, and the Netherlands.

The Italian language students met online once weekly over 12 weeks and their meetings were mediated by the videoconferencing program Adobe® Connect. This online environment enables students and teachers to communicate with each other orally, through writing (on a “whiteboard” for content material and in a chat-log for interaction), and visually when individual students’ webcams are activated. It also records everything that is visible on-screen during these sessions as well as audio-records participants’ spoken interactions. The researchers adapted CA transcription conventions to accommodate the interactional affordances of the online learning environment to represent the use of multiple languages and modalities and the possibility for multiple conversation floors to occur simultaneously. They selected varied font styles to represent the different languages that were used in the (sometimes co-occurring) written and oral interactions (see Transcription Key below the excerpt).

The following excerpt represents audio-recorded verbal interaction between two students and a simultaneously produced chat-log utterance from the teacher. On this occasion, the participants had not activated their webcams, and thus the students did not have access to each other’s embodied actions. Messina Dahlberg and Bagga-Gupta’s transcription includes the participants’ original language choices along with their English glosses, a typical practice in scholarly publications. (Hellermann and Harris’s use of only the English translations of interlocutors’ utterances is less common.) Due to limited space, I will focus on only a part of the authors’ analysis of this stretch of interaction.

After some displayed confusion (lines 1–3), Luisa responds to Olle’s question asking whether Macedonian is similar to Serbian, first confirming that it is similar (4) before adding that it is a language similar to Russian (6). In the single utterance shown in line 6, Luisa switches from Italian to Macedonian to Italian to English and back to Italian. Olle shows his understanding of Luisa’s response by uttering, in Italian, that Macedonian is a Slavic language (7), and as he does so, the teacher simultaneously begins typing a comment in the chat-pod. This action is communicated via a message that appears (in English) on the screen (7) as the oral interaction between Olle and Luisa continues. Olle then repeats his utterance that Serbian is a Slavic language (9a) at the same time as the teacher’s chat-log turn appears on the screen (9b). This written utterance confirms Olle’s earlier utterance that Macedonian is a *lingua slava* (7), but her production uses the correct morphology, the feminine –a ending on *slava*. It is possible that this written information is noticed by Olle at the same time that he verbally utters his turn *slava e una lingua slava* (9a),



thereby treating the teacher's written chat-log message as an implicit correction to his earlier utterance in which he used the masculine –o ending *slavo*. Luisa likewise orients to the correct Italian morphology in her uptake of the utterance *lingua slava* in line 10.

### Excerpt (Adapted from Messina Dahlberg & Bagga-Gupta, 2014, p. 474)

1 Olle: macedone è similar al serbo?

Is Macedonian similar to Serbian?

(5)

2 Luisa: ehm (.) ursäkta vem frågade?

Ehm 9.) sorry who asked?

(3)

3 Olle: eh eh io io voglio sapere se se similar al serbo (2) [macedone]

eh eh I I want to know if it is similar to Serbian (2) [Macedonian]

4 Luisa: [sì sì sì] ((laugh))

[yes yes yes] ((laugh))

5 Olle: [molto similar]

[very similar]

6 Luisa: [vorrei] (3) **da** (.) **da!** ((laugh)) sì sì è una lingua vicino a (.) vad sager m- (.)

[I would like] (3) **yes** (.) **yes** ((laugh)) yes yes it is a language near to (.) how do you say

similar to russi (.) eller russi?

similar to Russian (.) or Russian?

**632 E. R. Miller**

7 Olle: è una lingua slavo | Teacher starts typing |

It is a Slavic language |

8 Luisa: russo |

Russian |

9a Olle: slava è una lingua slava |

Slavic it is a Slavic language |

9b Tchr: | **CHAT** | < -----|

| lingua slava |

| Slavic language |

10 Luisa: **da** (.) si si lingua slava **da** (.) si ((laugh)) jag blir helt f orvirrad (.) ja

**yes** (.) yes yes Slavic language **yes** 9.0 ((laugh)) I get completely confused (.) yes

Transcription Key

- No emphasis Oral utterance in Italian
- Italics Oral utterance in Swedish
- Bold Oral utterance in Macedonian
- Italics Underlined Oral utterance in English
- (.) or (3) Micropause or pause in seconds
- [macedone] Simultaneous overlapping talk
- 1a Indicates simultaneity of turns in different modalities
- 1b
- | | Chat pod written language

The linguistic feature (the use of masculine vs. feminine morphemes –o and –a in Italian) is almost incidentally oriented to as the interlocutors work out their communicative goal of establishing intersubjectivity around the

topic of Macedonian being a Slavic language. Messina Dahlberg and Bagga-Gupta (2014) in fact contend that their investigations of the interactions in this online learning ecology “illuminate learning as a process that occurs at the boundaries of [interlocutors’] fluid shifting from one language variety to another” (p. 483). This translanguaging, they argue, “is not merely a way to solve issues of ‘insufficient’ knowledge in the target language but rather illustrates participants’ orientation towards language varieties and modalities that are accessible in the focused gathering” (p. 475). They also found that the teacher’s production of meta-language utterances (such as in line 9b) that appear in the chat-pod contributes to the participants’ positions as experts and novices, and thus, it seems participants’ identities or positions are constructed, in part, through how the distinctive communicative modalities are taken up.

## Challenges and Future Directions

It seems likely that Interaction Analysis will continue to be a fruitful methodology for language learning research. In conducting an informal survey of the 90 research articles published in 2015 in three leading journals in applied linguistics (*TESOL Quarterly*, *Applied Linguistics*, and *The Modern Language Journal*), I found that 17 (19%) of them used some form of Interaction Analysis (they minimally had to give analytic attention to turn-taking). As language learning scholarship seeks to shift away from its long-standing monolingual bias, Interaction Analysis is particularly well suited for pursuing research based on usage- or practice-oriented theories of language learning given its attention to the details of interactional practice, as demonstrated in Li’s enhanced ability to participate in the language practice sessions more expertly and confidently (Hellermann & Harris, 2015). It seems clear too that Interaction Analysis will increasingly be used to explore digitally mediated interactions and online learning spaces given the ongoing growth of such interactional activity. However, such efforts will continue to challenge researchers as they seek to account for the complex array of affordances, modalities, and meaning-making practices in these spaces. Likewise, the closer attention given to translanguaging will continue to pose challenges, not so much in accounting for the different languages in play, but in analyzing the linguistic resources as *translanguaging* rather than simply as mixing multiple languages. Messina Dahlberg and Bagga-Gupta’s (2014) careful use of multiple font styles to index Luisa’s translanguaging in a single utterance still demarcated shifts between *different* languages (i.e., Italian, Macedonian, and

English). As such, interaction analysts must continue striving to arrive at a “nuanced framework [and analytic methodology] with which to describe and analyze communication patterns which appear to have become, and continue to become, more dynamic, mobile, and complex” (Creese & Blackledge, 2015, p. 33).

And finally, Interaction Analysis must adapt to the still-nascent but growing exploration of new materiality theories which view the influence of materials on human meaning-making, not merely as mediating artifacts for people but as relational “assemblages” in which “both the material and human *interact* with one another and *influence* one another and *accomplish* social and material goals” together (Toohey & Dagenais, 2015, p. 304, italics added). This expanded focus points to yet another vexing issue. In order to do analytic justice to the complex assemblages of people, materials, actions, and knowledge that are part of language learning and meaning-making in interaction, the analysis of very short segments of interaction becomes increasingly complex, and the data that are published fill ever more page space, thereby constraining researchers to accounting for even briefer interactional moments in these sites. It is likely that two-dimensional journals and books will increasingly need to incorporate supplemental online sites so that readers can gain access to the video-taped data and thus can experience the “data” themselves.

Streeck (2010) noted that “the adventures that await those who enter the microcosm of human interaction” will open them to “challenges of multiple kinds” (p. 2). As interaction researchers apply and adapt their analytic methods to new and changing landscapes of interaction, they are uniquely sensitized to explore carefully and deeply *what is going on* at any moment of interactional meaning-making. We still have much to learn and we will always find interactions that still need to be explored.

## References

- Bamberg, M. G. (1997). Positioning between structure and performance. *Journal of Narrative and Life History*, 7(1–4), 335–342.
- Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge: Cambridge University Press.
- Bloome, D., Carter, S. P., Christian, B. M., Otto, S., & Shuart-Faris, N. (2005). *Discourse analysis and the study of classroom language and literacy events: A microethnographic perspective*. Mahwah, NJ: Lawrence Erlbaum.
- Canagarajah, S. (2014). Theorizing a competence for translanguaging practice at the contact zone. In S. May (Ed.), *The multilingual turn: Implications for SLA, TESOL and bilingual education* (pp. 78–102). New York: Routledge.

- Cole, K., & Zuengler, J. (Eds.). (2008). *The research process in classroom discourse analysis: Current perspectives*. New York: Routledge.
- Creese, A., & Blackledge, A. (2015). Translanguaging and identity in educational settings. *Annual Review of Applied Linguistics*, 35, 20–35.
- Davies, B., & Harré, R. (1990). Positioning: The discursive production of selves. *Journal for the Theory of Social Behaviour*, 20(1), 43–63.
- Edwards, D., & Potter, J. (1992). *Discursive psychology*. London: Sage.
- Erickson, F. (1982). Audiovisual records as a primary data source. *Sociological Methods & Research*, 11(2), 213–232.
- Erickson, F. (2004). *Talk and social theory*. Malden, MA: Polity Press.
- Erickson, F., & Mohatt, J. (1982). Cultural organization of participant structures in two classrooms of Indian students. In B. Spindler (Ed.), *Doing the ethnography of schooling* (pp. 132–174). New York: Holt, Rinehart & Winston.
- Eskildsen, S. W. (2012). L2 negation constructions at work. *Language Learning*, 62(2), 335–372.
- Ford, C. E., & Couper-Kuhlen, E. (2004). Conversation and phonetics: Essential connections. In E. Couper-Kkuhlen & C. E. Ford (Eds.), *Sound patterns in interaction: Cross-linguistic studies from conversation* (pp. 3–25). Amsterdam: John Benjamins.
- Ford, C. E., & Wagner, J. (1996). Interaction-based studies of language. *Pragmatics*, 6(3), 277–279.
- Garcez, P. M. (1997). Microethnography. In N. H. Hornberger & D. Corson (Eds.), *Encyclopedia of language and education* (pp. 187–196). Dordrecht: Kluwer.
- García, O., & Flores, N. (2014). Multilingualism and common core state standards in the United States. In S. May (Ed.), *The multilingual turn: Implications for SLA, TESOL and bilingual education* (pp. 147–166). New York: Routledge.
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.
- Gass, S. (2004). Conversation analysis and input-interaction. *Modern Language Journal*, 88(4), 597–602.
- Giles, D., Stommel, W., Paulus, T., Lester, J., & Reed, D. (2015). Microanalysis of online data: The methodological development of “digital CA”. *Discourse, Context & Media*, 7, 45–51.
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge, MA: Harvard University Press.
- Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.
- Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.
- Hellermann, J. (2008). *Social actions for classroom language learning*. Clevedon, UK: Multilingual Matters.
- Hellermann, J., & Harris, K. A. (2015). Navigating the language-learning classroom without previous schooling. In D. A. Koike & C. S. Blyth (Eds.), *Dialogue in multilingual and multimodal communities* (pp. 49–78). Amsterdam: John Benjamins.

- Hymes, D. (1964). Introduction: Toward ethnographies of communication. *American Anthropologist*, 66(6), 1–34.
- Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction*, 28(3), 171–183.
- Jefferson, G. (1984). Transcript notation. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. ix–xvi). Cambridge: Cambridge University Press.
- Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *The Journal of the Learning Sciences*, 4(1), 39–103.
- Kerbrat-Orecchioni, C. (2010). The case for an eclectic approach to discourse-in-interaction. In J. Streeck (Ed.), *New adventures in language and interaction* (pp. 71–98). Amsterdam: John Benjamins Publishing.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.
- Li, P., Eskildsen, S. W., & Cadierno, T. (2014). Tracing an L2 learner's motion constructions over time: A usage-based classroom investigation. *Modern Language Journal*, 98(2), 612–628.
- Lin, A. (2010). Researching intercultural communication: Discourse tactics in non-egalitarian contexts. In J. Streeck (Ed.), *New adventures in language and interaction* (pp. 125–144). Amsterdam: John Benjamins Publishing.
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition: Second language acquisition* (pp. 413–468). New York, NY: Academic Press.
- Makoni, S., & Pennycook, A. (2007). *Disinventing and reinventing languages*. Clevedon, UK: Multilingual Matters.
- Markee, N. (2000). *Conversation analysis*. Mahwah, NJ: Lawrence Erlbaum.
- Markee, N. (2008). Toward a learning behavior tracking methodology for CA-for-SLA. *Applied Linguistics*, 29(3), 404–427.
- May, S. (Ed.). (2014). *The multilingual turn: Implications for SLA, TESOL and bilingual education*. New York, NY: Routledge.
- Mehan, H. (1981). Ethnography of bilingual education. In H. T. Trueba, G. P. Guthrie, & K. H. Au (Eds.), *Culture and the bilingual classroom: Studies in classroom ethnography* (pp. 36–55). Rowley, MA: Newbury House.
- Meredith, J., & Potter, J. (2013). Conversation analysis and electronic interactions: Methodological, analytic and technical considerations. In H. L. Lim & F. Sudweeks (Eds.), *Innovative methods and technologies for electronic discourse analysis* (pp. 370–393). Hershey, PA: IGI Global.
- Messina Dahlberg, G., & Bagga-Gupta, S. (2014). Understanding glocal learning spaces. An empirical study of languaging and transmigrant positions in the virtual classroom. *Learning, Media and Technology*, 39(4), 468–487.
- Michael-Luna, S., & Canagarajah, S. (2008). Multilingual academic literacies: Pedagogical foundations for code meshing in primary and higher education. *Journal of Applied Linguistics*, 4(1), 55–77.

- Miller, E. R., & Zuenger, J. (2011). Negotiating access to learning through resistance to classroom practice. *Modern Language Journal*, 95(S), 130–147.
- Mori, J., & Hayashi, M. (2006). The achievement of intersubjectivity through embodied completions: A study of interactions between first and second language speakers. *Applied Linguistics*, 27(2), 195–219.
- Ochs, E. (1979). Transcription as theory. *Developmental pragmatics*, 10(1), 43–72.
- Ortega, L. (2005). For what and for whom is our research? The ethical as transformative lens in instructed SLA. *Modern Language Journal*, 89(3), 427–443.
- Ortega, L. (2014). Ways forward for a bi/multilingual turn in SLA. In S. May (Ed.), *The multilingual turn: Implications for SLA, TESOL and bilingual education* (pp. 32–53). New York: Routledge.
- Pérez-Milans, M. (2016). Language and identity in linguistic ethnography. In S. Preece (Ed.), *The Routledge handbook of language and identity* (pp. 83–97). New York: Routledge.
- Rampton, B. (2015). Contemporary urban vernaculars. In J. Nortier & B. A. Svendsen (Eds.), *Language, youth and identity in the 21st century: Linguistic practices across urban spaces* (pp. 24–44). Cambridge: Cambridge University Press.
- Rampton, B., Tusting, K., Maybin, J., Barwell, R., Creese, A., & Lytra, V. (2004). UK linguistic ethnography: A discussion paper. Retrieved from [www.ling-ethnog.org.uk](http://www.ling-ethnog.org.uk)
- Rymes, B. (2016). *Classroom discourse analysis: A tool for critical reflection* (2nd ed.). New York: Routledge.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.
- Seedhouse, P. (2004). Conversation analysis methodology. *Language Learning*, 54(S1), 1–54.
- Selting, M., & Couper-Kuhlen, E. (Eds.). (2001). *Studies in interactional linguistics*. Amsterdam: John Benjamins.
- Sert, O., & Seedhouse, P. (2011). Introduction: Conversation analysis in applied linguistics. *Language Teaching*, 5(1), 1–14.
- Sindoni, M. G. (2014). *Spoken and written discourse in online interactions: A multi-modal approach*. New York: Routledge.
- Streeck, J. (Ed.). (2010). *New adventures in language and interaction*. Amsterdam: John Benjamins Publishing.
- Streeck, J., & Mehus, S. (2005). Microethnography: The study of practices. In K. L. Fitch & R. E. Sanders (Eds.), *Handbook of language and social interaction* (pp. 381–404). Mahwah, NJ: Lawrence Erlbaum.
- Thorne, S. L., Sauro, S., & Smith, B. (2015). Technologies, identities, and expressive activity. *Annual Review of Applied Linguistics*, 35, 215–233.
- Toohy, K., & Dagenais, D. (2015). Videomaking as sociomaterial assemblage. *Language and Education*, 29(4), 302–316.

- Trognon, A., & Batt, M. (2010). Interlocutory logic: A unified framework for studying conversational interaction. In J. Streeck (Ed.), *New adventures in language and interaction* (pp. 9–46). Amsterdam: John Benjamins Publishing.
- Young, R. F. (2008). *Language and interaction*. New York: Routledge.
- Young, R. F., & Miller, E. R. (2004). Learning as changing participation: Discourse roles in ESL writing conferences. *Modern Language Journal*, 88(4), 519–535.
- Zheng, D., Newgarden, K., & Young, M. F. (2012). Multimodal analysis of language learning in s play: Linguaging as values-realizing. *ReCALL*, 24(3), 339–360.





# 28

## Multimodal Analysis

Jesse Pirini, Tui Matelau-Doherty, and Sigrid Norris

### Introduction

Researchers have developed a variety of ways for taking a multimodal perspective to study interactions between people, objects, and environments. Some approaches focus more on artefacts and texts, like images, objects, and buildings. Other approaches focus more on people interacting with people, environments, and objects. As a result, there is considerable differentiation regarding data collection and analysis amongst researchers taking a multimodal perspective.

Important differences in approaches to multimodal data analysis emerge through the way that approaches theorize modes, treat historical and psychological aspects in analysis, and consider objects, texts, and artefacts. Since data collection and analysis are tightly linked to these theoretical positions, it is important to keep these links in mind when conducting a multimodal research project.

In this chapter, we introduce the methods of data collection and analysis utilized in the most common approaches to multimodal analysis and discuss how these relate to the underlying theoretical positions of each approach.

---

J. Pirini (✉)

School of Management, Victoria University of Wellington,  
Wellington, New Zealand

e-mail: [jesse.pirini@vuw.ac.nz](mailto:jesse.pirini@vuw.ac.nz)

T. Matelau-Doherty • S. Norris

AUT Multimodal Research Centre, Auckland University of Technology,  
Auckland, New Zealand

e-mail: [tmatelau@gmail.com](mailto:tmatelau@gmail.com); [sigrid.norris@aut.ac.nz](mailto:sigrid.norris@aut.ac.nz)

Linking data analysis to theoretical positions ensures that research projects build upon a coherent theoretical basis.

## What Are the Approaches and What Is Analysed?

In this section, we briefly introduce the five most prominent approaches to multimodal data analysis in reverse chronological order. We begin our discussion with the most recently developed approach, Multimodal (Inter)action Analysis (Norris, 2004, 2011, 2013). The reason we begin here is that this is the only approach that was especially developed because of and for the analysis of multimodal action and interaction. We then discuss Mediated Discourse Analysis (Scollon, 1998, 2001), which, through its units of analyses, builds a cornerstone for the development of Multimodal (Inter)action Analysis and is also used in its own right, in Nexus Analysis (Scollon & Scollon, 2003) as well as in Geosemiotics (Scollon & Scollon, 2004), to examine multimodal data (Al Zidjaly, 2009; de Saint-Georges, 2014). Next, we outline Systemic Functional Multimodal Discourse Analysis (O' Toole, 2004; O'Halloran, 1998, 2004, 2008) and Social Semiotics (Jewitt, 2006; Kress, 1999; Kress & van Leeuwen, 1996, 2001, 2002), and we end this section by outlining Multimodal Conversation Analysis (Goodwin, 1979, 2000; Mondada, 2006, 2009). By tracing each approach's development from seminal publications, we identify how data collection and analysis proceed and how each approach theorizes the notion of multimodality.

It is noteworthy that there are additional approaches that could be considered. These include multimodal critical discourse analysis (Machin & Mayr, 2012), cognitive linguistics (Hart, 2016), multimodal metaphor (Forceville & Urios-Aparisi, 2009), and many others. Given the space limitations of this chapter, however, we will focus on what we perceive to be the most prominent approaches to multimodal analysis.

### Multimodal (Inter)action Analysis

Multimodal (Inter)action Analysis (Matelau, 2014; Norris, 2004, 2011, 2013; Norris & Pirini, 2017; Pirini, 2014a, 2014b, 2016, 2017; Pirini, Norris, Geenen, & Matelau, 2014) is a holistic approach to the analysis of multimodal action and interaction. With its strongest theoretical origins in Mediated Discourse Analysis (Scollon, 1998, 2001), the framework embraces the mediated action, defined as a social actor acting with/through mediational

means (Scollon, 1998, 2001; Wertsch, 1998) as a unit of analysis. Since a mediated action always involves a social actor, acting with and through mediational means, a central form of data is video recordings of social actors. Some examples of settings where researchers have used this framework include classroom interactions (Norris, 2014), family video-conferencing interactions (Geenen, 2017) (see Sample Study 28.1), management of attention/awareness in high school tutoring sessions (Pirini, 2014b), sign language in classrooms (Tapio, 2014), construction of ethnic identity within adult learning environments (Matelau, 2014), and the structure of lectures (Bernad-Mechó, 2017).

Further data sources are essential to a Multimodal (Inter)action Analysis, which reflect the sociolinguistic facets (Goffman, 1959, 1981; Gumperz, 1982; Hamilton, 1994; Schiffrin, 1987; Tannen, 1984) embraced by the approach. Researchers will usually observe interactions and take field notes at the same time as recording. Pirini's (2014b, 2016, 2017) studies of tutoring are based on data collected from tutoring sessions where Pirini was observing and taking field notes while also video recording. Similarly, Norris (2006) presents video-recorded data from an office setting. In Norris's project, she was present for much longer than the moments that were video recorded, enabling her to observe and collect field notes beyond the recordings. During observation, analysts often also collect photographs of texts and artefacts for further analysis.

Multimodal (inter)action analysts also interview participants, and they may speak to the participants about issues that are studied, and play back data to the participants to drive discussion (Tannen, 1984). Such triangulations allow researchers to gain a deeper understanding of what is being studied.

Once data collection has been completed, analysis begins with delineating interactions into higher-level actions to gain an understanding of the data. Actions of interest are then transcribed in detail. Developing detailed image-based multimodal transcripts is an important aspect in Multimodal (Inter)action Analysis. Then, multimodal (inter)action analysts apply a range of theoretical/methodological tools to the data depending on their interests. This includes analysing the distribution of participants' attention/awareness and their shifts into and out of focusing on various actions. These tools provide an understanding of the interaction(s) studied and are used to address research questions relating to multimodal interaction. Norris ([forthcoming](#)) outlines the phases of a Multimodal (Inter)action Analysis in more detail than we can provide here. Sample Study 28.1 walks readers through an example of a Multimodal (Inter)action Analysis studying family-child interaction over Skype.

### Sample Study 28.1

Geenen, J. (2017). Show and (sometimes) tell: Identity construction and the affordances of video-conferencing. *Multimodal Communication*, 6(1), 1–18.

#### Background

This study uses Multimodal (Inter)action Analysis as a methodological framework to study families interacting over video-conferencing. Geenen applies modal density, a methodological notion which positions actions in the foreground, mid-ground, and background of participants' attention/awareness. He shows that children agentively produce identity through showing and (sometimes) telling, taking advantage of the affordance of video to direct the attention/awareness of participants.

#### Research Questions

1. What does video-conferencing technology afford children in the production of their identity?
2. How might family members use this affordance when video-conferencing with children?

#### Method

Researchers recruited families with young children who communicate over video-conferencing with distant family members. With each family, researchers organized a suitable time to collect data of one such video-conferencing session.

Researchers visited the local family's home, with two video cameras and a laptop. Participants used the researcher laptop for video-conferencing and ran screen recording software to capture the video stream. The video cameras were positioned to record the local family as they used the laptop. Researchers were also present taking field notes.

Following the video-conferencing session, the researchers conducted a brief interview with a local family member, asking questions about their use of video-conferencing.

During analysis, Geenen noticed that often children would show things to the screen. He then sought out examples of this type of interaction in the data. After building a collection of data samples, he created multimodal transcripts to depict the higher-level actions where children showed things to the screen. After transcribing several examples, he applied modal density to analyse how showing and telling influenced the attention/awareness of adults and children.

#### Results

Geenen found that show and (sometimes) tell is a co-produced higher-level action which consists of multiple lower-level actions such as gaze, utterances, gestures, and movements of the device used in the video-conferencing. The multimodal transcripts show that this is a higher-level action co-produced by children and adults during video-conferencing.

Furthermore, children that show and (sometimes) tell during video-conferencing produce identity elements as the objects that they show have frozen actions embedded in their materiality. One example shown by Geenen involves a child displaying his hand-held gaming device to the interlocutor through video-conferencing technology. The child is not playing the game or instructing the inter-

locutor on how to play the game. Instead he is showing what he has done on the game already.

The frozen actions embedded in the hand-held device include previously playing the game and likely watching the film that the game is based on, revealing identity elements relating to the child's interests. In this way, show and (sometimes) tell is identity-telling.

Geenen argues that although show and (sometimes) tell can be initiated by the adult participants, a degree of agency is always evident in the lower-level actions performed by the children. This agentic production of identity elements during video-conferencing sessions is an affordance of video-conferencing technology.

Geenen proposes that adult participants can use show and (sometimes) tell to aid interaction when video-conferencing with children, especially younger children as the lower-level actions that make up show and (sometimes) tell are not dependent on the linguistic abilities of the child. Therefore, show and (sometimes) tell is an interactive strategy available to adult participants which can be used to enhance interaction with younger children during videoconferencing.

## Mediated Discourse Analysis

Mediated Discourse Analysis was developed by Ron Scollon (Scollon, 1998, 2001) and builds on anthropology, social psychology, and linguistics to highlight the primacy of social action in analyses of practices and discourses. Scollon adopted the concept of mediated action from Vygotsky (1978) and Wertsch (1998), to refer to a social actor acting with and through mediational means. From the perspective of Mediated Discourse Analysis, the actions social actors produce with mediational means (mediated actions) are the unit of analysis. Mediated Discourse Analysis is applied to study actions, practices and discourses but always starts with an analysis of concrete actions. Scollon's (2001) seminal study of the practice of handing focuses on the concrete action of handing objects from one person to another.

Importantly for multimodal analysis, Mediated Discourse Analysis also directly underpins Nexus Analysis (Scollon & Scollon, 2004) and Geosemiotics (Scollon & Scollon, 2003) and provides a part of the theoretical basis for Multimodal (Inter)action Analysis (Norris, 2004, 2011, 2013, forthcoming). Mediated discourse analysts embrace multimodality since their focus is upon social actions, which are inherently multimodal.

We will focus on a nexus analysis as an example of a Mediated Discourse Analysis here. A nexus is a highly complex intersection of practices, and researchers use the theoretical underpinnings of Mediated Discourse Analysis to study this intersection of practices and discourses. Pietikäinen, Lane, Salo, and Laihiala-Kankainen (2011) study endangered languages in the Arctic Circle. They first identify all kinds of material signs, such as street signs, signs

advertising businesses, and privately placed signs, as a nexus of practice. These signs are produced and positioned, people engage with them, and they embed traces of language policies, language hierarchies and norms relating to language use. Many of the signs use multiple languages, including minority, majority and global languages.

Collecting data for a nexus analysis requires first engaging the identified nexus of practice. In the case of Pietikäinen et al., many of the researchers had previously studied the language communities of interest. Building on this previous engagement, the researchers utilized participant observation through ethnographic fieldwork, to identify 20 sites relevant for language activities in these communities.

These sites include places like schools, grocery shops, and museums where language is used by minority language users, and also by all inhabitants and tourists. Signs in these areas were photographed, and field notes were taken providing background information on them. A total of 379 photographed signs were collected.

Then, the authors looked at who made the signs (municipal, organizations/companies/businesses, private), and analysed the different languages used by each group, seeking patterns of use and drawing conclusions from these findings. Central to this study is the theorization that the linguistic landscape (i.e., the signs) is an outcome of human action, and can therefore be used to explore historical, political, and economic processes and the relationship between various languages and their speakers. As a result of this perspective, data collection includes participant observation in order to deeply engage in the practices and discourses in the nexus of practice. Pietikäinen et al. used this deep engagement to guide the subsequent photography of signs. In nexus analytical studies, mediated discourse analysts may also collect interviews, questionnaires, and field notes; they may collect audio-video data and speak to the participants about issues that are studied.

## **Systemic Functional Multimodal Discourse Analysis**

Systemic Functional Multimodal Discourse Analysis originates from Systemic Functional Linguistics (Halliday, 1978), taking grammar as its theoretical basis. Since Systemic Functional Linguistics forms a basis also for Social Semiotics, we will briefly introduce the central theoretical tenets here.

Systemic Functional Linguistics considers language as a systematic structure of signs, which evolved to serve social functions. When people speak and write they make choices, selecting from a range of options. Each choice

acquires meaning in relation to possible choices that were not made. Thus, language operates as a semiotic resource where choosing amongst possible options can be used to make meaning. Multimodal approaches taking Systemic Functional Linguistics as a theoretical basis apply the same ideas about structure and function to other semiotic resources, like layout, framing, colour, and so forth.

Three main metafunctions of semiotic resources are central to a systemic functional approach. The first metafunction of semiotic resources is to represent internal and external experience (ideational metafunction). The second metafunction is to enact social relations, or meaning related to interpersonal relations (interpersonal metafunction). The third metafunction is to create a coherent flow of discourse (textual metafunction). These functions can be analysed separately, but it is important also to look at all three in conjunction, to understand how different functions come together to allow people to make sense of one another and the world (Eggins, 2004).

In terms of data, researchers using Systemic Functional Multimodal Discourse Analysis focus on texts. Drawing on Halliday (1978), a text is broadly defined as a meaning-making event. In practice, a text can refer to media like web pages, newspapers, and print and online advertisements. Researchers also study gesture, spoken language, and mathematical symbolism as texts (O'Halloran, 1998), while others explore buildings and art as texts (O'Toole, 2004). Researchers can collect hard or soft copy instances of print and online media texts. When interested in topics like mathematics teaching, buildings, and art, researchers can take photographs and use audio- and video recording devices to collect data.

Once data has been collected, the analyst starts by identifying and transcribing the semiotic resources deployed in a particular text and describing the ways that these semiotic resources interact to produce text-specific meaning. While data are not usually collected regarding how a participant engages with a text, transcription may be based upon assumed reading paths. Baldry and Thibault (2006) propose an approach called *cluster analysis*, which involves identifying clusters of semiotic resources that indicate possible reading paths.

After identifying clusters, and implied reading paths, researchers can analyse the semiotic resources used in each cluster. The analysis can be constructed as a table, with each cluster to the left and the semiotic resources used to the right.

As shown in Baldry and Thibault (2006, p. 29), the analysis of clusters may show that the text in the middle of a page depicts the *details*, while the text to the left and right depicts the *principle*. The principle and details of a system may be further reinforced with images. Further analyses can be conducted to

describe how a specific text operates as part of a larger genre. For example, a researcher might collect a corpus of flyers, of which the London transport system flyer is one. Analysis across the corpus can uncover patterns of semiotic resource use common to the genre of flyers.

Analyses using this approach can become very complex, as researchers identify detailed catalogues of semiotic resources and their interactions. Software has been developed to deal with the growing complexity of analysis. The Multimodal Analysis Image Software (O'Halloran, Marissa, & Tan, 2014) supports data management, annotation, search, and semi-automated analysis through identifying aspects of data such as faces, changes in shot, and motion of objects. Recent moves attempt to integrate Systemic Functional Multimodal Discourse Analysis software with data mining techniques to conduct automated analysis of large data sets. O'Halloran et al. (2016) propose using such techniques to study how meaning arises from text and image relations in media used by violent extremist groups and comparing these to representations of those same images by various media outlets.

## Social Semiotics

Social Semiotics also has its origin in Systemic Functional Linguistics (Halliday, 1978), and with this origin, Social Semiotics takes grammar as a theoretical basis. Social semioticians collect, document, and catalogue semiotic resources (van Leeuwen, 2005). Semiotic resources are defined as actions and artefacts, which are produced both physiologically, through the body, and technologically, through objects, tools and environments (van Leeuwen, 2005). These resources are used functionally, to achieve the social uses and needs of producers. Taking grammar as a theoretical basis suggests an underlying structure of the semiotic resources, which present meaning potentials to producers, from which semiotic choices can be made.

Semiotic potential can refer to all the theoretically possible and potential uses of a semiotic resource. However, in practice, researchers investigate actual semiotic potential, which refers to all uses of a semiotic resource which are considered known and relevant by its users (van Leeuwen, 2005). The actual semiotic potential can also include all potential uses that might be uncovered by users through needs and interests. Thus, social semioticians focus on the past uses of a semiotic resource, its current uses, and potential future use.

Kress and Van Leeuwen (1996) are seminal in the development of contemporary Social Semiotics by exploring a grammar of images. Kress and Van Leeuwen's social semiotic analysis of images examines the way that metafunc-



tions are realized through semiotic resources in images, for example, how the interpersonal relationship between the image and the viewer operates. The notion of grammar refers to the underlying system which is used by the producer and viewer of the image as resources to produce, for example, interpersonal meaning. Kress and Van Leeuwen identify several ways that the interpersonal metafunction is realized through images. These include the gaze direction of people depicted in images and the composition of the image, which cohere to position the viewer in a certain relationship to the images.

An analyst taking a social semiotic perspective usually begins with identifying semiotic resources, used in an area of social life they are interested in. They then proceed to develop an inventory by collecting, documenting, and cataloguing these semiotic resources and their histories. Inventories can also be produced, or added to, from secondary sources. These sources include descriptions or photographs from specialist and historical literature.

Analysts can also build on their inventories by closely studying how previously identified semiotic resources are used in a given context. Van Leeuwen (2005) points out that inventorizing semiotic resources may produce an understanding of semiotic principles, which are then *inflected* through use in particular situations. He focuses on the example of framing. As a general semiotic principle, framing separates some elements and joins others. When applied in open-plan offices, framing draws on discourses of productivity and teamwork to regulate people's behaviour.

## Multimodal Conversation Analysis

Multimodal Conversation Analysis has its origins in Conversation Analysis (Sacks, 1972; Sacks, Schegloff, & Jefferson, 1974). Early research in Conversation Analysis noted that mostly only one person speaks at a time (Sacks, 1972) and theorized interaction as a series of turns constructed by interlocutors. Rather than treating language as a semiotic resource with various potential meanings, conversation analysts emphasize the sequential and temporal unfolding of social action. They seek to understand the endogenous order of social activity. This order is produced through publicly displayed resources of the body and communicative affordances of the material setting, which participants monitor moment by moment.

Conversation Analysis was first based upon recordings of telephone calls (Sacks, 1967; Schegloff, 1967), and was soon applied to face-to-face interactions (Sacks, 1972; Sacks et al., 1974). From early on, Conversation Analysis has included some consideration of embodied conduct (Goodwin, 1979;

Sacks & Schegloff, 2002); however, the focus has overwhelmingly been upon audible talk, with the addition of bodily conduct. Only more recently has the notion of Multimodal Conversation Analysis become more prevalent (Deppermann, 2013; Mondada, 2006, 2009). Researchers continue to focus on the sequential and temporal unfolding of social action and include embodied and material resources beyond language. Within Conversation Analysis there is ongoing robust discussion regarding a multimodal perspective and the relevance of a notion of mode (Deppermann, 2013), and talk is still often considered the most important feature of interaction (Haddington, Mondada, & Nevile, 2013).

Data are collected for Multimodal Conversation Analysis from naturally occurring interactions. Mondada (2013) points out that naturally occurring data ensures the endogenous order of social activity is not influenced by researcher imposed topics, tasks, or contexts. Furthermore, Mondada argues that although recording devices are continuously present during data collection, they are not always relevant for or commented on by participants.

Data collection in Multimodal Conversation Analysis begins with deciding what and how to record in the context of interest. These decisions require technical, social, and ethical considerations, alongside collaboration with participants (Mondada, 2013). Participant observation is used as preparation for data collection, more so than as a form of data collection in its own right. Researchers establish relationships with participants, determine events to be recorded, and make practical considerations for recording such as camera placement.

We will use Zemel and Koschmann (2014) as an example of a multimodal conversation analytic project. Zemel and Koschmann studied learnability and instructed experience during surgery in a teaching hospital. The data for their study, a 50-second excerpt, comes from the SIU Surgical Education Video Corpus. In this excerpt, an instructional opportunity comes about, is demonstrated for the resident, the resident then enacts what has been learned, and so on. Detailed conversation analytical verbal transcripts lead the way in this kind of multimodal analysis, and five particular images are used to illustrate the multimodal aspects of the points made.

Sidnell (2012) describes the classic conversation analytic procedure, which begins with noticing something of interest in social interaction. This noticing usually comes from exploring a data set through repeated observation. The analyst then seeks out other instances of the phenomenon and identifies its boundaries. A corpus of these instances is produced from audio- and video

recorded data. The final stage of analysis involves describing the features of the phenomenon that are generic and context independent.

## Challenges and Controversial Issues

This chapter shows that there are a variety of well-developed approaches to multimodal analysis. This is not a challenge in itself, as no single approach is suitable for all things. The challenge lies in identifying and remaining aware of how theoretical differences between approaches emerge during data collection and analysis. Here, we identify some of the central theoretical divergences between approaches to multimodal analysis and highlight their relation to data collection and analysis. These include the way historical and psychological aspects are treated in analysis and the status of objects, texts, and artefacts. Lastly, we discuss the definition of mode.

## Historical and Psychological Aspects of Interaction

Multimodal (Inter)action Analysis embeds a historical and psychological aspect into analysis, through the mediated action as a unit of analysis. Within Multimodal (Inter)action Analysis, modal density and the foreground/background continuum of attention/awareness (Norris, 2004) identify a rich landscape of attention/awareness for social actors, allowing for the analysis of simultaneity in action and interaction. Participant observation, field notes, and interviews are forms of data collection that provide further access to the historical and psychological aspects of social action. They help the analyst to understand the actions social actors produce.

Mediated Discourse Analysis also embeds history and psychology, and these aspects are primarily theorized through focusing on the notion of mediation and the primacy of concrete action. As with Multimodal (Inter)action Analysis, mediation implicates mediational means through which actions are produced. These mediational means have psychological and material aspects. In the case of Mediated Discourse Analysis, participant observations are used in data collection to deeply understand spaces where language activities of particular importance occur.

Social Semiotics also includes a historical aspect, by examining how participants engage with semiotic resources that researchers have identified and collecting historical records of semiotic resources (van Leeuwen, 2005). The

psychological aspect of Social Semiotics is not well developed, beyond proposing some level of agency in semiotic action. Social semioticians will use observation and video recording to collect instances where participants use semiotic resources. These data can be analysed to show how semiotic resources can be *inflected* (van Leeuwen, 2005) through use.

In Systemic Functional Multimodal Discourse Analysis, the historical aspect is present in the notion of structured semiotic resources. Regarding a psychological aspect, Systemic Functional Multimodal Discourse Analysis pays little attention to social actors, focusing on cataloguing semiotic resources (Baldry & Thibault, 2006; O'Halloran, 2004). The social actor is theoretically present in artefacts through semiotic choices and reading paths. However, analysis of social actors engaging with artefacts is not required when cataloguing semiotic resources and the notion of a social actor's psychological aspect intersecting with semiotic resources is not addressed. Data collection therefore consists of developing corpora of texts or artefacts and analysing the way semiotic resources function within them.

Multimodal Conversation Analysis focuses on in situ interaction and remains strongly connected to the notion of sequentiality, although some multimodal work raises questions here (Deppermann, 2013). History in Conversation Analysis does not go much further than the preceding turn, or at times larger sequential structure. Because the focus is upon the in situ produced turns, the psychology of interlocutors is not taken into account. Sidnell (2012) points out that Conversation Analysts do not deny differences between people that might influence interaction, but they do not assume that these differences are consequential to the interaction. Rather, for any psychological or historical aspects to enter into analysis, they must be present in the data. Thus, as we have pointed out above, participant observation is used in Multimodal Conversation Analysis as a form of preparation for data analysis, but these observations would not usually constitute data.

## Status of Objects and Artefacts

Multimodal (Inter)action Analysis, Mediated Discourse Analysis, and Conversation Analysis do not promote analysis of objects and artefacts beyond the way interlocutors refer to and act with them. In contrast, Systemic Functional Multimodal Discourse Analysis and, albeit to a lesser extent, Social Semiotics focus on cataloguing the semiotic resources in artefacts and objects. Social Semiotic studies at times include observation and recording of participants' interactions with artefacts of interest, but primacy is usually given to

the semiotic resources identified by the researcher (e.g., Sample Study 28.2). These resources then form the basis for understanding meaning.

### Sample Study 28.2

Wilson, A. A., & Landon-Hayes, M. (2016). A social semiotic analysis of instructional images across academic disciplines. *Visual Communication, 15*(1), 3–31.

#### Background

Wilson and Landon-Hayes use a Social Semiotic approach to analyse the images used by six middle-school teachers across multiple disciplines: language arts, social science, earth science, and mathematics. They use ideas from visual grammar to analyse the interpersonal and ideational metafunction of images used and their relationship to how knowledge and truth are constituted within these disciplines.

#### Research Question

What discipline-specific patterns were realized through images used during earth science, language arts, mathematics, and social sciences instruction?

#### Method

Wilson and Landon-Hayes collected data about patterns used in disciplinary images by working with six teachers. They used purposive sampling, selecting effective and innovative teachers based on recommendations. Each teacher taught in two different subject areas.

The researchers collected four types of data. They observed 354 total lessons and took field notes. At the same time, they photographed all images viewed by the students, including those in textbooks and projected onto the Smartboard. They also collected artefacts from the lessons, such as printouts of PowerPoint presentations and handouts. Lastly, they conducted interviews with each teacher, discussing why they used such images. Teachers ranked the utility and importance of images, compared to other semiotic resources used during instruction.

#### Analysis

Interviews were not formally analysed but provided contextualization for the study. The artefacts and images were uploaded to NVivo 9 and coded for the techniques they used relating to the ideational and interpersonal metafunctions. For the interpersonal metafunction, the codes included techniques like framing, vantage point, and angle. For the ideational metafunction, the codes included techniques like orientation, subject, and use of colour. An initial set of codes were developed based on the literature (e.g., Kress and Van Leeuwen, 2006) and applied to a random subset of the data. The codes were expanded and modified based on the researchers' observations during this initial coding.

The first author then analysed and coded the entire data set, and the second author analysed 10 percent of the data set. Agreement between authors was over 85 percent, which the authors use as a measure of "trustworthiness" of their coding process.

The coding phase of data analysis illustrates how frequently particular techniques are used in images across the data set. A second phase was then carried out to analyse how multiple techniques were used together to produce meaning

in individual images. Two images from each teacher's instruction for each discipline were selected and analysed in detail. The coding from phase one established that these images were representative of the overall trends in the teacher's data set.

### Results

Wilson and Landon-Hayes report their final results by first reporting the prevalence of particular techniques within each discipline. They enrich this reporting with their analyses of meaning produced through multiple techniques. For example, images used for mathematical instruction were least likely to include people, and when they did a distance was maintained between the viewer and the subject. Mathematics images were also decontextualized from the natural environment through the use schematic images with limited use of colour.

Wilson and Landon-Hayes conclude that highlighting assumptions that exist within each discipline can help students understand the metadiscourse of each discipline which can lead to greater success. However, it can also lead to a more critical use of images that can counter some of the dominant assumptions within these disciplines.

## Definition of Mode

The differences between approaches discussed above are visible in the way that mode is defined. We identify two major theoretical positions. The first comes from Multimodal (Inter)action Analysis. Norris (2013) defines mode as a system of mediated action. "Mode" is not viewed as a resource that exists apart from social actors and their embodied and psychological interaction with the physical environment. As we point out above, Multimodal (Inter)action Analysis is designed to integrate physical, historical, psychological, and embodied dimensions in analysis.

Multimodal Conversation Analysis does not have a clear definition of mode. While it is in some respects aligned with the definition of mode in Multimodal (Inter)action Analysis, it lacks the dimensions of psychology and history, clearly embedded in the definition of mode as system of mediated action as discussed above. Yet, in Multimodal Conversation Analysis, "mode" (often referred to as a resource) is also not considered an analytic entity. Rather, a multimodal perspective in Conversation Analysis could be more accurately characterized as including embodied conduct and the material environment alongside talk, to understand the localized production of social order. Thus, the relevance, or not, of different "modes," as in Multimodal (Inter)action Analysis, emerges from the data and is only relevant as far as participants orient towards particular types of movements or talk which might be characterized in other approaches as cohering into "modes."

The second theoretical position on mode comes from a Systemic Functional Linguistic basis, which suggests an underlying structure (grammar) of semiotic resources, language being one. Social Semiotics and Systemic Functional Multimodal Discourse Analysis identify many other accepted modes (Jewitt, 2013) and apply metafunctional analysis to them. Importantly, mode in these terms refers to socially and culturally shaped resources for meaning-making. Thus, they exist in the world and are shared across groups.

This key difference between these two perspectives on mode is a central challenge to working across multimodal approaches. We would like to point out, however, that the challenge lies more in remaining aware of this difference and the impact it has on data collection and particularly analysis, than in trying to dissolve it.

## Conclusion

In this chapter, we have discussed the five most prominent approaches to multimodal data analysis. First, we introduced the origins of each approach and then discussed how data is collected and analysed. We started with Multimodal (Inter)action Analysis, since this is the only approach that was developed for the analysis of multimodal action and interaction. The other approaches that we discussed were developed out of studies of language to take a multimodal perspective. These were introduced in reverse chronological order, starting with Mediated Discourse Analysis, then Systemic Functional Multimodal Discourse Analysis, followed by Social Semiotics, and finishing with Multimodal Conversation Analysis.

We intend for readers to understand the types of data that are collected by researchers using each approach and how analysis proceeds. More importantly, however, we hope that readers will use this chapter as a launch pad to deepen their understanding regarding how the theoretical basis of a particular approach drives decisions about data collection and forms of analysis.

This hope and intention develops out of concern that the many approaches to multimodal analysis in fact represent many different theoretical perspectives, which raise important theoretical issues. We identified three such issues and discussed these in more detail. They are historical and psychological aspects in interaction, the status of objects, texts, and artefacts, and the definition of mode. Readers are encouraged to further explore data, analysis, and theoretical position in more detail, and an annotated bibliography is provided to support these explorations. Finally, we seek to encourage recognition of the diversity in theoretical positions, in order to facilitate robust and coherent multimodal analysis.

## Resources for Further Reading

Baldry, A., & Thibault, P. J. (2006). *Multimodal transcription and text analysis: A multimedia toolkit and associated on-line coursebook*. London and Oakville: Equinox.

This text explores how a Systemic Functional Linguistics approach can be used to analyse a diverse range of multimodal texts. Chapter titles include the printed page, the web page, and film texts and genres. Baldry and Thibault use a range of tools to analyse a large number of texts, demonstrating the tools for students interested in this approach.

Jewitt, C., Bezemer, J. J., & O'Halloran, K. L. (2016). *Introducing multimodality*. London: Routledge.

This text (and accompanying online study guide) introduces three key approaches in multimodality: Systemic Functional Linguistics, Social Semiotics, and Conversation Analysis. Each chapter details the theoretical underpinnings of each approach and methods used for analysis. An additional chapter briefly introduces more recently developed frameworks and combined multimodal concepts with other approaches.

Norris, S. (2004). *Analyzing multimodal interaction: A methodological framework*. London: Routledge.

This text introduces Multimodal (Inter)action Analysis as a methodological framework to use in the study of social interaction. Norris explains the theoretical foundations and methodological tools used within Multimodal (Inter)action Analysis. These methodological tools are explained in detail and applied to video-ethnographic and image-based data collected from various research projects, demonstrating their use and applicability.

Norris, S., & Jones, R. H. (2005). *Discourse in action: Introducing mediated discourse analysis*. New York: Routledge.

This edited volume provides an overview of the theoretical underpinnings of Mediated Discourse as well as a range of studies exemplifying the application of this approach. The book is organized into key concepts relating to Mediated Discourse, and within each section the concept is first defined



before being explored through its application to real studies and data. The sections end with discussion points and ideas for small projects that students can engage with.

Van Leeuwen, T. (2005). *Introducing social semiotics*. London/New York: Routledge.

This text provides an overview of Social Semiotic theory alongside practical examples and exercises. It is organized into three sections: the first focuses on semiotic principles, the second on application and analysis, and the last section examines the relationships across modes. A diverse range of multimodal texts are used as examples, demonstrating the interdisciplinary application of Social Semiotics.

## References

- Al Zidjaly, N. (2009). Agency as an interactive achievement. *Language In Society*, 38(2), 177–200.
- Baldry, A., & Thibault, P. J. (2006). *Multimodal transcription and text analysis: A multimedia toolkit and associated on-line coursebook*. London and Oakville: Equinox.
- Bernad-Mechó, E. (2017). Metadiscourse and topic introductions in an academic lecture: A multimodal insight. *Multimodal Communication*, 6(1), 659.
- Deppermann, A. (2013). Multimodal interaction from a conversation analytic perspective. *Journal of Pragmatics*, 46(1), 1–7.
- de Saint-Georges, I. (2014). Mediated discourse analysis, “embodied learning” and emerging social and professional identities. In C. Maier & S. Norris (Eds.), *Interactions, images and texts* (Vol. 11, 1st ed., pp. 347–356). Boston, MA: De Gruyter.
- Eggs, S. (2004). *Introduction to systemic functional linguistics* (2nd ed.). London: A&C Black.
- Forceville, C., & Urios-Aparisi, E. (2009). *Multimodal metaphor*. Berlin: Walter de Gruyter.
- Geenen, J. (2017). Show and (sometimes) tell: Identity construction and the affordances of video-conferencing. *Multimodal Communication*, 6(1), 1–18.
- Goffman, E. (1959). *The presentation of self in everyday life*. Gloucester, MA: Peter Smith Pub Incorporated.
- Goffman, E. (1981). *Forms of talk*. Philadelphia, PA: University of Pennsylvania Press.

- Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), *Everyday language: Studies in Ethnomethodology* (pp. 97–121). New York: Irvington Publishers.
- Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10), 1489–1522.
- Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.
- Haddington, P., Mondada, L., & Nevile, M. (2013). Being mobile: Interaction on the move. In P. Haddington, L. Mondada, & M. Nevile (Eds.), *Interaction and mobility, language and the body in motion*. Berlin, Boston: De Gruyter.
- Halliday, M. A. K. (1978). *Language as a social semiotic: The social interpretation of language and meaning*. Baltimore, MD: University Park Press.
- Hamilton, H. E. (1994). *Conversations with an Alzheimer's Patient: An interactional sociolinguistic study*. Cambridge: Cambridge University Press.
- Hart, C. (2016). The visual basis of linguistic meaning and its implications for critical discourse studies: Integrating cognitive linguistic and multimodal methods. *Discourse & Society*, 27, 335–350.
- Jewitt, C. (2006). *Technology, literacy and learning: A multimodal approach*. New York: Psychology Press.
- Jewitt, C. (2013). Multimodal methods for researching digital technologies. In *The SAGE handbook of digital technology research*. Thousand Oaks, CA: Sage.
- Kress, G. (1999). Multimodality. In B. Cope & M. Kalantzis (Eds.), *Multiliteracies: Literacy learning and the design of social futures*. New York, NY: Routledge.
- Kress, G., & van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. New York, NY: Routledge.
- Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London: Edward Arnold.
- Kress, G., & van Leeuwen, T. (2002). Colour as a semiotic mode: Notes for a grammar of colour. *Visual Communication*, 1(3), 343–368.
- Kress, G., & van Leeuwen, T. (2006). The semiotic landscape. In P. Griffiths, A. J. Merrison, & A. Bloomer (Eds.), *Language in use—A reader* (pp. 344–350). London: Routledge.
- van Leeuwen, T. (2005). *Introducing social semiotics*. New York, NY: Psychology Press.
- Machin, D., & Mayr, A. (2012). *How to do critical discourse analysis: A multimodal introduction*. London: SAGE.
- Matelau, T. (2014). Vertical identity production and Maori identity. In S. Norris & C. Maier (Eds.), *Interactions, texts and images: A reader in multimodality* (Vol. 1, 1st ed., pp. 255–266). New York: Mouton de Gruyter.
- Mondada, L. (2006). Participants' online analysis and multimodal practices: Projecting the end of the turn and the closing of the sequence. *Discourse Studies*, 8(1), 117–129.
- Mondada, L. (2009). Emergent focused interactions in public places: A systematic analysis of the multimodal achievement of a common interactional space. *Journal of Pragmatics*, 41(10), 1977–1997.

- Mondada, L. (2013). The conversation analytic approach to data collection. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 210–228). Oxford: Wiley-Blackwell.
- Norris, S. (2004). *Analyzing multimodal interaction: A methodological framework*. London: Routledge.
- Norris, S. (2011). *Identity in interaction: Introducing multimodal interaction analysis*. London: Mouton de Gruyter.
- Norris, S. (2013). What is a mode? Smell, olfactory perception, and the notion of mode in multimodal mediated theory. *Multimodal Communication*, 2(2), 155–169.
- Norris, S. (2014). Learning tacit classroom participation. WCTLA. *Procedia Social and Behavioral Sciences*. Amsterdam: Elsevier.
- Norris, S. (forthcoming). *Systematically working with multimodal data: Research methods in multimodal discourse analysis*. Hoboken, NJ: John Wiley and Sons.
- Norris, S., & Pirini, J. P. (2017). Communicating knowledge, getting attention, and negotiating disagreement via videoconferencing technology: A multimodal analysis. *Journal of Organizational Knowledge Communication*, 3(1), 23–48.
- Norris, S. (2006). Multiparty interaction: A multimodal perspective on relevance. *Discourse Studies*, 8(3), 401–421.
- O’ Toole, M. (2004). Opera Ludentes: The Sydney opera house at work and play. In K. O’Halloran (Ed.), *Multimodal discourse analysis: Systemic functional perspectives* (pp. 11–27). London: Continuum.
- O’Halloran, K. (1998). Classroom discourse in mathematics: A multisemiotic analysis. *Linguistics and Education*, 10(3), 359–388.
- O’Halloran, K. (2004). *Multimodal discourse analysis: Systemic functional perspectives*. London: Continuum.
- O’Halloran, K. (2008). Systemic functional-multimodal discourse analysis (SF-MDA): Constructing ideational meaning using language and visual imagery. *Visual Communication*, 7(4), 443–475.
- O’Halloran, K., Marissa, K. L. E., & Tan, S. (2014). Multimodal analytics: Software and visualization techniques for analyzing and interpreting multimodal data. In *The Routledge handbook of multimodal analysis* (pp. 386–396). London: Routledge.
- O’Halloran, K. L., Tan, S., Wignell, P., Bateman, J. A., Pham, D.-S., Grossman, M., & Moere, A. V. (2016). Interpreting text and image relations in violent extremist discourse: A mixed methods approach for big data analytics. *Terrorism and Political Violence*, 1–21.
- Pietikäinen, S., Lane, P., Salo, H., & Laihiala-Kankainen, S. (2011). Frozen actions in the arctic linguistic landscape: A nexus analysis of language processes in visual space. *International Journal of Multilingualism*, 8(4), 277–298.
- Pirini, J. (2014a). Introduction to multimodal (inter)action analysis. In S. Norris & C. Maier (Eds.), *Interactions, texts and images: A reader in multimodality* (Vol. 1, 1st ed., pp. 77–92). New York: Mouton de Gruyter.

- Pirini, J. (2014b). Producing shared attention/awareness in high school tutoring. *Multimodal Communication*, 3(2), 163–179.
- Pirini, J. (2016). Intersubjectivity and materiality: A multimodal perspective. *Multimodal Communication*, 5(1), 1–14.
- Pirini, J. (2017). Agency and Co-production: A multimodal perspective. *Multimodal Communication*, 6(2), 1–20.
- Pirini, J., Norris, S., Geenen, J., & Matelau, T. (2014). Studying social actors: Some thoughts on ethics. In S. Norris & C. Maier (Eds.), *Interactions, texts and images: A reader in multimodality* (Vol. 1, 1st ed., pp. 233–243). New York: Mouton de Gruyter.
- Sacks, H. (1967). The search for help: No one to turn to. In E. Schneidman (Ed.), *Essays in self-destruction*. New York: Science House.
- Sacks, H. (1972). An initial investigation of the usability of conversational data for doing sociology. In *Studies in social interaction*. New York: Free Press.
- Sacks, H., & Schegloff, E. A. (2002). Home position. *Gesture*, 2(2), 133–146.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.
- Schegloff, E. A. (1967). *The first five seconds: The order of conversational openings*. Berkeley: Department of Sociology, University of California.
- Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.
- Scollon, R. (1998). *Mediated discourse as social interaction: A study of news discourse*. London and New York: Longman.
- Scollon, R. (2001). *Mediated discourse: The nexus of practice*. London and New York: Routledge.
- Scollon, R., & Scollon, S. B. K. (2003). *Discourses in place: Language in the material world*. London: Routledge.
- Scollon, R., & Scollon, S. B. K. (2004). *Nexus analysis: Discourse and the emerging internet*. New York, NY: Psychology Press.
- Sidnell, J. (2012). Basic conversation analytic methods. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 77–100). Oxford: Wiley-Blackwell.
- Tannen, D. (1984). *Coherence in spoken and written discourse*. Norwood: ALEX Pub. Corp.
- Tapio, E. (2014). The Marginalisation of finely tuned semiotic practices and misunderstandings in relation to (signed) languages and deafness. *Multimodal Communication*, 3(2), 245.
- Vygotsky, L. S. (1978). *Mind in society*. Cambridge: Harvard University Press.
- Wertsch, J. V. (1998). *Mind as action*. New York: Oxford University Press.
- Zemel, A., & Koschmann, T. (2014). “Put your fingers right in here”: Learnability and instructed experience. *Discourse Studies*, 16(2), 163–183.

## Part IV

### Selected Research Topics and Areas in Applied Linguistics

There are 13 chapters in this part of the Handbook. In each chapter, the authors discuss current and core issues, challenges and controversial issues, and limitations and future directions. Resources for further reading are also provided.

- In Chap. 29 (Instructed Second Language Acquisition), Shawn Loewen presents important constructs, such as implicit and explicit L2 knowledge, noticing, and discourse, that constitute the focus of instructed second language acquisition (ISLA) research. Loewen also addresses various methods that have been used to investigate these constructs.
- In Chap. 30 (Bilingualism and Multilingualism), Tej Bhatia presents a state-of-the-art discussion of various facets of bilingualism and multilingualism. This discussion also entails defining, describing, and measuring bilingualism and bilingual language development. In addition, Bhatia reviews key notions and terms pertinent to the conceptualization of bilingualism and the bilingual mind/brain in order to better understand the complexity and challenges of designing, conducting, and interpreting research projects on bilingualism.
- In Chap. 31 (Forensic Linguistics), Samuel Larner explores some of the challenges and controversies related to carrying out forensic linguistics research. Highlighting that data are central to defining what counts as forensic linguistics, rather than any one methodological approach, Larner also considers the areas of action research, disciplinary engagement, and knowledge mobilization.

- In Chap. 32 (World Englishes), Peter De Costa, Jeffery Maloney, and Dustin Crowther provide an overview of methodologies key to World Englishes inquiry, an empirical field focused on the linguistic and sociocultural dimensions of global English usage. The authors outline three major streams of World Englishes inquiry: Kachruvian-inspired world Englishes (WE), English as an international language (EIL), and English as a lingua franca (ELF). They also discuss the primary elicitation tools employed within and across the three separate streams.
- In Chap. 33 (Heritage, Community, and Indigenous Languages), Shereen Bhalla and Terrence Wiley address the current and historical aspects of data related to the population and perspective of heritage, community, and indigenous language populations in the United States. Bhalla and Wiley also focus on historical patterns of immigration, the role of language, and the subsequent impact on heritage, community, and indigenous language programs.
- In Chap. 34 (Translation and Interpreting), Claudia Angelelli addresses the most salient issues in translation and interpretation (T&I) and qualitative and quantitative methods most commonly used to investigate phenomena in T&I studies against an evolving backdrop of technological advancement and increased communicative needs.
- In Chap. 35 (Identity), Ron Darvin discusses identity as a foundational research area in applied linguistics and examines various methodologies that foreground the role of language in the construction and negotiation of identities. Darvin provides a conceptual and methodological foundation for scholars to pursue identity studies and demonstrates how identity is and will continue to be a rich and productive research area in applied linguistics.
- In Chap. 36 (Gesture Research), Gale Stam and Kimberly Buescher provide an overview of the last several decades of research in second language acquisition (SLA), both inside and outside of the classroom, that examine the speech and gestures of second language (L2) learners and second and foreign language teachers. Stam and Buescher also discuss various research designs and methods, such as how gestures are coded to address the studies' research questions.
- In Chap. 37 (Language Policy and Planning), David Johnson and Crissa Stephens review conceptual and methodological developments in language policy and planning (LPP). Johnson and Stephens also trace the trends and changes in LPP research methodology and underline the importance of making connections across diverse layers and levels, institutional contexts,

and discursive scales to better understand how language policies and plans impact social practices.

- In Chap. 38 (Second Language Pragmatics), Soo Jung Youn discusses current research issues in second language (L2) pragmatic learning, instruction, and assessment. Along with the conceptual and methodological issues, Youn offers a survey of research issues in pragmatic learning, instruction, and assessment, in each case synthesizing the research findings available across diverse contexts focusing on current research.
- In Chap. 39 (Language Testing and Assessment), April Ginter and Kyle McIntosh discuss core concepts in the field of language testing and assessment and provide an overview of the most commonly used research methods in this field. Ginter and McIntosh also discuss issues of test validity as they relate to decision-making purposes in various domains and underscore the growing emphasis on fairness and accountability.
- In Chap. 40 (Linguistic Landscape), David Malinowski unpacks the epistemological and methodological tensions underlying this emerging body of work before going on to describe several current areas of innovation and productive challenges for the future.
- Finally, in Chap. 41 (Researching Academic Literacies), David Bloome, Gilcinei Carvalho, and Sanghee Ryu identify three theoretical perspectives of learning—academic literacy, academic literacies as academic socialization, and academic literacies as chronotopic, dialectical social practices—and examine how each perspective frames a logic of inquiry for conducting research. The authors also explore new ways to investigate new and evolving situations that influence meaning-making.



# 29

## Instructed Second Language Acquisition

Shawn Loewen

### Introduction

Research on instructed second language acquisition (ISLA) is critical in helping provide answers to theoretical and pedagogical questions about second language (L2) acquisition and development. The primary question that ISLA researchers want to investigate is, at its most basic, *What is the best way to learn an L2?* To ask the question in a more nuanced manner, it is helpful to draw on Loewen's (2015, p. 2) definition of ISLA as "a theoretically and empirically based field of academic inquiry that aims to understand how the systematic manipulation of the mechanisms of learning and/or the conditions under which they occur enable or facilitate the development and acquisition of an additional language." Consequently, ISLA research addresses, among other things, (a) cognitive processes, such as input processing and output production, involved in L2 learning, (b) implicit and explicit teaching conditions that assist L2 development, and (c) contextual factors, such as social environments and individual differences. However, for ISLA research to be able to investigate such issues, it is imperative that the research methods used are sound. In other words, the answers are only as good as the research methods used to address the questions.

Traditionally, ISLA researchers have used quantitative methods to address their research questions (Lazaraton, 2000; Loewen & Gass, 2009). However,

---

S. Loewen (✉)

Department of Linguistics and Languages, Michigan State University,  
East Lansing, MI, USA

e-mail: [loewens@msu.edu](mailto:loewens@msu.edu)



there is an increasing amount of qualitative research (e.g., Benson, Chik, Gao, Huang, & Wang, 2009; De Costa, Valmori, & Choi, 2017) as well. Both types of research can provide interesting, albeit different, insights into L2 acquisition; however, because of the predominance of quantitative research in ISLA (and the presence of qualitatively focused chapters in this handbook), this chapter will focus primarily on quantitative methods in its discussion of the central issues related to conducting ISLA research. Furthermore, due to the broad nature of ISLA research, it is impossible to cover every aspect in detail. Consequently, this chapter endeavors to serve as a resource and reference that points readers to additional sources for more extensive information.

## Core Issues

### What Is Being Measured?

A core issue in ISLA research methodology is the nature of the constructs that are being examined. Because the ultimate goal of L2 learning is generally the development of some type of L2 knowledge or proficiency, it is necessary to understand what that proficiency consists of before it can be measured. Two central conceptualizations of L2 proficiency (though by no means the only ones) are that it is comprised of both implicit and explicit L2 knowledge (Ellis, 2005; Rebuschat, 2013), which can also be described as proceduralized and declarative knowledge (DeKeyser, 2015). Although there are theoretical differences between the two sets of terms, what is important for this chapter is the primary characteristics that are shared in both theoretical perspectives. For example, implicit L2 knowledge is unconscious and readily accessible; learners are not aware of this knowledge, but they draw on it for spontaneous L2 production. In contrast, explicit L2 knowledge is conscious and accessed through controlled processing. Learners are able to verbalize this knowledge, and they can draw on it when they have sufficient time to do so. L2 learners' proficiency is generally comprised of both types of knowledge, although not necessarily in equal proportions. Typically, L2 learners who have spent numerous years solely in traditional classrooms have greater amounts of explicit knowledge, while the development of implicit knowledge tends to occur in more communicative contexts, such as classes underpinned by task-based language learning or exposure to speakers of the language outside the classroom. Once researchers have decided which type (or types) of knowledge they want to measure, then they will need to consider how best to measure it.

Although evidence of L2 learning is probably at the heart of ISLA research, there are other constructs that are also of interest, some of which involve cognitive processes. Noticing, awareness, and attention, for example, have been the focus of considerable investigation, due particularly to Schmidt's (2001) noticing hypothesis, which states that very little learning happens without noticing. Although there is debate concerning the exact nature of noticing, awareness, and attention, the important point for ISLA research methods is that they are cognitive processes that must be measured in order to consider their influence on L2 acquisition.

Another construct of interest in ISLA research is L2 learner discourse and interaction. Researchers have been concerned with learner discourse for a variety of reasons, but a primary one is the emphasis that has been placed on communication, due initially, in large part, to Krashen's Monitor Model (1982). Subsequently, learner interaction has been a primary consideration in focus on form (Long, 1996) and task-based language learning (Ellis, 2003) research because interaction is viewed both as a context within which L2 learning occurs and as evidence that L2 acquisition has taken place. While communicative approaches to L2 learning have often been interested in learner interaction, there has also been a concern with the complexity, accuracy, and fluency of both monologic and dialogic learner production (Housen, Kuiken, & Vedder, 2012).

In sum, the more clearly researchers conceptualize the objects of interest and investigation, the better they can then devise methods of measurement. The following section explores the ways in which the previously discussed constructs have been investigated in ISLA research.

## How Is It Being Measured?

After deciding what to measure, ISLA researchers need to consider how best to measure it. Consequently, it is important to consider the specific methodological characteristics of ISLA studies. First, it is possible to make a broad distinction between observational and interventionist research designs. In observational studies, researchers attempt to gain greater understanding of a specific phenomenon or context by observing it, often in its naturally occurring settings. Thus, researchers make no attempt to manipulate the object of investigation. Examples of observational studies include those interested in the frequency and types of spontaneously occurring negotiation for meaning and corrective feedback that occur during meaning-focused L2 interaction. Such studies have been conducted in a variety of instructional contexts,

including French immersion high schools in Canada (e.g., Lyster & Ranta, 1997), ESL classes in New Zealand (e.g., Ellis, Basturkmen, & Loewen, 2001; Loewen, 2004), English conversation classes in Korea (Sheen, 2004), and online computer-mediated communication (e.g., Roushad, Wigglesworth, & Storch, 2016, see Sample Study 29.1). In such studies, researchers examine teacher-student and/or student-student interaction in pedagogical contexts. After the observations, the researchers describe and tally the different types of negotiation and feedback that occurred during the interaction. No specific attempt is made by the researchers to control or direct the interaction.

### Sample Study 29.1

Roushad, A., Wigglesworth, G., & Storch, N. (2016). The nature of negotiations in face-to-face versus computer-mediated communication in pair interactions. *Language Teaching Research*, 20, 514–534.

#### Research Background

This descriptive study uses the interaction approach to investigate the interaction that occurs between L2 learners in face-to-face (FTF) and synchronous computer-mediated communication (SCMC) contexts. Roushad et al. apply etic categories to analyze learner discourse in an effort to reveal characteristics that may be beneficial for L2 learning.

#### Research Problems/Gaps

1. How does the frequency of negotiations for form and meaning compare between FTF and written SCMC modes with same-proficiency pairs? Is there a difference between the two modes in their negotiation type (form vs. meaning)?
2. How does the quality of negotiations (successful uptake and modified output) compare between FTF and written SCMC modes with same-proficiency pairs?

#### Research Method

- Descriptive study.
- Twenty-four intermediate English learners at an Australian college, placed in self-selected pairs.
- Learners performed two information exchange/decision-making tasks: one face-to-face and the other using written SCMC.
- Participants were encouraged to correct their partners during the interaction.
- Interactions were recorded and coded for negotiation type, including clarification request, recast, and metalinguistic corrective feedback. Transcripts were also coded for successful uptake and output modification.
- Analysis involved tallying the number of negotiations per 1000 words to compensate for the difference in time that it took to complete the tasks. Wilcoxon signed rank tests were used to compare the two conditions.

**Key Results**

1. The frequency of negotiations in FTF and SCMC conditions was compared. Negotiation for meaning occurred more frequently in FTF interaction, but no difference was found for negotiation for form.
2. The frequency of successful uptake and output modifications was also compared between the two communication modalities. More successful uptake and output modification occurred in the FTF condition.

**Comments**

This study explores the effects of technology on L2 interaction, which is an important issue due to the ever-increasing role that technology plays in instructed L2 development. Because this study is noninterventionist, it seeks to observe and describe the interaction that occurs in the two contexts rather than manipulate it. Such descriptive studies serve as important baselines for understanding pedagogical phenomena, which then allows researchers to consider how manipulation of the factors involved might improve the developmental potential of the activities.

In contrast to observational studies, interventionist research includes direct efforts by researchers to manipulate the research environment, often through quasi-experimental studies. In these cases, researchers provide learners with one or more types of carefully controlled instructional materials with the goal of determining whether, and to what extent, the treatment, as the instructional materials are often called, leads to L2 learning. To assess learning, researchers often administer a pretest and one or more posttests to determine the effects of the treatment over time. The expectation is that if a specific type of instruction is beneficial for learning, then learners will show improvement from pretest to posttest as a result of being exposed to the instruction.

In addition to examining learner gains from pretest to posttest, researchers often employ comparison or control groups that do not receive the instruction in question. If the treatment group improves their performance over time, but the comparison group does not, researchers can be more confident that the improvement was due to the instruction and not some other moderating or confounding variable. Continuing to use corrective feedback research as an example, several studies have examined the effects of different types of corrective feedback on specific linguistic structures. For example, Ellis, Loewen, and Erlam (2006) compared the effectiveness of recasts and metalinguistic feedback in response to ESL learners' incorrect use of past tense. Ellis et al.'s study involved a pretest, one hour of treatment, an immediate posttest one day after the treatment, and a delayed posttest two weeks later. Ellis et al.

also included a control group that did not receive any corrective feedback. Results indicated that learners receiving metalinguistic feedback improved statistically more than learners in the other groups. (Sample Study 29.2 also provides an example of an interventionist study with Cerezo, Caras, and Leow's (2016) investigation of the effects of explicit L2 instruction.)

### Sample Study 29.2

Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish *gustar* structures. *Studies in Second Language Acquisition*, 38, 265–291.

#### Research Background

This study is a prototypical, quasi-experimental interventionist ISLA study comparing the effectiveness of different types of L2 instruction.

#### Research Problems/Gaps

1. Does the type of instruction (guided induction vs. deductive instruction vs. control) have differential effects on the development of a structurally and cognitively complex L2 structure (Spanish *gustar* structures)? If so, do these effects hold after two weeks?
2. What cognitive processes are employed while processing a complex L2 form during guided induction in an online environment?

#### Research Method

- Participants were 70 university students from intact beginning level Spanish classes. The study had two treatment groups. One group ( $n = 24$ ) received computerized guided induction regarding the Spanish verb *gustar*. The other ( $n = 26$ ) received deductive instruction. A third group ( $n = 20$ ) served as the control group and received no instruction. The treatment lasted between 15 to 20 minutes.
- Learners completed one oral and one written productive test, as well as a receptive multiple choice recognition test. A pretest was administered immediately before the treatment, and an immediate posttest was given immediately after the treatment. A delayed posttest was administered two weeks after the treatment. The tests and treatment were all designed to assess and develop, respectively, explicit knowledge of the target structure.
- A repeated measures ANOVA was used to analyze the test scores. Means and standard deviations were reported for all three groups on all three tests at all three times, and line graphs were used to graphically present the data. ANOVA results were included in the text.
- In addition to the treatment, about half of the participants in the guided induction group were asked to think aloud while engaged in the treatment task. Participants' statements were recorded and transcribed. The verbal protocols were then coded for statements that reflected various categories such as "rule formulation," and "metacognition."

**Key Results**

Both treatment groups outperformed the control group on nearly every post-test. Furthermore, the guided instruction group performed statistically better than the deductive instruction group on several of the posttests. For the think-aloud protocols, the categories of “rule formulation,” and “metacognition” were the most commonly mentioned by participants.

Cerezo et al. conclude: “Our study showed that whereas both traditional deductive instruction in a classroom setting and guided induction via a video game effectively promoted the development of explicit knowledge of the complex *gustar* structures, the latter had an edge, particularly on productive assessment measures that were more cognitively taxing and for the more difficult aspects of the target structure” (p. 288).

**Comments**

This quintessential ISLA study addresses two of the most important constructs: L2 development and learners’ cognitive processes. L2 development is measured quasi-experimentally, using a pretest/posttest design as well as a control group. Furthermore, the study specified the type of L2 knowledge that was being targeted, namely, explicit knowledge. In terms of cognitive processes, the study used concurrent measures to access learners’ thoughts while participating in the treatment activities. Consequently, the authors are able to provide insights into what pedagogical features may be of greater or lesser benefit to learners.

Within interventionist studies, there has been considerable concern with how to measure implicit, and to a lesser extent explicit, L2 knowledge (e.g., Ellis, 2005; Norris & Ortega, 2000). Because explicit knowledge is verbalizable and available under untimed conditions, it is relatively easy to measure. Metalinguistic knowledge tests, untimed grammaticality judgment tests, cloze tests, vocabulary meaning tests are all typically considered good measures of explicit knowledge.

However, implicit knowledge is more difficult to measure, in part because it is not available for conscious introspection, and because learners may rely more on explicit knowledge if they are allowed sufficient time to do so. Nevertheless, implicit knowledge is arguably the more important type of knowledge to measure because it underlies L2 communicative competence which is typically the goal of L2 learning (e.g., Ellis, 2005). The best measure of implicit knowledge is probably spontaneous L2 production; however, ensuring learner production is not always easy, especially if there are specific linguistic items that researchers are targeting. Learners are adept at avoiding the production of difficult structures; nevertheless, the absence of production does not necessarily indicate a lack of knowledge. Thus, tests involving picture description or story-retelling that target specific linguistic structures have

been used to measure implicit knowledge; however, researchers have to ensure that learners do not take excessive time and draw on their explicit knowledge. Other attempts to access implicit knowledge include oral elicited imitation tasks (Erlam, 2006) in which learners hear a stimulus sentence and are asked to repeat it correctly. Self-paced reading and word monitoring tasks, in which researchers measure and compare learners' reaction times when reading grammatical and ungrammatical sentences, have also been proposed as being good measures of implicit knowledge (e.g., Vafaei, Kachisnke, & Suzuki, 2015). Sensitivity to grammatical violations involves greater reading time; however, because learners are focused on the meaning of the sentences, they do not consciously focus on linguistic knowledge. Instead, they rely on their implicit knowledge.

Although ISLA researchers are primarily interested in L2 knowledge and proficiency, there are other constructs of interest because of their potential impact on learning. In terms of cognitive processes, noticing is important and has typically been measured in two ways. One is through online, concurrent measures such as think-aloud protocols (Bowles, 2011), in which learners verbalize their thoughts while they are conducting a task. The benefit of online measures is that they can provide real-time indications of learners' cognitive processes, so such measures are not prone to memory decay. However, there is concern about the reactivity of think-aloud tasks, with learners performing another task (thinking aloud) while they are engaged in the original task (e.g., reading or writing); the act of thinking aloud might increase the cognitive difficulty of doing both tasks, thereby altering participants' performance on the original task. Another drawback is that thinking aloud cannot be done during oral production activities.

The other method of investigating cognitive processes involves off-line, retrospective measures. One such ISLA technique is stimulated recall (Gass & Mackey, 2017) in which learners are provided with recorded excerpts from previous activities and asked to state what they were thinking at the time of the excerpt. Post-activity interviews are another type of retrospective measure that can be used to gain information about learners' noticing. A benefit of retrospective measures is that they do not interfere with the original task and can also be used with oral activities. However, there is concern that memory decay may affect the veridicality or accuracy of recall. Another concern is that learners may not recall what they were thinking or noticing at the time of the stimulus, but rather will comment on their current cognition or invent spurious comments to satisfy the researcher.

In terms of interaction and discourse, there are several methods of analysis. One perspective is to take an emic approach in which researchers attempt to

make sense of interaction from the interlocutors' perspective without imposing any outside categories on the data. Conversation analysis (e.g., Markee, 2000) is a prominent method of analysis in this vein. Another method, and perhaps one more common in ISLA research, is to approach the data with preexisting categories that researchers are looking for in the data. As discussed earlier, research on negotiation for meaning and corrective feedback examines learner interaction in order to identify instances of specific types of interaction, and while some inductive analysis may be conducted, generally researchers search the data for specific discourse moves, such as clarification requests, confirmation checks, and recasts (e.g., Bowles, Toth, & Adams, 2014). Another etic type of analysis is using complexity, accuracy, and fluency measures on either dialogic or monologic learner utterances (Ahmadian & Tavakoli, 2011).

Finally, it should be noted that there are no perfect research methods or designs in ISLA research. The IRIS database ([iris-database.org](http://iris-database.org)), a repository of L2 research instruments, can be helpful in gaining an understanding of the different types of data collection tools that previous studies on a given topic have used (see Marsden, Mackey, & Plonsky, 2016). In addition, researchers should be aware of the options for measuring a given construct, as well as knowing the advantages and disadvantages of using specific instruments and methods. Having this knowledge will allow researchers to focus on construct validity, which is the degree of match between what is being measured and how it is being measured. Because the primary goal of ISLA is to determine if L2 learners are making gains in L2 knowledge, it is first important to consider which type of knowledge researchers want to investigate. Then it is important to devise methods that will accurately and reliably assess such knowledge. For example, if implicit L2 knowledge is the object of investigation, then a meta-linguistic knowledge test would be a poor means of assessment. Rather some measure of implicit L2 knowledge is called for.

The importance of construct validity has been exemplified in ISLA research perhaps most prominently by Norris and Ortega's (2000) meta-analysis, which compared the effectiveness of explicit and implicit L2 instruction for L2 learning. The meta-analysis results indicated that explicit instruction was generally more beneficial and more durable for L2 acquisition; however, Norris and Ortega conceded, and others (e.g., Doughty, 2003) pointed out, that the majority of instruments used to measure L2 learning in the meta-analyzed studies were biased toward explicit instruction because they provided relatively good measures of explicit L2 knowledge. Thus, the results of the meta-analysis were called into question due to the research methods that were used in the original studies.



## Challenges and Controversial Issues

There are many challenges in ISLA research; however, several of the larger issues pertain to the generalizability and quality of research. Furthermore, there are challenges in reconciling different research paradigms. These issues will be considered in some detail.

A primary concern for ISLA research is the generalizability of laboratory research to classroom contexts (e.g., Eckerth, 2009; Gass, Mackey, & Ross-Feldman, 2005). There is a trade-off in conducting research in these two contexts. For classroom research, the benefit is high ecological validity because the research context closely matches the context in which the findings might be applied. However, classrooms, especially if they are not constituted specifically for research purposes, often do not provide the degree of methodological control desired, especially for interventionist studies. Students may be absent for testing or treatment; teachers may not cover the necessary material; nonrandomly selected student samples may be biased, to name a few concerns. An alternative is to conduct research in more controlled laboratory settings, in which students or groups of students volunteer outside of class time or are taken out of their regular classrooms for research purposes. A benefit of laboratory research is that researchers can control numerous moderating and confounding variables, such as the administration of the treatment. However, such control may make the research too far removed from the classroom context with which ISLA researchers are supposedly concerned.

Another challenge is to maintain the quality of research, especially quantitative research (e.g., Norris, 2015; Plonsky, 2013, 2014). There has been increasing concern with the quality of ISLA research because poor quality can lead to ill-informed theoretical and pedagogic conclusions. In particular, research reformers have criticized the field's (over)reliance on null hypothesis significance testing (NHST), which treats the results of quantitative quasi-experimental studies in a dichotomistic manner: either a result is statistically significant or it is not, usually based on a  $p$  value of less than or greater than 0.05 (Norris, 2015). NHST does not indicate the magnitude of effects found in the research, and  $p$  values are strongly influenced by sample size, which is a notorious problem in classroom research. The average ISLA sample size is roughly 19, which is less than optimal for most statistical analyses (Plonsky, 2014). In light of these and other shortcomings, there has been a call for more nuanced approaches to statistical tests in which results are interpreted based on effect sizes, which are less affected by sample size, and provide an indication of the magnitude of an effect.

In addition to the quality of ISLA research, there has been an issue in the reporting of research, with important information omitted on numerous published works (e.g., Larson-Hall & Plonsky, 2015). As far back as Norris and Ortega's (2000) meta-analysis, concerns have been raised about missing descriptive statistics, standard deviations, and other information that is important both for readers to assess the credibility of a study and for meta-analysts to include it in research syntheses. Norris and Ortega, for example, had to exclude 29 of the 78 (38%) relevant studies from their analysis due to insufficient statistical information reported in the primary studies. Although reporting practices, especially the inclusion of effect sizes and standard deviations, have improved with time, there is still considerable room for improvement.

Another challenge in ISLA research is addressing the different ontological and epistemological paradigms represented in quantitative and qualitative research. Quantitative research typically strives to uncover accurate representations of reality using a positivist or postpositivist paradigm. Qualitative research, on the other hand, concerns itself more with describing and detailing the situated perspectives in a given context, using postmodern or constructionist viewpoints (De Costa et al., 2017). As previously mentioned, quantitative research has predominated in ISLA; nevertheless, qualitative research has been increasing. Furthermore, although some researchers claim superiority for certain types of research, it is more common for ISLA researchers to acknowledge the (different) insights that both perspectives can bring. As an example, there has been an increasing interest in mixed methods research (e.g., Riazi & Candlin, 2014), which combines the strengths of both quantitative and qualitative research.

Finally, there have been continued calls for replication research as a means of advancing knowledge of ISLA (e.g., Porte, 2012), and while there has been an increasing number of replication studies, it is not a common practice in ISLA, perhaps due to the pressure from various sources, such as tenure review committees and journal editors, to produce new and original research.

## Limitations and Future Directions

There are at least two general areas for future direction in ISLA research. The first pertains to the continued advancement of technology in terms of both its pedagogical use and its methodological potential. The second direction relates to ISLA researchers' knowledge of methodological practices and how such knowledge is acquired and maintained.

A primary consideration for future research is the role that new technologies will play, both in L2 pedagogy and ISLA research. The increasing use of technology for L2 learning, both inside and outside the classroom, is an important area of investigation because ISLA theories need to be able to account for L2 development in all instructed contexts. One area that has seen some research is the use of synchronous computer-mediated communication by L2 learners, and interaction in such contexts can contribute to L2 development (e.g., Sauro, 2011; Smith, 2004; Ziegler, 2016). Another area that has an established research tradition is computer-assisted language learning (CALL) (e.g., Chapelle, 2010); however, as technology changes in both of these areas, there will be continued need for ISLA researchers to investigate these new instructional contexts. For example, the recent and increasing use of mobile devices for L2 learning calls for research in this area (e.g., Burston, 2015).

In addition to new L2 pedagogical tools, ISLA research will need to continue to draw on advancements in methodological techniques and technologies to provide new insights into existing questions. For example, eye tracking, which measures learners' eye movements as they read or view material on a computer screen, has become a popular method to investigate a variety of ISLA questions (e.g., Winke, Godfroid, & Gass, 2013). Eye-movement data can provide information about how learners process language, as well as what they pay attention to. Thus, eye tracking data can provide an additional source of data, in addition to more traditional types of noticing data such as concurrent (think-alouds) or retrospective (stimulated recall) learner reports.

Other technologies that provide information about physical processes are being used in ISLA research and show potential for further development. For instance, measurements of brain activity, including event-related potentials (ERPs) and functional magnetic resonance imaging (fMRI), can provide information about brain functioning during L2 activity (Tolentino & Tokowicz, 2011). Such information may eventually provide insights into the relationship between implicit and explicit L2 knowledge as researchers discover which parts of the brain are active under conditions in which learners are more likely to draw on one or the other type of knowledge. Another example is the physiological data that can be used to investigate learner emotions, both positive and negative, and their influence on L2 learning and use. In addition to traditional instruments such as the Foreign Language Classroom Anxiety Scale (Horwitz, Horwitz, & Cope, 1986), measures of heart rate and skin conductivity can provide information on learner affect (e.g., Gregersen, MacIntyre, & Meza, 2014).

The other broad area for continued advancement is in the development of competent and knowledgeable researchers. As previously mentioned, there

have been multiple calls for improvements in conducting and reporting quantitative analyses in ISLA research. However, very little is known about the general state of methodological and statistical literacy among ISLA researchers, nor about how such literacy is attained and maintained by graduate students and faculty. Two studies (Lazaraton, Ediger, & Riggenbach, 1987; Loewen et al., 2014) have assessed statistical knowledge in the field, but both relied on self-report data. Just recently, Gonulal (2016), in his study of graduate student statistical knowledge, developed a *Statistical Literacy Assessment for Second Language Acquisition* instrument that measures the ability to interpret statistical knowledge. Such research is an initial step in the right direction, but the field still has little idea about the types of training that graduate students receive in conducting both quantitative and qualitative research. If sound research methods are the basis of knowledge about ISLA theory and pedagogy, then it is imperative to understand how individuals can acquire such knowledge.

Finally, one methodological phenomenon, which reflects the maturing nature of ISLA research, and which shows no sign of diminishing, is research synthesis and meta-analysis. Norris and Ortega (2000) published one of the first meta-analyses in ISLA, and now there are scores of them. A benefit of meta-analyses and other types of research synthesis is that the pool of includable studies continues to grow. For example, many of the early meta-analyses had only a minimal number of studies that they could include, some as small as 10 or 15. However, with the continued increase of ISLA research, meta-analyses are able to include a much larger number of studies, thus yielding a more stable and nuanced set of results. Furthermore, meta-analysts have produced some of the loudest calls for methodological rigor in conducting and reporting ISLA research. In the end, their syntheses are only as good as the initial studies that comprise the meta-analysis.

In summary, this chapter has only scratched the surface of research methods in ISLA. Numerous books have been written on this topic in general and indeed on some of these issues in particular. Consequently, this review is meant to be a springboard and resource for further exploration of this topic rather than a comprehensive summary of it. In my coverage, I have tried to address representative issues, but of course, some researchers may feel that I have ignored or underrepresented their areas of research. I acknowledge that this review reflects my own understanding of ISLA and the important research issues therein. Nevertheless, it is an exciting time to be researching ISLA, as we endeavor to raise the bar to further our knowledge of the best ways to acquire a second language.

## Resources for Further Reading

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol: Multilingual Matters.

This edited volume details a large-scale project to develop and validate tests that are good measures of implicit and explicit L2 knowledge, respectively. Chapters include an overview of the constructs of implicit and explicit L2 knowledge, as well as individual chapters on the tests that were used: grammaticality judgment tests, an oral elicited imitation test, a metalinguistic knowledge test. Several chapters also describe studies that compared the tests to other proficiency tests such as TOEFL and IELTS or other ISLA research studies that used the tests to measure the effectiveness of different instructional techniques.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York: Routledge.

A discipline-specific guide to quantitative research, it incorporates both more conceptual statistical issues, along with practical, step-by-step procedures for conducting commonly used statistical test.

Loewen, S., & Plonsky, L. (2016). *An A–Z of applied linguistics research methods*. New York: Palgrave.

A glossary of research methodology that provides definitions, descriptions, and references for more than 200 methodological terms ranging from *alpha* to *z* score. In addition to providing a definition and/or description, the book references studies that exemplify the term in question. Numerous illustrations of terms such as *bar chart* and *scatterplot* are also provided. Finally, each entry contains several references that provide either an overview of a methodological issue or an example of the term as it is used in an actual study.

Mackey, A., & Gass, S. M. (2012). *Research methods in second language acquisition: A practical guide*. Malden, MA: Wiley-Blackwell.

A hands-on guide for conducting research, using a variety of different methods and theoretical perspectives.

Norris, J. M., Ross, S. J., & Schoonen, R. (Eds.). (2015). Special issue: Currents in language learning series: Improving and extending quantitative reasoning in second language research. *Language Learning*, 65(S1).

A special issue containing ten chapters that address numerous methodological issues, from specific statistical procedures, such as mixed effects modeling, to quality in conducting and reporting quantitative research.

Plonsky, L. (Ed.). (2015). *Advancing quantitative methods in second language research*. New York: Routledge.

An edited volume that introduces readers to the joys of advanced statistical procedures. In addition to chapters on statistical tests like factor analysis, multivariate regression, and structural equation modeling, there are chapters that address the use and misuse of advanced statistics.

## References

- Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful on line planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, 16, 129–149.
- Benson, P., Chik, A., Gao, X., Huang, J., & Wang, W. (2009). Qualitative research in language teaching and language learning journals, 1997–2006. *Modern Language Journal*, 93, 79–90.
- Bowles, M. (2011). *The think aloud controversy in language acquisition research*. New York: Routledge.
- Bowles, M., Toth, P., & Adams, R. (2014). A comparison of L2-L2 and L2-heritage learner interactions in Spanish language classrooms. *Modern Language Journal*, 98, 497–517.
- Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learner outcomes. *ReCALL*, 27, 4–20.
- Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish gustar structures. *Studies in Second Language Acquisition*, 38, 265–291.
- Chapelle, C. (2010). The spread of computer-assisted language learning. *Language Teaching*, 43, 66–74.
- De Costa, P., Valmori, L., & Choi, I. (2017). Qualitative research. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 522–540). New York: Routledge.

- DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 94–112). New York: Routledge.
- Doughty, C. (2003). Instructed SLA: Constraints, compensation, and enhancement. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (2nd ed., pp. 256–310). Malden, MA: Blackwell Publishing.
- Eckerth, J. (2009). Negotiated interaction in the L2 classroom. *Language Teaching*, 42, 109–130.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.
- Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, 27, 141–172.
- Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL. *Language Learning*, 51, 281–318.
- Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, 28, 339–368.
- Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27, 464–491.
- Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research*. New York: Routledge.
- Gass, S. M., Mackey, A., & Ross-Feldman, L. (2005). Task-based interactions in classroom and laboratory setting. *Language Learning*, 55, 575–611.
- Gonulal, T. (2016). *Statistical literacy among second language acquisition graduate students*. Unpublished PhD manuscript, Michigan State University, East Lansing, MI.
- Gregersen, T., MacIntyre, P., & Meza, M. (2014). The motion of emotion: Idiodynamic case studies of learners' foreign language anxiety. *Modern Language Journal*, 98, 574–588.
- Horwitz, E. K., Horwitz, M. B., & Cope, J. A. (1986). Foreign language classroom anxiety. *Modern Language Journal*, 70, 125–132.
- Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency. Complexity, accuracy, and fluency in SLA*. Amsterdam and Philadelphia: John Benjamins.
- Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon.
- Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(S1), 127–159.
- Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly*, 34, 175–181.
- Lazaraton, A., Riggenbach, H., & Ediger, A. (1987). Forming a discipline: Applied linguists' literacy in research methodology and statistics. *TESOL Quarterly*, 21, 263–277.



- Loewen, S. (2004). Uptake in incidental focus on form in meaning-focused ESL lessons. *Language Learning*, 54, 153–187.
- Loewen, S. (2015). *Introduction to instructed second language acquisition*. New York: Routledge.
- Loewen, S., & Gass, S. (2009). Research timeline: The use of statistics in L2 acquisition research. *Language Teaching*, 42, 181–196.
- Loewen, S., Lavolette, B., Spino, L., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, 48, 360–388.
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego: Academic Press.
- Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19, 37–66.
- Markee, N. (2000). *Conversation analysis*. Mahwah, NJ: Erlbaum.
- Marsden, E., Mackey, A., & Plonsky, L. (2016). Breadth and depth: The IRIS repository. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York: Routledge.
- Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417–528.
- Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, 65(S1), 97–126.
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35, 655–687.
- Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, 98, 450–470.
- Porte, G. (2012). *Replication research in applied linguistics*. New York: Cambridge University Press.
- Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, 63, 595–626.
- Riazi, A. M., & Candlin, C. N. (2014). Mixed-method research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching*, 47, 135–173.
- Roushad, A., Wigglesworth, G., & Storch, N. (2016). The nature of negotiations in face-to-face versus computer-mediated communication in pair interactions. *Language Teaching Research*, 20, 514–534.
- Sauro, S. (2011). SCMC for SLA: A research synthesis. *CALICO Journal*, 28, 369–391.
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge: Cambridge University Press.



- Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research*, 8, 263–300.
- Smith, B. (2004). Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition*, 26, 365–398.
- Tolentino, L. C., & Tokowicz, N. (2011). Across languages, space, and time: A review of the role of cross-language similarity in L2 (morpho)syntactic processing as revealed by fMRI and ERP methods. *Studies in Second Language Acquisition*, 33, 91–125.
- Vafae, P., Kachisnke, I., & Suzuki, Y. (2015). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, 39, 59–95.
- Winke, P., Godfroid, A., & Gass, S. M. (2013). Introduction to the special issue: Eye-movement recordings in second language research. *Studies in Second Language Acquisition*, 35, 205–212.
- Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, 38, 553–586.



# 30

## Bilingualism and Multilingualism

Tej K. Bhatia

### Introduction

While the news Brexit from the European Union (June 2016) has sent a tsunami around the globe, which called for revisiting the future and value of globalization, the linguistic winds of globalization are unlikely to dramatically shift the multilingualism landscape as a phenomenon for the following two underlying reasons: (1) Globalization is not an entirely new global phenomenon; it was present in the past in some form or other (e.g., India among others); and (2) bilingualism/multilingualism is a natural and dynamic phenomenon that has been shaped/reshaped with time dating back to precolonial era around the globe (Bhatia, 2017).

In this chapter, the term bilingualism is used as an umbrella that includes both bilingualism and multilingualism for ease of expository expression and efficiency while cautioning readers that bilingualism and multilingualism are two entirely different beasts because of the qualitative (e.g., input types and input conditions) and quantitative differences in becoming bilingual and multilingual.

Contrary to the widespread misconception in some monolingual, or largely English-speaking, countries (e.g., the US), bilingualism is neither an irregular nor a restricted phenomenon. A cursory examination of the global linguistic situation reveals bilingualism is widespread around the globe, manifesting itself in a wide array of diverse languages (e.g., language families, typologically

---

T. K. Bhatia (✉)

Languages, Literatures and Linguistics, Syracuse University, Syracuse, NY, USA  
e-mail: [tkbhatia@syr.edu](mailto:tkbhatia@syr.edu)

distinct and a variety of contact languages). Bilinguals and multilinguals certainly outnumber monolinguals (Bhatia, 2017). Yet it is ironic that monolingualism has been the driving force in research in linguistics for quite some time and is often used as a yardstick to characterize bilingualism. A case in point is bilingual language assessment methods and tools.

## Bilingualism and Multilingualism: Core Issues

On conceptual grounds, two views have primarily been the key drivers and the determinants of research methodology and research practices. Is a bilingual a composite of two monolinguals? Does the bilingual brain comprise two monolinguals crowded into a limited space? Are two bilinguals clones of each other? For some researchers, the answers to these questions have traditionally been affirmative. Such a view of bilingualism is termed the “fractional view.” According to this view, monolingualism holds the key to the understanding of bilingualism. However, a more balanced and accurate picture of bilingualism emerges from the “holistic” view of bilingualism. According to this view, a bilingual person is not a mere sum of two monolinguals, nor is the bilingual brain a composite of two monolingual brains (Dehaene, 1999; Grosjean, 1989). The reason for this position is that the cooperation, competition, and coexistence of the bilingual’s two languages make a bilingual a very complex and colorful individual, which distinguishes him or her from monolinguals, both on quantitative and qualitative grounds (Grosjean, 2010). Furthermore, two bilinguals are not clones of each other because bilingualism is a lifelong language-learning phenomenon—one that is constantly evolving and changing. Individuals become bilinguals by virtue of being exposed to a second or foreign language at various stages of their life for a variety of motivations ranging from emotional to external gains. Therefore, “one size fits all” is not true for bilinguals; they cannot be grouped as a single homogeneous group.

The fractional view is the underlying reason for the frequent misconception and negative perception of bilinguals on the one hand and misguided research on bilingualism and intelligence on the other. The fractional view subscribes to the notion that learning/teaching two languages during childhood contributes to information overload, which is detrimental to the linguistic and cognitive development of children (and perhaps of adults, too). The linguistic and cognitive deficit perception of bilingualism served as a basis for the overwhelmingly negative perception of bilinguals: low intelligence, confused, stuttering, retardation, left-handedness, and even schizophrenia. Fortunately,

later researchers—namely, Peal and Lambert (1962)—corrected the negative perception. In view of their research, the fractional view of bilinguals has been rapidly becoming obsolete because their findings failed to show any significant performance differences between bilinguals and monolinguals on IQ tests, which, in turn, showed how methodologically flawed research led to the erroneous findings and conclusions that monolingual children were smarter than bilingual ones. The most common risks of conducting research on bilinguals and interpreting research data and findings for bilingual language assessment will be discussed in the section devoted to methodological considerations and future challenges.

### **Key Concepts: Describing Bilingualism**

A polyglot is a by-product of three types of bilingualism: individual, social, and political bilingualism. A complex interplay of these types plays an important role in the life of bilinguals. Individual bilingualism is determined by a host of factors (e.g., marriage, immigration, and education), different situations (e.g., input conditions, input types, input modalities, and age), and yields differential end results in terms of differential stages of fossilization and learning curve (U-shape or nonlinear curve during their grammar and interactional development). Social (e.g., language attitudes, stigma attached to language mixing, and language switching) and political factors (e.g., government language policies) further add complexity to measuring bilingualism.

Another major misconception about bilinguals stems from a too-narrow conception of their linguistic knowledge and language use as defined by scholars such as Bloomfield (1933), who claimed a bilingual is one who has “native-like” control over two languages (i.e., “balanced” bilingual). The notion of balanced bilingual overlooks the fact that most bilinguals exhibit a differential control of their two languages. In other words, the linguistic competency bar is set too high for bilinguals.

### **Measuring Bilingualism: Basic Methodological Procedures**

It is not surprising that a widely accepted definition or measure of bilingualism does not exist. Instead, a rich range of scales, dichotomies, and categories are employed to describe bilinguals. A bilingual who can speak and understand two languages is called a “productive” bilingual, whereas a “receptive”

bilingual is an individual who can understand, but cannot speak, a second language. A child who has acquired two languages before the age of five at home (“natural setting”) is called a “simultaneous” or “early” bilingual, whereas those who learn a second language after the age of five, either at home or in school setting, are described as “late” or “sequential” bilinguals. Other labels and dichotomies such as “fluent” vs. “nonfluent,” “balanced” vs. “non-balanced,” “primary” vs. “secondary,” and “partial” vs. “complete” are based upon different types of language proficiency (speaking, writing, listening) or an asymmetrical relationship between the two languages. “Compound” vs. “coordinate” bilingualism refers to the differential processing of language in the brain (Weinreich, 1953). These labels and dichotomies demonstrate the complex attributes of bilingualism, which make the task of defining and measuring bilinguals a daunting one (see Edwards, 1994 for details). Without probing such typologies (labels and/or dichotomies) in detail, the result is inadequate methodologies and flawed outcomes.

The basic procedures adopted to achieve the goal of measuring bilingualism are to conduct both naturalistic/descriptive research as well as experimental research. The naturalistic studies examine bilingual cognitive and linguistic performance in natural settings (e.g., bilingual language switching in actual interactional mode; diary studies of bilingual language acquisition). In contrast with the naturally observed phenomena, an experimental approach aims to manipulate variables to gain insights into how changing variables influence bilingual linguistic performance on tasks such as picture naming, picture narration across two languages, or language mixing. Additionally, experimental studies can lead to correlational measures. The relative strengths and weaknesses of naturalistic and experimental studies will be detailed in the sections on neuro- and psycholinguistic approaches, bilingual language acquisition, and language mixing, with special reference to naturalistic and elicited data in the study of grammatical aspects of language mixing by bilinguals.

Additionally, the typologies of bilinguals are equally important for determining an operational definition of bilingualism in order to devise better methods of measurements and lend reliability and validity of research studies. Insight into typologies equally critical on methodological grounds (e.g., selection of participants and languages for data gathering, determining control groups) and on theoretical grounds in order to come to terms with the question of what is being measured (independent variable) and what is being manipulated (dependent variable); see Phakiti (2014) for experimental research basics and designs.

## Approaches to Researching Bilingualism

Since bilingualism is a very diverse, dynamic, and complex field, there is no single theory or approach that can adequately address the theoretical and methodological issues to gain deeper understanding of the multidimensional facets of a bilingual's production, processing, and comprehension of languages. Bilinguals are currently being studied from cognitive, linguistic, neurological, psycholinguistic, sociolinguistic, and educational linguistic perspectives, to name a few. In spite of their distinct disciplinary, specific theoretical and methodological goals, they converge in their ultimate goal of gaining insights into the bilingual mind/brain by accounting for the following key competencies/characteristics of the bilingual mind/brain: bilingual language choice, control and processing; bilingual language modes; bilingual language mixing and creativity; and bilingual language acquisition.

### Neuro- and Psycholinguistic Approaches

The primary focus of a neurolinguistic approach is to probe bilingual language processing and the organization of language in the bilingual brain. Neurolinguists study typical subjects, with no speech disturbance, and atypical subjects (i.e., impaired—aphasic patients) to gain insight into the neurological dimension of language control—specifically choices and consequences. For comparing and contrasting bilingual language processing, monolingual subjects serve as a bench mark. Both longitudinal and experimental studies are performed.

Rather than examining the social consequences of language choice, neurolinguists focus on language impairments witnessed in aphasic patients (e.g., failure of language matching in language tasks and interaction). Unlike monolinguals, the task of language choice is more complex for bilinguals since their brains need to choose between different languages. It is a widely held belief, at least in some monolingual speech communities, that the process of language choice for bilinguals is random (Ritchie & Bhatia, 2013). The factors that control language choice belong to four categories: participants, situation, topic of discourse, and function of interaction. The following three models address the issues of bilingual brain functions—the declarative/procedural model, the inhibitory control model, and hierarchical/translational model (Green, 1986; Abutalebi & Green, 2008; Ijalaba, Obler, & Chengappa, 2013). Furthermore

the differential domain allocation (e.g., one language for “public” world and another one for “private” world) is central to the bilingual language choices and control, which provides a window into the bilingual language organization in the brain (see Altarriba & Moirier, 2006; Dewaele, 2013; Grosjean, 2010; Pavlenko, 2005). In order to fully comprehend the bilingual brain, the effects of differential domains of bilinguals’ two or more languages must be addressed in depth, which is often slighted in neurolinguistic studies.

Methodological and theoretical issues concern both differential and shared hemispheric involvement of the bilingual brain. Neuroimaging techniques such as functional magnetic resonance imaging (fMRI) and eye tracking (Marian, Spivey, & Hirsch, 2003) provide real-time online evidence about the activity taking place in the different regions of the brain during language processing (Hull & Vaid, 2005). Other experimental methods include the event-related potential (ERP) technique (e.g., examining the brain waves resulting from linguistic stimuli); see Friederici and Thierry (2008). In spite of the advances in neuroimaging and experimental techniques, research findings and conclusions come to contradictory conclusions concerning hemispheric involvement; moreover, many studies support that the two languages are associated with different cortical structures (e.g., Kovelman et al., 2008; Marian, Spivey, & Hirsh, 2003), while other studies support overlapping cortical representation (in the left inferior frontal gyrus) using positron emission tomography (PET) neuroimaging technique (Klien et al., 1995). These differences are due, in part, to differences in methodologies (limitation of the imaging tools), differences in tasks and stimuli, participants (e.g., early bilinguals vs. late bilinguals—age of acquisition), and language typologies (e.g., SOV vs. SVO languages, similarity or difference between the selected languages, role of scripts). Additionally, Grosjean (2008) and Meuter (2005) rightly claim that the failure to ensure natural conditions responsible for the activation of bilingual language mode is a common methodological shortcoming of bilingual aphasic/neurolinguistic research and language testing. Nevertheless, such contradictory findings present an opportunity for neurolinguists to devise better ways of examining hemispheric involvement among different bilingual groups. It is worth pointing out that while fMRI and ERP are largely noninvasive technique, PET requires the use of radioactive tracer; therefore it is not suited for children.

Beyond language organization in the bilingual brain, even more urgent concern of neurolinguists and psycholinguists is to study the two language modes in bilinguals. A number of theories and models have been proposed to describe bilingual language processing in the bilingual mind/cognitive architecture (e.g., coordinate vs. compound bilinguals (Weinreich, 1953), Revised

Hierarchical Model (e.g., Kroll & Dussias, 2013)). Green and Wei (2014) provide insights into the differential neural organization involved in bilingual language switching and language mixing. As is the case with studies devoted to the language organization in the bilingual brain, studies devoted to language processing yield contradictory results, too, essentially for the same reasons identified above. In order to overcome the limitations of the neurolinguistic research paradigm and capture the constantly evolving nature of bilingualism, the most recent model of bilingual language representation and processing is a computational, interactive model of bilingualism, called the Bilingual Language Interaction Network for Comprehension of Speech (BLINCS; Shook & Marian, 2013).

## Bilingual Language Acquisition

In order to gain insights into the bilingual mind/brain and theoretical concerns (e.g., human capacity to acquire languages), researchers need to address a variety of methodological issues specifically grounded in childhood and adult bilingualism. For example, selecting bilingual children for longitudinal and/or cross-sectional studies, controlling dependent and independent variables such as age, gender, social class, ethnicity, and language pairs, among others. Input patterns (single language vs. mixed language) and conditions (naturalistic vs. formal) play a key role in bilingual language development than in monolingual language development. Providing a natural environment or inputs in monolingual/dominant language speech communities is not a challenging task. However, this is not the case in bilingual communities.

Based on the recommendation of educators, among others, bilingual families usually adopt a “one-parent/one-language” strategy with different combinations, such as language allocation based on time and space, for example, using one language in the morning and the other in the evening or one language in the kitchen and another in the living room. This is done to maintain the minority language. In spite of their obvious potential benefits for language maintenance, such strategies fall short in raising bilingual and bicultural children for a number of reasons, including imparting pragmatic and communicative competence and providing negative and positive evidence to children undergoing heritage language development with sociolinguistically real verbal interactional patterns (Bhatia & Ritchie, 1999). Therefore, De Houwer (2007) rightly points out that it is important for children to be receiving language input in the minority language from both parents at home.



## Childhood Bilingualism

The foundation of research on childhood bilingualism was laid in the pioneering work of Leopold (1939–1950). His study was aimed at describing the pattern of English-German language development of the investigator's own daughter, Hildegrad, since childhood. The method used for this longitudinal and naturalistic study was a diary study. Meticulous effort went into Leopold's recording of speech samples. In methodological terms, the main merits of the parent-researchers model are as follows:

1. Parent-researchers model has unlimited continuous access to a subject. Therefore, the risk of attrition is very low.
2. Such studies do not require any formal permission for human subjects (e.g., Institutional Review Board [IRB] permission).
3. Furthermore, such studies are capable of offering rare insights into bilingual language modes.

For example, Leopold reported that Hildegard, while in Germany, came to tears at one point when she could not activate her mother tongue, English. Similarly, Burling (1959), in his longitudinal diary study of his son, Stephen, examines the psychological factors that make a child swing into monolingualism. In contrast, longitudinal naturalistic studies carried out by nonparent researchers are dependent upon the limitations of time and availability of subjects; therefore, such studies experience higher attrition rate than in the parent-researchers model. Also, IRB permission is mandatory. In terms of collect data, the natural consequence of limited time and availability of subjects is the low yield of data.

In spite of their multiple merits, naturalistic diary studies are colored by limitations of what Labov (1972, p. 209) calls “the observer's paradox” (that is, the investigators' own biases in selecting and recording data run the risk of inadvertently collecting nonrepresentative data). Also, it is observed that parents make a lot of transcriptional errors. More powerful methods of ensuring large data is the employment of corpora (e.g., CHILDES and TalkBank corpora) and experimental data. For instance, MacWhinney (2012) developed two widely used databases for cross-linguistic language acquisition research. The CHILDES system comprises a database of transcripts, programs for computer analysis of transcripts, methods for linguistic coding, and systems for linking transcripts to digitized audio and video. The CHILDES database offers a rich variety of computerized transcripts from language learners, ranging from bilingual children, older school-aged children, adult second language learners, children with various types of language

disabilities, and aphasics who are trying to recover from language loss. The system embodies data drawn from 26 languages.

One of the most intriguing aspects of the childhood bilingualism is how children learn to separate the two languages, particularly in a natural setting (i.e., a simultaneous bilingual) in initial stages. This task is not challenging for a monolingual child because only one language serves as a source of input. The two hypotheses which attempt to shed light on the question of language separation by children are the unitary system hypothesis and the dual system hypothesis.

According to the unitary system hypothesis, based on the longitudinal study of the language development of Lisa by Volterra and Taeschner (1978), the child's language undergoes three stages before she/he is able to separate two input languages. During the first two stages, the child experiences confusion—first, an inability to separate the two lexical sets as evidenced by lack of translational equivalents; this is followed by failure to distinguish the grammars of two languages, that is, the fusion stage. It is only during the third stage that the child becomes capable of separating the two sets of vocabularies and grammars.

Findings of recent research reveal that the unitary system hypothesis cannot sustain the scrutiny of the subsequent research, and the evidence motivating the three stages of bilingual language development is full of shortcomings and contradictions, both on methodological and empirical grounds. On methodological grounds, Volterra and Taeschner's subject, Lisa's speech production bears the following age ranges: Stage I: 1:6–1:11 (i.e., from 1 year and 6 months to 1 year and 11 months); Stage II: 2:5–3:3 (i.e., from 2 years and 5 months to 3 years and 3 months); and Stage III: 2:9–3:11 (from 2 years and 9 months to 3 years and 11 months). Clearly the age determinants of the three stages suffer from the problem of overlapping and gaps. Furthermore, the lack of translational equivalents from each language continued to persist even during Stage II and beyond in the speech of Lisa (see Bhatia & Ritchie, 1999, for details).

The dual system hypothesis states that bilingual children, based on their access to universal grammar (UG) and language-specific parameter setting, have the capacity to separate the two grammars and lexical systems right from the beginning. A wide variety of cross-linguistic studies (e.g., different input conditions—one-parent/one-language and mixed input condition—and different word order types) lend support to this hypothesis. In other words, contrary to the claims of the unitary system, the dual system hypothesis present empirical evidence that bilingual children do not go through the stages of lexical and grammatical confusion. In spite of this, the findings of childhood

bilingualism are far from conclusive due to the effects of qualitative and quantitative differences in input provided to a child in two languages. For instance, Leopold's study lends partial support to the unitary system hypothesis. For a more detailed data-driven treatment of the shortcomings of the unitary system hypothesis and the strengths of the dual system hypothesis on methodological grounds, see Bhatia and Ritchie (1999, pp. 591–614) and De Houwer (1990).

Broadly speaking, childhood bilingualism can manifest itself in two distinct patterns: (1) simultaneous bilingualism acquired between birth and four to five years and (2) sequential/successive bilingualism taking place either in schools or in peer groups and/or family settings. Sequential bilingualism can persist throughout the adulthood, thus representing a continuum of language development.

### **Adult Language Acquisition**

In essence, bilingual language acquisition represents a continuum of childhood-adulthood with varying degrees of language proficiency in the second language. What is notable is that in contrast to childhood bilingualism, adults who learn a second language after they have learned their mother tongue experience the learning of a second language as a laborious and conscious task; moreover, adults rarely achieve native-like competency in their second language. For these reasons, second language (L2) learning is viewed as fundamentally different from first language (L1) acquisition based on the critical period hypothesis (Lenneberg, 1967). Access to UG triggers implicit learning/knowledge of the structure that invokes the operation of a principle of UG. Implicit learning takes place unconsciously in a natural environment such as the acquisition of language by children. Explicit knowledge, on the other hand, is relatively conscious and more typical of adults, who require explicit teaching and conscious learning, thus requiring more effort and planning on the part of both teachers and learners. The access of UG to adults is not totally absent; adults have a partial access, if not full access, to UG. Because the implicit knowledge is unconscious, its investigation and measurements pose serious methodological challenges for bilingual research. A variety of methods ranging from observation and experiment studies to neuroimaging techniques are employed to study and measure implicit knowledge of bilinguals. For an experimental approach to investigate the role of implicit knowledge in adults, see Study Box 30.1.

### Sample Study 30.1

Bhatia, T. K., & Ritchie, W. C. (2001). Language mixing, typology, and second language acquisition. *The Yearbook of South Asian Languages and Linguistics*, 37–62.

Bhatia, T. K., & Ritchie, W. C. (2016). Multilingual language mixing and creativity. *Languages*, 1(1), 6. <http://dx.doi.org/10.3390/languages1010006>

#### Research Background

This experimental study explores a new light verb construction which is not the part of existing monolingual grammars of English and Hindi. The bilingual grammar triggers a new phenomenon, which requires the use of dummy verbs such as *karnaa* “to do” with switched embedded-language verbs drawn from English.

#### Research Problem/Gaps

1. What aspect of second language learning does not require any or very little conscious classroom teaching?
2. How bilinguals creatively abstract away from preexisting structures and/or patterns to devise novel ones?

#### Research Method

- Experimental approach to measure acquisition of mixed systems (i.e. Hinglish).
- Language Choice/Pair: Hindi, SOV language; English, SVO.
- Tool: questionnaire containing 29 preference items—24 experimental items and 5 distracters.
- Stimuli: pairs of sentences with Hindi light verb *kar* and the other sentence did not contain this verb. Each pair consisted of either (1) two monolingual sentences or (2) two sentences with code-switched verbs.
- Task: grammaticality judgment—which member of the pair was more grammatical or “correct” than the other.
- Subjects: 42 undergraduate students at a medium-sized US university with Hindi classroom training—maximum of one year.

#### Results

1. The findings of the experimental study underscored the role of unconscious/innate knowledge and implicit learning in the acquisition of a very complex mixed language system (Hinglish).
2. Without any explicit teaching, the subjects employed their implicit knowledge of the Functional Head Constraint to accomplish the grammaticality judgment tasks.
3. The study showed the second language learners’ linguistic competence without any exposure to mixed utterance came close to the grammar of Hindi-English bilinguals.

#### Comments

This experiment made sure that the subjects had no prior knowledge of the structure/stimuli at the time of test taking. For more specific details about the experimental procedures, see Bhatia and Ritchie (2001, pp. 53–55).

Besides UG and the critical period hypothesis, alternative explanations to the variable pattern of bilingual language acquisition are offered on entirely social psychological grounds (Siegel, 2014), which underscore the role of nurture particularly in bilingual language acquisition.

## Linguistic Approaches

The existence of grammars in contact and their consequences represent the core of theoretical and descriptive linguistic studies devoted to bilingualism. One of the natural consequences of grammars in contact is the phenomenon of language mixing/language switching (henceforth, code-mixing), which represents an integral aspect of bilingual verbal behavior. Bilingual children, as well as bilingual adults, use two languages within a single sentence and across sentences. The linguistically oriented studies devoted to the grammar of code-mixing are carried out within a variety of theoretical frameworks (e.g., structural linguistics, generative linguistics, and others). Current approaches are driven by Chomsky's key theoretical construct: linguistic competence and performance within the generative framework in search for the grammar(s) of code-mixing on both descriptive and explanatory grounds (see Chomsky, 1986; MacSwan, 2013).

The unique aspect of the methodological innovation in the Chomskyan competence approach is the use of introspective data in which the investigator can use himself or herself as a subject to elicit grammatical judgments. This marks a point of departure from seeking data by nonintrospective means, similar to experimental studies in neuro-, psycho- and sociolinguistics. For instance, sociolinguists favor nonintrospective empirical data from subjects external to the investigator (e.g., informants, consultants) employing qualitative and quantitative research methods.

On methodological grounds, elicited data and naturalistic data are sought to account for the grammatical facts of code-mixing. While numerous studies, including the recent work by MacSwan and McAlister (2010), employ elicited data techniques to seek grammatical judgments about the code-mixed utterances, naturalistic and experimental data are also employed in bilingual code-mixing research. The latter approach argues that the limitation of the elicited data technique is that it precludes interactional data, thus accounting for a fraction of the bilingual capacity to code-mix in addition to overlooking the social motivations of bilingual code-mixing. Furthermore, due to the social stigma attached to code-mixing, the grammaticality judg-

ment data is not reliable (Mahootian & Santorini, 1996). Naturalistic/experimental data can fill these two gaps in addition to providing insights into the linguistic and social motivations of code-mixing. That is why a number of researchers have begun to develop large corpora of code-mixing data. For example, Hlavac 2000 (<http://www.siarad.org.uk/>) offers a large corpus of hundred Croatian-English bilinguals. Similarly, the *ESRC Centre for Research on Bilingualism in Theory & Practice* at the University of Wales, Bangor, has developed a large body of Welsh-English code-mixing corpus material, which is transcribed and annotated using the CHAT and CLAN application. Also, Language Interaction Data Exchange System (LIDES) by Gardener-Chloros (2009) is another useful database for research on code-mixing.

There are two major limitations of naturalistic/corpora data: (1) the problem of negative evidence (e.g., naturalistic data fail to shed light on ungrammatical utterances—such data offer only the positive evidence); (2) the problem of induction: the absence of a specific pattern implies that it is grammatically impossible. Absence of a form/structure in naturalistic studies does not mean that the structure is not the part of bilingual verbal competence. Grammatical elicitation techniques can overcome these limitations; see MacSwan and McAlister (2010).

## Sociolinguistic and Social Psychological Approaches

Sociolinguistic and social psychological approaches are equally diverse in nature. They aim to explore the social motivations of bilingual language use (e.g., language identity and mechanisms of identity construction, language ideology, language accommodation, language attitudes, social evaluation of speech, language maintenance, and language shift, among others). The scope of sociolinguistic methodology employed in the study of code-mixing, language attitudes, and implicit social biases among others is much broader, with a variety of research tools. One of the notable experimental tools that aim to uncover underlying social biases involved in social evaluation of speech is the groundbreaking matched guise testing (given in Study Box 30.2). Even though this technique was developed approximately half a century ago, it serves as a foundation to test language and dialect biases. Also, this testing technique finds its validity in investigating social discrimination cases (e.g., housing discrimination faced by minorities; see Erard, 2002).

### Sample Study 30.2

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, 60, 44–51.

#### Research Background

This study is a prototypical experimental aimed at tapping unexpressed subjects' language attitudes toward certain social, geographical, ethnic varieties in bilingual communities.

#### Research Method

- Listeners/subjects hear and evaluate speakers' two guises (accents, dialects etc.) without knowing they were spoken by the same speaker.
- Stimuli: Speakers are chosen who could be passed as native in two different varieties/language: English and French in this experiment. The voices of English and French bilinguals were recorded speaking a short paragraph in both languages with same content.
- Task: Subjects ("judges") are asked to rate recorded speakers' voices on personal traits like social class, intelligence, good looks, height, leadership, sense of humor among others.
- Tool: Likert scale and/or cloze test items among others.

#### Key Results

Lambert et al. found that the English speakers evaluated the English voices more favorably than French voices. Surprisingly, the French group evaluated the French guise less favorably than the English guise.

#### Comments

Ingenious research tool; replicated by several cross-cultural and cross-linguist studies. After this initial study, this technique was applied to other sociolinguistic situations involving other varieties of language, for example, AAVE (African-American Vernacular English) and SAE (Standard American English), Chicano English and SAE, Israeli Hebrew as spoken by Arabs vs. Israeli Hebrew as spoken by Jews, and so on. More than hundreds of studies have been inspired by this work. Six-seven-step process goes in matched guise stimuli.

Based on qualitative and quantitative research methods, the linguistic and sociolinguistic motivations for language mixing, both in children and adults, show that the following array of factors prompt bilinguals to mix two languages: semantic domains, stylistic effects, clarification, elaboration, interlocutor's identification, discourse strategies of participants/topics, and other complex sociopsychological reasons such as attitudes, societal values, and personality prompt bilinguals to mix their two languages. The list of motivations is by no means exhaustive; see Ritchie and Bhatia (2013) for more details.

## Social and Political Bilingualism

Social bilingualism refers to the interrelationship between linguistic and nonlinguistic factors, including the social evaluation of bilingualism. In other words, while social bilingualism embodies the linguistic dimensions of individual bilingualism, other factors, including social, attitudinal, educational, and historical considerations, primarily determine the nature of this form of bilingualism. For instance, in some societies (e.g., India) bilingualism is valued and receives positive evaluation and, thus, is encouraged, eventually leading to stable and natural bilingualism. In other societies, however, bilingualism is seen as a negative and divisive force, and it is often suppressed or even banned in public and educational arenas.

Political bilingualism refers to the language policies of a country that suppress language rights of minorities in some countries. The negative effects (e.g., biases) of social and political bilingualism has influenced language testing procedures and tools in the past and continue to face similar problem in spite of the fact language testing has come a way since the IQ testing. In essence, language assessment methodology places bilingual children at risk of low school performance. Clinical and experimental studies often measure the performance of bilingual children on one of their languages, thus giving a fractional or distorted assessment of their linguistic and cognitive development/performance.

## Methodological Considerations and Challenges

On the basis of the preceding discussion of global assessment, key conceptual issues, and approaches, it is clear that serious consideration of these issues is fundamental to designing bilingual research. The long, dark history of research on bilingualism and intelligence testing based on misconstrued research tools and data analysis demonstrate the inadequacy of research paradigms based on fractional and biased views of bilinguals. Peal and Lambert's research (1962) showed how careful methodological consideration of independent variables such socioeconomic class can open a new chapter on bilingualism research leading to a positive view of bilinguals. Their findings revolutionized research on the conception of bilinguals and impacted research on the cognitive, linguistic, and literacy effects of bilingualism (Bialystok, 2013; Hakuta, 1986). The Peal and Lambert study has been replicated around the world, lending support to the methodology and conclusions drawn.

Below, I further underscore what variables to consider when researching bilingualism (e.g., bilingual language organization, language acquisition and use) together with key challenges for future research.



## Language Pairing

Consideration of language pairing is crucial to determine the characteristics of bilinguals and their languages to be tested. For instance, genetically related languages, typologically distinct languages, and different contact languages (e.g., pidgins and creoles, diglossia) can lend differential support for the unitary hypothesis, the dual system hypothesis, and the dominance hypothesis. Such differences need to be reconciled by the utility of rigorous, multifactorial approaches to input types and input conditions.

## Typology of Bilinguals

The typology of bilinguals (e.g., balanced vs. unbalanced; coordinate vs. compound; simultaneous vs. sequential and others) holds the key to the selection of participants (based on individual, social, and political bilingualism), as well as designing tasks and stimuli, which yields reliable and valid results on the measuring and testing of the bilingual's linguistic capacity to use two or more languages.

## Language Modes and Choices

The complex interplay of cognitive, linguistic, and sociocultural factors (e.g., bilingual self-reporting, language accommodation, language identity, mechanism for identity construction, language attitudes, etc.) must receive careful consideration that may bear on the outcome of the study. For instance, any attempt to seek grammaticality judgments on language sentence-internal language mixing is more likely to yield different results in societies with a long history of stable bilinguals on one hand and societies with unstable or recent modes of mixing on the other. Seeking naturalistic data for code-mixing research is not free from methodological challenges with respect to the presence of an external factor—may that be an investigator or a recording instrument—since subjects tend to change their speech under these circumstances, see Labov (1972).

## Language Naming

Language naming poses another challenge for bilingual research that calls for an understanding of the discrepancy often found between actual and reported linguistic situations. For instance, Chinese is not a single language; the label

actually refers to multiple languages such as Mandarin and Cantonese. Hindi and Urdu are not two distinct languages on structural grounds. A deeper understanding of language naming is imperative for avoiding pitfalls in bilingualism research. Language naming is not an intrinsically linguistic distinction but encompasses nonlinguistic distinctions (e.g., identity) as well.

## Future Challenges and Conclusion

The study of bilingualism has posed, and continues to pose, serious challenges to the overall field of linguistics and its interdisciplinary allies. The conceptual and methodological challenges stemming from the divergent theories and research questions/methods are many and multidimensional. In spite of the advances made in the study of bilingualism based on ongoing scrutiny of the conceptual, theoretical, and methodological concerns and the lesson learned from past mistakes, fundamental, theoretical, and methodological problems still haunt the field (McLaughlin, 1984). The lack of terminological and conceptual agreement over the key concepts related to mixed systems (e.g., borrowing, code-mixing, code-switching, pidgins, creoles, and diglossia) and age of language acquisition can interfere with drawing meaningful conclusions and comparisons in interdisciplinary and cross-linguistic research. Finally, the logical and variational problem of bilingual language competence, comprehension, and use is so complex that no single theory or methodology is likely to account for it. The path to trilingualism is even more complex than growing up with two languages (Wang, 2008). Finally, in spite of the advances in neurolinguistics, neuroimaging techniques fail to lend descriptive and explanatory power to inform the current theoretical models of linguistic analyses of bilingualism.

## Resources for Further Reading

Bhatia, T., & Ritchie, W. (Eds.). (2013). *The handbook of bilingualism and multilingualism*. Oxford: Blackwell Publishing Ltd.

This handbook provides state-of-the-art overviews of central issues of bilingualism and plurilingualism and represents a new integration of interdisciplinary research by a team of internationally renowned scholars. The handbook is organized in four parts and comprises 36 chapters, covering neurolinguistics, psycholinguistics, sociolinguistics, and educational aspects of bilingual-

ism. The chapter devoted to methodology is particularly useful for researchers interested in multidisciplinary research methodological approaches to bilingualism.

Grosjean, F. (2010). *Bilingual: Life and reality*. Cambridge, MA: Harvard University Press.

An introductory text dealing with a variety of topics, including lifelong dimensions of bilingual; an interdisciplinary perspective is the hallmark of this book.

Kroll, J., & De Groot, A. M. B. (Eds.). (2005). *Handbook of bilingualism: Psycholinguistic approaches*. Oxford: Oxford University Press.

This handbook fills an important gap by exploring the psycholinguistics, neuro- and cognitive linguistic basis of bilingualism. Methodological and experimental issues are also explored. Like the other handbooks in the field, it is aimed at scholarly and cross-disciplinary audience.

Myers-Scotton, C. (2006). *Multiple voices: An Introduction to bilingualism*. Oxford: Blackwell.

Provides an excellent foundation to topics such as bilingual education, second language teaching and language policies; the production and processing treatment of language mixing is quite technical for an introductory reader.

Weinreich, U. (1953). *Languages in contact: Findings and problems*. The Hague: Mouton.

This work represents a classic in bilingualism. It presents a comprehensive study of grammars in contact, underpinning the mechanism and structural factors in the language development of bilinguals. The role of nonlinguistic factors also receives substantive treatment.

## References

Abutalebi, J., & Green, D. (2008). Control mechanism in bilingual language production: Neural evidence from language switching studies. *Language and Cognition Processes*, 23, 557–582.

- Altarriba, K., & Moirier, R. (2006). Bilingualism: Language, emotions and mental health. In T. Bhatia & W. Ritchie (Eds.), *The handbook of bilingualism* (pp. 250–280). Oxford: Wiley Blackwell.
- Bhatia, T., & Ritchie, W. (1999). The bilingual child: Some issues and perspectives. In W. Ritchie & T. Bhatia (Eds.), *Handbook of child language acquisition* (pp. 569–643). San Diego, CA: Academic Press.
- Bhatia, T. K. (2017). *Bilingualism and multilingualism from a socio-psychological perspective*. Oxford Research Encyclopedia of Linguistics. New York: Oxford University Press.
- Bhatia, T. K., & Ritchie, W. C. (2001). Language mixing, typology, and second language acquisition. *The Yearbook of South Asian Languages and Linguistics*, 37–62.
- Bialystok, E. (2013). The impact of bilingualism on language and literacy development. In T. Bhatia & W. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (pp. 624–648). Oxford: Wiley Blackwell.
- Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.
- Burling, R. (1959). Language development of a Garo and English-speaking child. *Word*, 15, 48–68.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin and use*. New York: Praeger.
- De Houwer, A. (1990). *The acquisition of two languages from birth: A case study*. Cambridge: Cambridge University Press.
- De Houwer, A. (2007). Parental language input patterns and children's bilingual use. *Applied Psycholinguistics*, 28, 411–424.
- Dehaene, S. (1999). Fitting two languages into one brain. *Brain*, 122, 2207–2208.
- Dewaele, J. (2013). *Emotions in multiple languages*. New York: Palgrave.
- Edwards, J. (1994). *Multilingualism*. London: Routledge.
- Erard, M. (2002). Language matters. *Legal Affairs* (July/Aug). Retrieved June 8, 2017, from [https://www.legalaffairs.org/issues/July-August-2002/story\\_erard\\_julaug2002.msp](https://www.legalaffairs.org/issues/July-August-2002/story_erard_julaug2002.msp)
- Friederici, A., & Thierry, G. (Eds.). (2008). *Early language development: Bridging brain and behavior*. Amsterdam: John Benjamins.
- Gardener-Chloros, P. (2009). *Code-switching*. Cambridge: Cambridge University Press.
- Green, D. (1986). Control, activation, and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language*, 27, 210–223.
- Green, D., & Wei, L. (2014). A control process model of code-switching. *Language, Cognition and Neuroscience*, 29, 499–511.
- Grosjean, F. (1989). Neurolinguists! The bilingual is not two monolinguals in one person. *Brain and Language*, 36, 3–15.
- Grosjean, F. (2008). Studying bilinguals: Methodological and conceptual issues. In T. Bhatia & W. Ritchie (Eds.), *The handbook of bilingualism* (pp. 32–63). Oxford: Wiley Blackwell.
- Grosjean, F. (2010). *Bilingual: Life and reality*. Cambridge, MA: Harvard University.
- Hakuta, K. (1986). *Mirror of language*. New York: Basic Books, Inc.

- Hlavac, J. (2000). *Croatian in Melbourne: Lexicon, switching and morphosyntactic features in the speech of second-generation bilinguals*. Ph.D. dissertation, Monash University.
- Hull, R., & Vaid, J. (2005). Clearing the cobwebs from the study of the bilingual brain: Converging evidence from laterality to electrophysiological research. In J. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 480–496). Oxford: Oxford University Press.
- Ijalaba, E., Obler, L. K., & Chengappa, S. (2013). Bilingual aphasia: Theoretical and clinical considerations. In T. Bhatia & W. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (pp. 61–83). Oxford: Wiley Blackwell.
- Klein, D., Milner, B., Zatorre, R., Meyer, E., & Evans, A. (1995). The neural substrates underlying word generation: A bilingual functional-imaging study. *Proceedings of the National Academy of Science, USA*, 92, 2899–2903.
- Kovelman, I., Baker, S. A., & Petitto, L. A. (2008). Bilingual and monolingual brains compared: A functional magnetic resonance imaging investigation of syntactic processing and a possible “neural signature” of bilingualism. *Journal of Cognitive Neuroscience*, 20, 153–169.
- Kroll, J., & Dussias, P. (2013). The comprehension of words and sentences in two languages. In T. Bhatia & W. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (pp. 216–243). Oxford: Wiley Blackwell.
- Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley Press.
- Leopold, W. (1939–1950). *Speech development of a bilingual child: A linguist's record 4 vols*. Evanston, IL: Northwestern University Press.
- MacSwan, J. (2013). Code-switching and grammatical theory. In T. Bhatia & W. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (pp. 323–350). Oxford: Wiley Blackwell.
- MacSwan, J., & McAlister, K. (2010). Naturalistic and elicited data in grammatical studies of codeswitching. *Studies in Hispanic and Lusophone Linguistics*, 3, 521–532.
- MacWhinney, B. (2012). *The CHILDES project: Tools for analyzing talk*. New York: Psychology Press.
- Mahootian, S., & Santorini, B. (1996). Code switching and the complement/adjunct distinction. *Linguistic Inquiry*, 27, 464–479.
- Marian, V., Spivey, M., & Hirsch, J. (2003). Shared and separate systems in bilingual language processing: Converging evidence from eye-tracking and brain imaging. *Brain and Language*, 86, 70–82.
- McLaughlin, B. (1984). Early bilingualism: Methodological and theoretical issues. In M. Paradis & Y. Lebrun (Eds.), *Early bilingualism and child development* (pp. 19–45). Lisse, The Netherlands: Swets and Zeitlinger.

- Meuter, R. (2005). Language selection in bilinguals: Mechanism and processes. In J. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 349–370). Oxford: Oxford University Press.
- Pavlenko, A. (2005). *Emotions and multilingualism*. Cambridge, MA: Cambridge University Press.
- Peal, E., & Lambert, W. E. (1962). Relation of bilingualism to intelligence. *Psychological Monographs*, 76, 1–23.
- Phakiti, A. (2014). *Experimental research methods in language learning*. New York: Bloomsbury.
- Ritchie, W., & Bhatia, T. (2013). Social and psychological factors in language mixing. In T. Bhatia & W. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (pp. 375–390). Oxford: Wiley Blackwell.
- Shook, A., & Marian, V. (2013). The bilingual language interaction network for comprehension of speech. *Bilingualism: Language and Cognition*, 16, 304–324.
- Siegel, J. (2014). Social context. In C. Doughty & M. Long (Eds.), *Handbook of second language acquisition* (pp. 178–223). Oxford: Wiley Blackwell.
- Volterra, V., & Taeschner, T. (1978). The acquisition and development of language by bilingual children. *Journal of Child Language*, 5, 311–326.
- Wang, X. (2008). *Growing up with three languages*. Toronto: Multilingual Matters.
- Weinreich, U. (1953). *Languages in contact: Findings and problems*. The Hague: Mouton.



# 31

## Forensic Linguistics

Samuel Larner

### Introduction

Forensic linguistics is the sub-discipline of applied linguistics that explores the relationship between language, law, and crime. This chapter aims firstly to introduce the field by outlining key areas of research and ways to subdivide the field, before then considering issues surrounding data collection and ethics. The growing importance of statistics is then considered. The chapter argues that a prime motivation of forensic linguistics research is to effect social change and so the areas of action research, disciplinary engagement, and knowledge mobilisation will be discussed, before concluding that despite being a relatively young field in comparison to other areas of applied linguistics, forensic linguistics is a credible sub-discipline which offers plenty of opportunities for exciting and impactful research.

### Characterising Forensic Linguistics Research

Forensic linguistics provides numerous opportunities for research into any area where language, law, and crime intersect. It would be impossible to offer a full and comprehensive survey of all possible areas here, but commonly researched areas include:

---

S. Larner (✉)

Department of Languages, Information and Communications, Manchester Metropolitan University, Manchester, UK

e-mail: [S.Larner@mmu.ac.uk](mailto:S.Larner@mmu.ac.uk)

- the study of legal language and legal genres (e.g., Bhatia, 1993; Finegan, 2010; Tiersma, 2000)
- the comprehensibility of legal language (e.g., Rock, 2007; Tiersma, 2010)
- the language of police interviews (e.g., Haworth, 2010; Heydon, 2005) and trial discourse (e.g., Conley & O'Barr, 1998; Cotterill, 2003; Heffer, 2005)
- producing linguistic evidence of authorship (e.g., Grant, 2010; Larner, 2014b; Wright, 2013) and disputed confessions (e.g., Coulthard, 1994b)
- investigating language crimes such as stalking (e.g., Gales, 2015), bribery (e.g., Shuy, 2013), and malicious communication such as internet trolling (e.g., Hardaker, 2015)

With such diversity in the areas covered by forensic linguists, it is unsurprising that attempts have been made to organise the field into discrete areas. Both Coulthard and Johnson (2007) and Cotterill (2013) independently assert similar ways to characterise the field into two main divisions. “Descriptive” forensic linguistics is concerned with analysing and describing language as it is used at any stage of the legal process (for instance, the way that victims of sexual assault are cross-examined during trials), whilst “investigative” forensic linguistics, by contrast, is concerned with analysing language that has been implicated in a crime (such as attributing a ransom demand to its likely author based on linguistic similarities between the ransom demand and other texts written by the suspect(s)). In investigative forensic linguistics, linguists analyse data with a view to forming an expert opinion that might assist law enforcement agencies and the courts. Linguists produce evidence in a wide range of cases in addition to answering questions of authorship, such as determining the veracity of confessions, clarifying disputed meanings, and determining the level of similarity between competing trademarks.

Coulthard and Johnson (2010) revised their original (2007) distinction between descriptive and investigative forensic linguistics to more fully capture the extent of forensic linguistics research and to more clearly differentiate between analyses of written and spoken language. To this end, they proposed that forensic linguists are concerned with (1) written legal language, (2) interaction in the legal process, and (3) providing testimony as expert witnesses. This final category would appear to be distinct from the first two, seemingly being practice-based rather research-based. To provide some context, historically, academics did not train as forensic linguists. For instance, prior to the 1990s, Malcolm Coulthard, the world's first professor



of forensic linguistics, was perhaps most recognised for his expertise in spoken and written discourse analysis (Tomblin, 2013). Linguists were occasionally approached by lawyers with a specific question they wanted answering as part of their evidence, and Coulthard (1994a) explains that investigative casework “was usually undertaken as an intellectual challenge and almost always required the creation, rather than simply the application, of a method of analysis” (p. 27). By comparison, specialism in forensic linguistics (through, for instance, doctoral research) is a relatively recent trend, with an increasing amount of research attention being paid to the validity and reliability of methods in authorship analysis (e.g., Grant, 2007; Turell, 2010) and the role of expert linguistics testimony in evidential contexts (e.g., Coulthard, 2004, 2010; Solan & Tiersma, 2004). As such, there is now a strong research tradition surrounding the most appropriate ways to produce linguistic evidence, and the category of “providing evidence as an expert witness” therefore legitimately encapsulates research activity rather than being purely practice-based.

Whilst an individual forensic linguistics researcher may show preference for one particular methodological approach, the wider field of forensic linguistics draws on the full range of these, rather than one particular approach dominating. Conversation analysis, discourse analysis, critical discourse analysis, genre analysis, and pragmatic analysis are all frequently used, with some researchers adopting an integrated approach, for example, Haworth (2006), who drew on conversation analysis, critical discourse analysis, and pragmatics to explore the balance of power and control in a police interview context. Based on his own casework, Coulthard (1994a) argued “that any improved methodology must depend, to a large extent, on the setting up and analysing of corpora” (p. 40), and in recent years, corpus linguistics has become especially prominent in both descriptive and investigative research. The role of corpora in forensic linguistics has been widely discussed (Cotterill, 2013; Kredens & Coulthard, 2012), and Larner (2015) has argued that since critical discourse analysis is already central to exploring language, power, and disadvantage in legal contexts, forensic linguistics research is likely to move toward corpus-assisted discourse analysis and corpus-based critical discourse analysis. The key point is that the field of forensic linguistics cannot easily be categorised by a single research approach or methodology. Rather, what defines the field is either (1) the collection and analysis of forensic data and/or (2) the potential application of the research to a particular forensic context.

## Identifying and Collecting Data

Since the field is largely defined by the data being investigated, issues around data collection and ethics are at the forefront of forensic linguistics research. Negotiating access to authentic, attested forensic data can be difficult, time-consuming, and costly. As an alternative, many researchers use data from highly publicised trials, where transcripts have already been prepared and are freely available (e.g., Cotterill, 2003; Haworth, 2006; Tkačuková, 2015). An additional range of online archives also provide access to different types of data. For instance, London's Central Criminal Court, the Old Bailey, hosts an archive of 197,745 historical proceedings held between 1674 and 1913 ([www.oldbaileyonline.org](http://www.oldbaileyonline.org)), and Brigham Young University has made available a corpus of 32,000 US Supreme Court decisions totalling 130 million words, spanning 1790–present (<http://corpus.byu.edu/scotus>). The website [www.famous-trials.com](http://www.famous-trials.com) hosts a range of materials, including case histories and media coverage, as well as legal documents such as court transcripts and transcripts of emergency calls, from a selection of high-profile trials between 399 BC and 2013.

Whatever form the data takes or from wherever it is collected, it is important to remember that as legal language, it will constitute institutional discourse, and so it is essential to understand the institutional context in which it was produced before beginning any analysis. For instance, it would be unwise to analyse police interview discourse without first understanding the relevant legislation and best practice that governs how those interviews should be conducted within the relevant jurisdiction.

The above consideration relates to authentic and attested forensic data—that is, data which genuinely occurred in some forensic context. An alternative approach, appropriate for some types of research, is to use non-forensic data: that is, data which have not originated in a forensic context but which can be argued to replicate forensic data. This is clearly less appropriate for some types of project—most typically, those studies that could be characterised as descriptive forensic linguistics. For instance, a project to explore narratives produced by victims of sexual abuse during police interviews could only claim ecological validity—the extent to which findings are generalizable to real-world contexts—if the data were authentic (although cf. Grant and MacLeod (2016), who argue that experimental data can be used instead of authentic forensic data in some contexts). However, in investigative forensic linguistics, the data are often classified as “forensic” simply because they are implicated in a crime, but in reality, they are typical everyday texts such as

e-mails (see Sample Study 31.1). For instance, in cases of authorship analysis, a text of unknown authorship which is somehow implicated in a crime (e.g., a potentially inauthentic suicide note, an anonymous terrorist threat letter, or an essay suspected to be plagiarised) is compared against a corpus of texts known to have been written by the suspect(s). The corpus of known texts will largely consist of anything the forensic linguist can get access to, including e-mails between friends and colleagues, professional letters, text messages, and so on. Crucially, these will be historical texts, produced over a period of time and not for the purposes of the specific authorship investigation. Therefore, research which seeks to explore methods of authorship analysis can legitimately use data which does not originate from authentic forensic contexts (e.g., Lerner, 2014b) but which has the potential to be “forensic”—in other words, it is not inconceivable that the data could be implicated in a crime.

### Sample Study 31.1

Wright, D. (2013). Stylistic variation within genre conventions in the Enron email corpus: Developing a text-sensitive methodology for authorship research. *International Journal of Speech, Language and the Law*, 20, 45–75.

#### Research Background

Individual linguistic variation—known as *idiolect*—underpins the majority of authorship analysis casework and research. However, there is limited empirical evidence into how it can best be identified. Wright argues that focussing on an author’s linguistic behaviour within specific genre conventions provides a strong platform for investigating idiolect. Specifically, this research sets out to explore the extent to which authors show individual variation in the way they formulate greetings and farewells in professional e-mails. The e-mails were written by employees of Enron, a former American energy company, which was investigated and found guilty of accounting fraud.

#### Research Problems

1. How distinctive are four authors’ linguistic choices in e-mail greetings and farewells?
2. How distinctive are those linguistic choices in comparison to a larger corpus of other authors?

#### Research Method

- *Type of Research*: Quantitative and qualitative.
- *Setting and Participants*: 2622 e-mails written by four employees. A further 40,236 e-mails written by 126 employees for comparison.
- *Instruments/Techniques*: Corpus linguistics techniques were used to extract and analyse patterns in greetings and farewells.

- *Data Analysis:* Greetings and farewells were analysed qualitatively and computationally for lexical form, punctuation, line spacing, and capitalisation. Greetings and farewells used by the four authors were then compared against the other 126 employees to determine distinctiveness. Likelihood ratios (a statistical test) were calculated to test rarity, expectancy, and distinctiveness of the features.

#### **Key Results**

Wright identified 19 greeting and 12 farewell forms that were distinctive to each of the four authors. When compared to e-mails by the other 126 authors, 18 of the greetings and 9 of the farewells were shown to be distinctive to the author who used them. Overall, Wright found that greetings were more distinctive than farewells but the distinctiveness of individual forms became considerably higher when individual greetings and farewells were combined for each author.

#### **Comments**

Wright's results indicate that the ways in which authors write e-mail greetings and farewells "are powerful in distinguishing between the writings of a small subset of (in this case) four authors" (p. 72), suggesting that exploring individual author variation within genre conventions has strong potential for characterising idiolect. Importantly, Wright's methods have ecological validity because e-mails constitute a large proportion of the texts available to forensic linguists carrying out casework. Furthermore, the e-mails he analysed constituted a huge body of evidence in the investigation of this company, meaning the data he used was attested, naturally occurring forensic data.

## **Ethical Issues in Data Collection**

Having identified data collection as a core issue in forensic linguistics research, it is now prudent to discuss the fact that this same core issue is perhaps the most challenging, and indeed controversial, to be faced by the researcher. To ensure ecological validity, there is often a need to use data from participants who have been designated as vulnerable (e.g., minors, victims of sexual abuse). As such, fully considering and navigating ethical issues can be especially problematic (De Costa, 2016). Arguably the biggest challenge is to use data from participants who in many cases are unable to provide informed consent. To illustrate this point, research has explored interactions between the police and victims of sexual abuse (e.g., MacLeod, 2016). Gaining access to such data is challenging in itself, since it is not generally freely available and, as sensitive personal data, access will be heavily restricted. Negotiating access to such data

is a lengthy process, which requires strong and productive relationships to be established with relevant organisations. Moreover, at the point when a person reports that they have been sexually abused, they do not give their consent to have the transcripts of their interviews used for specific research purposes. Nor are they forewarned that any interaction with the police could potentially be used in research more generally. Since informed consent is central to empirical research, researchers must consider whether their use of such data is ethical. Indeed in the European Union, the collection, processing, and storage of any sensitive personal data are bound by the principles outlined in the General Data Protection Regulation (<https://eugdpr.org>), so in addition to ethical principles, researchers are bound by legislation to collect informed consent.

In situations where it is not possible to gain informed consent, it is potentially still possible to collect, analyse, and store such data, provided that the data are used only for research purposes, that the research will not cause damage or distress to any participant, and that the research will not influence decisions made about the participant (e.g., whether they are permitted access to legal services). It should additionally be noted that only data which is directly relevant to the research can be collected: In an investigation of the linguistic strategies used by victims reporting sexual abuse to the police, it may be appropriate to collect information about the gender and age of the victim along with the transcript of their police interview, but their occupation, physical description, or contact details would be irrelevant and therefore must not be collected.

To further justify collecting data without informed consent from the participants, researchers must also demonstrate that the findings will be in the substantial public interest. This may be difficult if the researcher intends to produce only an academic journal article. However, if a stronger dissemination strategy is proposed, involving communicating with personnel from within the institution (in this example, police officers), with a view to effecting change through producing guidelines on best practice for these particular interactions, then a stronger case may be made. In this way, gaining support from the external organisation who provide the data is important for showing that there is belief in the value of the research and that they are awaiting the research outcomes in order to improve their practice (see discussion of knowledge mobilisation, below).

A further potential issue surrounds the use of a privacy notice. Ordinarily, it would be expected that anyone whose sensitive personal data could potentially be used in research should be informed before the data is collected. However, again, there are some circumstances where this is not practically

appropriate, as in the example provided above. The requirement for a privacy notice may be circumvented, provided that the researcher is not collecting the data personally (i.e., it is collected by another party, in this example, the police), and if that other party does not collect accurate and reliable contact details of participants, or any at all, it would involve disproportionate effort to provide individuals with a privacy notice. This would not be the case in the example above but might apply to fraudulent and hoax calls to the police, for instance, where contact details could not be collected reliably.

The discussion of sensitive personal data is not presented here as a way to simplify or circumvent the ethics process, but rather to demonstrate that collecting forensic data is fraught with both ethical and legal difficulties that need to be fully understood before embarking on a forensic linguistics research project. Additionally, due to the difficulty of gaining ethical approval to collect some types of forensic data, it should be noted that some researchers have faced problems in how research based on such data may be disseminated. Some journal editors are unwilling to publish research based on authentic and attested forensic data where informed consent was not collected from participants, and some doctoral theses have been placed under embargo in order to protect the rights of the participants. It is therefore important for forensic linguistics researchers to consider this at the early planning stage, to ensure that their research can most effectively be disseminated.

## Quantitative Analysis and Statistics

In terms of methods of data analysis, the challenges that a forensic linguistics researcher may face are not necessarily different from other areas of applied linguistics; that is, a forensic linguist applying conversation analysis to courtroom data is unlikely to face challenges that differ from a linguist carrying out conversation analysis in other areas of applied linguistics (e.g., data derived from healthcare settings or educational contexts). However, in the domain of investigative forensic linguistics, the main controversy rests with whether the methods used to determine authorship are rigorous enough, with some forensic linguists questioning whether methods are sufficiently developed to be used as evidence in court (Kniffka, 2007; Shuy, 2006). For this reason, quantification, and particularly statistical analyses of data, has become increasingly important as a way to add scientific rigour to methods, leading to more objective and reliable results than can typically be obtained through qualitative research.

A potential problem with the use of statistics is that linguists are often trained in the humanities rather than the sciences, and so it is possible that scientific principles may be misunderstood and/or misapplied, particularly issues regarding sampling, reliability, and validity. For instance, Chaski's (2001) authorship research was heavily criticised by Grant and Baker (2001) on the basis of the reliability and validity of the features investigated, whilst the CUSUM Technique<sup>1</sup> has been largely discredited, because, amongst other reasons, the analysis was not replicable (Hardcastle, 1997). Grant (2007) is a pioneer of highlighting the need for linguists to exercise caution in this regard, particularly of stressing the need for ensuring that statistical analysis is fully integrated into the research design, rather than being an afterthought, which "leads to weak statistical analysis, doubtful conclusions, and a lack of apparent seriousness in attempts at quantification" (p. 3).

A further problem with the use of statistics concerns the sorts of texts that forensic linguists analyse; typically, these are very short. As such, the statistical testing of data may be problematic since statistics require minimum thresholds before the tests carry validity. A solution is to draw on the types of statistical tests devised for small data sets in other disciplines. Grant (2010), when establishing the authorship of SMS text messages, drew upon a statistical test used in psychology to establish case linkage in crimes; Jaccard's coefficient, which is suited to short texts, and following Grant's work, is therefore growing in popularity amongst forensic linguistics researchers (e.g., Johnson & Wright, 2014; Larnar, 2014b).

Jaccard's coefficient establishes the strength of correlation between whether a series of particular features are present in a sample of texts, rather than the frequency with which particular features occur, meaning that low frequency scores are unproblematic. In the example of case linkage from psychology, features may include the presence or absence of offender behaviours such as whether the perpetrator used a weapon or blindfolded the victim and how the perpetrator left the scene of the crime and avoided detection (e.g., by wearing a mask and destroying semen) (Woodhams, Grant, & Price, 2007). In the authorship of text messages, this may include the presence or absence of specific features such as particular spellings, abbreviations, and formatting conventions (Grant, 2010).

Whilst it is true that some linguists remain faithful to a particular analytical paradigm—quantitative or qualitative—Coulthard and Johnson (2007) comment that "[i]t is not unusual for the expert to use more than one approach" (p. 173) and Solan and Tiersma (2004, 2005) advocate the use of "eclectic approaches" which combine the strengths of both qualitative and quantitative methods.

## Action Research, Disciplinary Engagement, and Knowledge Mobilisation

Although not a feature of all forensic linguistics research, a large portion can be classified as action research. Robson (2011) explains that the purpose of action research is to “make a difference to the lives and situations of those involved in the study and/or others” (p. 175). He further explains that improvement is at the heart of action research: “There is, firstly, the improvement of a *practice* of some kind; secondly, the improvement of the *understanding* of a practice by its practitioners; and thirdly, the improvement of the *situation* in which the practice takes place” (p. 188, *original emphasis*). Coulthard, Johnson, and Wright (2017) explain that as an applied discipline, forensic linguistics “has real-world applications and its findings can be applied in professional practice” (p. 14). As such, for many forensic linguists, a key objective of research is to improve some aspect of the criminal justice system (see Sample Study 31.2). To do this successfully and effectively, forensic linguists who wish to effect some form of positive change need to consider building relationships and networks with those in positions to enact change—such as judges, police officers, or politicians—early on in the research planning phase.

### Sample Study 31.2

MacLeod, N. (2016). “I thought I’d be safe there”: Pre-empting blame in the talk of women reporting rape. *Journal of Pragmatics*, 96, 96–109.

#### Research Background

Investigative interviews between the police and suspects or victims are characterised by a power imbalance, where the police officer, acting in an institutional role, controls the interaction. MacLeod argues that despite reforms in interview procedure, there is a mismatch between agendas, with police officers orienting to institutional practices, of which interviewees are typically unaware. In this article, MacLeod analyses police interview discourse, focussing on victims of rape and how they blame themselves for what happened. Her aim is to present “a systematic examination of how rape myths operate in the talk of victims themselves.” (p. 98).

#### Research Problems

1. To what extent are patterns of self-blame evident in the talk of women reporting themselves as victims of rape?
2. To what extent are these patterns rooted in victim-blaming ideology?

#### Research Method

- *Type of research*: Qualitative.
- *Setting and participants*: Transcripts from six video-recorded interviews between police officers and women reporting that they had been raped.



- *Instruments/techniques*: Data were analysed using discourse analysis based on discursive psychology, which explores how psychological phenomena such as intentions and motives are managed in talk.
- *Data analysis*: Points where speakers provided an account of their own behaviour in the transcripts were analysed; examples were categorised into key thematic categories, for example, the victims' reckless behaviour, their use of drugs and alcohol, their prior relationship with the suspect, and whether they provided "appropriate" resistance to the attack.

### Key Results

MacLeod's analysis revealed that female victims "often anticipate a requirement to account for their reported actions" (p. 107) and that in demonstrating "an awareness of the cultural norms and expectations surrounding sexual violence," female victims of rape "pre-empt and mitigate potential blame implications as a routine part of their contributions to the on-going talk" (p. 107). She argues that since the same patterns of self-blame recurred throughout the data, this is evidence that rape myths about blame attribution are deeply ingrained in society. Importantly, victims' accounts of self-blame were rarely challenged by the interviewers, instead being treated as relevant contributions.

### Comments

This research illustrates the important relationship between researchers and the police; indeed, that the police provided such sensitive data reveals a willingness to participate in, and learn from, research. MacLeod specifically discusses the implications of her findings for police practice: If women display patterns of self-blame in interviews, they will be less well positioned to challenge the implications of this during legal proceedings, leading to disadvantage in terms of access to justice. Disseminating the findings of this research, perhaps through police officer training, could serve to improve the experience and outcome of rape victims in the investigative process.

Due to the focus on action research, coupled with the fact that several disciplines are interested in the criminal justice system, interdisciplinarity is at the heart of a great deal of forensic linguistics research. Through a comparison of two forensic linguistics handbooks, Lerner (2014a) foregrounds the issue of disciplinary engagement in forensic linguistics, arguing that whilst there is a strong sense of interdisciplinarity in the field, the emphasis to date has largely been placed on the fact that some academics have previous professional experience (e.g., former police officers, lawyers) or that academics from other professions (e.g., psychologists, sociologists, criminologists) have published in the forensic linguistics literature, rather than actually exploring "the ways in which methods and ideas from different disciplines have really been integrated and synthesised—the goal of true interdisciplinary research" (p. 196). To ensure greater social impact, Lerner (2014a) argues that it is time to move beyond interdisciplinary research toward transdisciplinary research, defined as "the cooperation of academics, stakeholders, and practitioners to solve complex societal ... problems

of common interest with the goal of resolving them by designing and implementing public policy” (Repko, Szostak, & Buchberger, 2014, p. 36).

What this really speaks to is the issue of jointly developing research projects between the academic community and practitioners (such as the police). Moreover, there is a need to share knowledge derived from academic research effectively with the anticipated users of that knowledge and, crucially, with those in a position to change policy—the domain of knowledge mobilisation. Bannister and Hardill (2013) explain that knowledge mobilisation is about “turning research into action through exchange and linkage” (Bannister & Hardill, 2013, p. 5). They further state that the goal “is to develop a culture of partnership between academic researchers and decision-makers to assist in strengthening the development of policy, practice and social innovation, or the co-production of knowledge” (p. 5). Viewed in this way, it is clear to see how the applied field of forensic linguistics can enable stronger research partnerships which directly address problems in the justice system, enabling linguists to achieve considerable impact from their research.

## Conclusion

Despite being a relatively young field of linguistic enquiry, with the first dedicated journal being founded in 1994 (*Forensic Linguistics*, now *The International Journal of Speech, Language and the Law*), alongside a professional organisation, the International Association of Forensic Linguists, the field is experiencing exponential growth. A bilingual (English/Portuguese) journal was launched in 2014 (*Language and Law/Linguagem e Direito*). In addition, two handbooks (Coulthard & Johnson, 2010; Tiersma & Solan, 2012), several textbooks (e.g., Coulthard et al., 2017; Solan & Tiersma, 2005), along with numerous edited collections and monographs on all aspects of the criminal justice system have been published. Through engaging with exciting areas of research, forensic linguists have a real opportunity to effect change in society, making forensic linguistics a truly applied discipline.

## Resources for Further Reading

Coulthard, M., Johnson, A., & Wright, D. (2017). *An introduction to forensic linguistics: Language in evidence* (2nd ed.). Abingdon, Oxon: Routledge.

Into its second edition, this introductory textbook is an invaluable resource, which expertly speaks to undergraduate and postgraduate students, as well as

forensic linguistics researchers, both established and new. The book is organised into two parts: (1) the language of the legal process, which covers a range of descriptive forensic linguistics topics such as the language of the law, emergency calls to the police, and police interviews and trial discourse; and (2) language as evidence, which covers a range of investigative forensic linguistics topics, including authorship attribution, plagiarism, and forensic phonetics. Each chapter contains recommended further reading and a series of research tasks. All topics covered are comprehensive and concepts are well exemplified through a wide range of examples drawn from attested forensic data.

Coulthard, M., & Johnson, A. (Eds.). (2010). *The Routledge handbook of forensic linguistics*. Abingdon, Oxon: Routledge.

This edited collection is organised into three sections: (1) the language of the law and the legal process, (2) the linguist as expert in legal processes, and (3) new debates and new directions. The 39 chapters contained within cover a wide range of topics in descriptive and investigative forensic linguistics, and as such, this handbook provides an extensive survey of the field. Each of the chapters is written by a leading authority, with each chapter ending with recommended additional reading. As such, it is an excellent resource for those new to the field. The introductory chapter is particularly noteworthy for presenting the current debates in forensic linguistics, whilst the concluding chapter predicts the future direction of the field. It is significant—and testament to the authority of this handbook—that these two chapters remain current, despite being written in 2010.

## Note

1. The CUSUM technique claimed to be a scientific method for identifying authorship in which the cumulative sums of features such as the number of two- and three-letter words and number of words beginning with a vowel were calculated, providing each author with a so-called linguistic “fingerprint.”

## References

- Bannister, J., & Hardill, I. (2013). Knowledge mobilisation and the social sciences: Dancing with new partners in an age of austerity. *Contemporary Social Science*, 8, 167–175.
- Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. London and New York: Longman.

- Chaski, C. (2001). Empirical evaluations of language-based author identification. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 8, 1–65.
- Conley, J. M., & O’Barr, W. (1998). *Just words: Law, language and power*. London: University of Chicago Press.
- Cotterill, J. (2003). *Language and power in court: A linguistic analysis of the O. J. Simpson trial*. Basingstoke: Palgrave Macmillan.
- Cotterill, J. (2013). Corpus analysis in forensic linguistics. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. London: Wiley.
- Coulthard, M. (1994a). On the use of corpora in the analysis of forensic texts. *Forensic Linguistics*, 1, 27–43.
- Coulthard, M. (1994b). Powerful evidence for the defence: An exercise in forensic discourse analysis. In J. Gibbons (Ed.), *Language and the law* (pp. 414–427). London: Longman.
- Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25, 431–447.
- Coulthard, M. (2010). Experts and opinions: In my opinion. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 473–486). Abingdon, Oxford: Routledge.
- Coulthard, M., & Johnson, A. (2007). *An introduction to forensic linguistics: Language in evidence*. Abingdon, Oxon: Routledge.
- Coulthard, M., & Johnson, A. (2010). *The Routledge handbook of forensic linguistics*. Abingdon: Routledge.
- Coulthard, M., Johnson, A., & Wright, D. (2017). *An introduction to forensic linguistics: Language in evidence* (2nd ed.). Abingdon, Oxon: Routledge.
- De Costa, P. I. (Ed.). (2016). *Ethics in applied linguistics research: Language researcher narratives*. New York: Routledge.
- Finegan, E. (2010). Legal writing: Attitude and emphasis. Corpus linguistic approaches to “legal language”: Adverbial expression of attitude and emphasis in Supreme Court opinions. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 65–77). Abingdon, Oxon: Routledge.
- Gales, T. (2015). The stance of stalking: A corpus-based analysis of grammatical markers of stance in threatening communications. *Corpora*, 10, 171–200.
- Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *The International Journal of Speech, Language and the Law*, 14, 1–25.
- Grant, T. (2010). Text messaging forensics: Txt 4n6: Idiolect free authorship analysis? In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 508–522). Abingdon, Oxon: Routledge.
- Grant, T., & Baker, K. (2001). Identifying reliable, valid markers of authorship: A response to Chaski. *Forensic Linguistics: The International Journal of speech, Language and the Law*, 8, 66–79.
- Grant, T., & MacLeod, N. (2016). Assuming identities online: Experimental linguistics applied to the policing of online paedophile activity. *Applied Linguistics*, 37, 50–70.

- Hardaker, C. (2015). 'I refuse to respond to this obvious troll': An overview of responses to (perceived) trolling. *Corpora*, 10, 201–229.
- Hardcastle, R. A. (1997). CUSUM: A credible method for the determination of authorship? *Science and Justice*, 37, 129–138.
- Haworth, K. (2006). The dynamics of power and resistance in police interview discourse. *Discourse & Society*, 17, 739–759.
- Haworth, K. (2010). Police interviews in the judicial process: Police interviews as evidence. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 169–181). Abingdon, Oxon: Routledge.
- Heffer, C. (2005). *The language of jury trial: A corpus-aided analysis of legal-lay discourse*. Basingstoke: Palgrave Macmillan.
- Heydon, G. (2005). *The language of police interviewing: A critical analysis*. Basingstoke: Palgrave Macmillan.
- Johnson, A., & Wright, D. (2014). Identifying idiolect in forensic authorship attribution: An n-gram textbite approach. *Language and Law/Linguagem e Direito*, 1, 37–69.
- Kniffka, H. (2007). *Working in language and law: A German perspective*. Basingstoke: Palgrave Macmillan.
- Kredens, K., & Coulthard, M. (2012). Corpus linguistics in authorship identification. In P. Tiersma & L. Solan (Eds.), *The Oxford handbook of language and law* (pp. 504–516). Oxford: Oxford University Press.
- Larner, S. (2014a). A comparative review of *The Routledge handbook of forensic linguistics* and *The Oxford handbook of language and law*. *Language and Law/Linguagem e Direito*, 1, 194–197.
- Larner, S. (2014b). A preliminary investigation into the use of fixed formulaic sequences as a marker of authorship. *International Journal of Speech, Language and the Law*, 21, 1–22.
- Larner, S. (2015). From intellectual challenges to established corpus techniques: Introduction to the special issue on forensic linguistics. *Corpora*, 10, 131–143.
- MacLeod, N. (2016). "I thought I'd be safe there": Pre-empting blame in the talk of women reporting rape. *Journal of Pragmatics*, 96, 96–109.
- Repko, A., Szostak, R., & Buchberger, M. (2014). *Introduction to interdisciplinary studies*. London: Sage.
- Robson, C. (2011). *Real world research* (3rd ed.). West Sussex: Wiley.
- Rock, F. (2007). *Communicating rights: The language of arrest and detention*. Basingstoke: Palgrave Macmillan.
- Shuy, R. (2006). *Linguistics in the courtroom: A practical guide*. Oxford: Oxford University Press.
- Shuy, R. (2013). *The language of bribery cases*. Oxford and New York: Oxford University Press.
- Solan, L., & Tiersma, P. (2004). Author identification in American courts. *Applied Linguistics*, 25, 448–465.
- Solan, L., & Tiersma, P. (2005). *Speaking of crime: The language of criminal justice*. London: The University of Chicago Press.

- Tiersma, P. (2000). *Legal language*. London: University of Chicago Press.
- Tiersma, P. (2010). Instructions to Jurors: Redrafting California's jury instructions. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 251–264). Abingdon, Oxon: Routledge.
- Tiersma, P., & Solan, L. (Eds.). (2012). *The Oxford handbook of language and law*. Oxford: Oxford University Press.
- Tkačuková, T. (2015). A corpus-assisted study of the discourse marker *well* as an indicator of judges' institutional roles in court cases with litigants in person. *Corpora*, 10, 145–170.
- Tomblin, S. (2013). Coulthard, Malcolm. In C. Chapelle (Ed.), *The Encyclopedia of applied linguistics*. London: Wiley.
- Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, 17, 211–250.
- Woodhams, J., Grant, T., & Price, A. (2007). From marine ecology to crime analysis: Improving the detection of serial sexual offences using a taxonomic similarity measure. *Journal of Investigative Psychology and Offender Profiling*, 4, 17–27.
- Wright, D. (2013). Stylistic variation within genre conventions in the Enron email corpus: Developing a text-sensitive methodology for authorship research. *International Journal of Speech, Language and the Law*, 20, 45–75.



# 32

## World Englishes

Peter De Costa, Jeffrey Maloney, and Dustin Crowther

### Introduction

As a worldwide language, English has without doubt reached unprecedented heights of global importance (Seidlhofer, 2011; Trudgill, 2002). With conversations across multiple country boundaries now not only inevitable (Appiah, 2006), but, for many, a daily occurrence (MacKenzie, 2011), there has been a push to rethink and rearticulate current approaches toward World Englishes research (De Costa & Crowther, 2018; Kumaravadivelu, 2015). Yet, this proposed reconceptualization, one promoting an increased focus on phenomena related to globalization such as super-diversity and translanguaging (Bolton, 2018; Saraceni, 2015), has remained focused primarily on the conceptual side

---

P. De Costa (✉)

Department of Linguistics and Languages, Michigan State University,  
East Lansing, MI, USA  
e-mail: [pdecosta@msu.edu](mailto:pdecosta@msu.edu)

J. Maloney

Department of Languages and Literature, Northeastern State University,  
Tahlequah, OK, USA  
e-mail: [jeffmaloney87@gmail.com](mailto:jeffmaloney87@gmail.com)

D. Crowther

Department of English, Oklahoma State University,  
Stillwater, OK, USA  
e-mail: [dustincrowther@gmail.com](mailto:dustincrowther@gmail.com)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_32](https://doi.org/10.1057/978-1-137-59900-1_32)

(through the theorization of World Englishes, English as an international language, and English as a lingua franca), with little focus on the methodological approaches used to generate the data that inform and shape this reconceptualization. To address this methodological gap, this chapter, after providing a brief overview of three key streams of World Englishes research, describes several of the specific tools common to such inquiry.

## Streams of World Englishes Inquiry: WE, EIL, and ELF

Here we use World Englishes (big “W”) as an umbrella term for three specific streams of inquiry. The first of these is world Englishes (small “w”) inquiry stemming from a Kachruvian perspective (see below), differentiated from the umbrella term throughout our chapter using the acronym (WE). The two other streams of inquiry include English as an international language (EIL) and English as a lingua franca (ELF; Galloway & Rose, 2015). Though each stream is unique in overall focus, each derives from similar themes of inquiry. All three are “concerned with similar subject matter such as language attitudes and ideology, language policy and pedagogy, pragmatics, identity, culture and language” with a primary focus on “the linguistic and sociocultural dimensions of global uses and users of English” (Baker, 2015, p. 11).

WE “has been widely used to refer to localized forms of English found throughout the world, particularly in the Caribbean, parts of Africa, and many societies in Asia” (Bolton, 2012, p. 13). Much of the research in this area stems from the work of Larry Smith and Braj Kachru, with a heavy focus on Kachru’s (1991) concentric circle model. Within this model, English is seen as composing three distinct varieties of English: inner circle (first language usage, as in the UK, US), outer circle (second language usage, as in India, Singapore), and expanding circle (foreign language usage, as in Brazil, China). In light of ongoing globalization, the extent to which three distinct varieties truly exist has been questioned (e.g., Bruthiaux, 2003), and Bolton (2012, 2018) and Saraceni (2015) have both argued the need for an expanded scope of WE inquiry.

Such expansion allows for the consideration of EIL and ELF, both of which consider English when used as a medium within multilingual contact. While both have received substantial empirical attention, scholars have been less than precise in how distinct the two streams are (Galloway & Rose, 2015). Baker (2015) has argued that a primary difference is in the status each ascribes to specific varieties of English. EIL views multilingual contact as challenging due to the likelihood that “more than one variety of English is represented in such situations because each speaker brings a variety that he or she is most familiar with” (Matsuda, 2012, p. 7). From an ELF perspective, English usage is less



reliant on any specific language norm (Seidlhofer, 2011), with interaction more “likely to include borrowing, code-switching, and other types of crosslinguistic interaction” (MacKenzie, 2014, p. 4) as means to facilitate mutual intelligibility. In essence, EIL considers how the English variations of two interlocutors promote and hinder mutual intelligibility, while ELF places a greater focus on how interlocutors go about attaining this mutual intelligibility. Despite this difference in primary focus, the two are defined by similar ideologies (Galloway & Rose, 2015), putting at the forefront both native/non-native speaker and nonnative/nonnative speaker communication (Low, 2015).

Considering their shared focus, WE, EIL, and ELF resonate quite strongly with each other in terms of thinking and researching the global spread of English (Low, 2015). As such, it should not come as a surprise that there is significant overlap in the methodological approaches employed when conducting empirical research. Similarly, as each stream also maintains its own primary focus, there are methodological tools that are more conducive to a single stream. While providing a brief overview of many of the tools incorporated within World Englishes research, along with the research questions they address, the remaining portion of this chapter will focus on the methodologies most commonly used in WE, EIL, and ELF empirical inquiry.

## A Methodological Approach to World Englishes Inquiry

An increased interest in global usage of English has brought an increase in the number of similarly focused volumes (e.g., Galloway & Rose, 2015; Jenkins, 2015; Seargeant, 2012). While these volumes provide readers with an understanding of many of the themes that have defined such inquiry (e.g., identity, pronunciation, variation), little focus is placed on how these themes are investigated (though see De Costa, Crowther, & Maloney, *forthcoming*, for one such example). Considering the similarity in themes underlying the three streams of World Englishes (e.g., language attitudes and ideology, language policy and pedagogy, pragmatics, identity, culture, and language: Baker, 2015), it should not be surprising that World Englishes inquiry, whether targeting WE, EIL, or ELF, has relied primarily on corpora-, qualitative-, and discourse-based research methods as such tools are ideal for the study of natural social life (Saldaña, 2011). Approaches incorporated across streams have primarily included corpus analysis to uncover language norms (e.g., Matsumoto, 2010; Yao & Collins, 2012) and the use of ethnographic tools (e.g., Chew, 2013; Cogo, 2012), though data has also been collected through

the use of survey- and questionnaire-based techniques (e.g., Timmins, 2002; Wei & Su, 2015), as well as through acoustic analysis (e.g., Roeder, 2012). This chapter provides an overview of how such tools have been incorporated into World Englishes research, along with the themes they have been used to investigate. We first consider the methodologies adopted from a WE perspective before reviewing those used by EIL and ELF researchers.

## WE Methodologies

Beginning in the mid-1960s, WE has developed as a field concerned with investigating different varieties of English as legitimate systems. As Kachru (1976) argued, “The strength of the English language is in presenting the *Americanness* in its American variety, and the Englishness in its British variety. Let us, therefore, appreciate and encourage Third World varieties of English, too” (p. 236, emphasis in original). Following Kachru’s lead, WE has established itself as an influential field with numerous publications (e.g., Kirkpatrick, 2010; Saraceni, 2015; Seargeant, 2012) and, in 1981, the formation of its own journal, *World Englishes*, all established with the intent of continuously moving the field forward.

WE research has covered a wide range of areas. It has a rich history of examining how English is utilized within and across different cultures, though, up till now, a majority of publications have focused on the *what* of WE (what a variety is) and not so much on the *how* (how inquiry is accomplished). Since the examination of WE is broad and varied, it is easy to see why this might be the case. Bolton (2005), in his brief review of the then-current approaches to WE, outlined three major groupings: studies with a primary *linguistic* orientation, those with both *linguistic* and *sociopolitical* orientations, and those primarily focused on *sociopolitical* and *political* issues. Seargeant (2012) offers a fourth grouping, that of approaching WE within local and translocal *cultural* contexts. As Bolton (2005) points out, these different groupings help to frame the range of research that has been conducted, with each potentially serving to inform the others.

What follows is an outline of the major methodologies adopted by researchers in WE and a brief explanation of their contributions to the field. Each section presents one or two studies that illustrate how the method was utilized and can serve to inform future research. While we recognize the contribution of many other studies to WE literature, the studies discussed here can serve as an introduction to how the specific method has been employed.

## Corpus Linguistics

One method that has served to inform each of the four strands previously identified is corpus linguistics. WE literature is replete with analyses of established corpora (e.g., ICE<sup>1</sup>) or corpora constructed for research purposes. They include both spoken and written corpora that have been gathered from a range of sources, including traditional mass media outlets (e.g., radio, television, newspaper) and online platforms (e.g., social media, forum postings). Interviews, focus groups, and other interactional events have also been employed, with data being recorded and/or transcribed for reference. Corpora have been widely used to address purely linguistic interests, such as pragmatic markers and phonological variability, as well as sociolinguistic concerns, including language change/evolution and cultural attitudes towards global English variation. A clear example of how corpora have been used to consider *linguistic* interests across varieties of English can be seen in Yao and Collins (2012), who compared five *inner* circle Englishes to five *outer* circle varieties to better understand regional variation of a specific English language structure, the present perfect. Their approach allows them to establish a clearer picture of the impact of colonization on the evolution of different English varieties (see Study Box 32.1 for more detail).

### Sample Study 32.1

Yao, X., & Collins, P. (2012). The present perfect in world Englishes. *World Englishes*, 31, 386–403.

#### Research Background

It is noted that the English present perfect structure is one of the most widely researched in linguistic literature. Yao and Collins observe that while other Indo-European languages are adopting the present perfect as a default for referring to the past and the preterit is being used in more formal registers, the trend is the opposite for English. Work focused on this phenomenon at one point in time has yielded inconsistent findings on regional variations, which Yao and Collins point out as possibly due to methodological issues.

#### Research Problem/Gaps

This study is focused on examining the present perfect in English, which has been relatively restricted to examinations of its variation within only inner circle countries. Norms of usage outside of inner circle countries have been examined, but only in reference to their deviations from native speakers' patterns. This fact may be a possible explanation for the inconsistent findings of regional variation.

#### Research Method

- Corpus study of WE.
- Drew on corpora to compare the stylistic variations in usage of the present perfect of five inner circle (British, American, Australian, Canadian, New Zealand) and five outer circle (Philippine, Hong Kong, Singapore, Kenyan, Indian) English varieties.

- The corpora utilized were drawn from those available through the International Corpus of English (ICE). Two hundred texts were selected under multiple categories and genres from each ICE corpus for comparison.
- Raw frequency across both inner circle and outer circle varieties were compared as well as ratios of usage of contracted auxiliaries, and present perfect and the preterit. The ratios between the present perfect and the preterit were also compared across genres, their co-occurrence with “already,” “yet,” “always,” “never,” and “ever.”

### Key Results

Yao and Collins established a continuum on which to place different WE varieties with regard to the usage of the English present perfect. Findings indicated that for “inner” circle varieties, earlier decolonization aligned with greater decline in the usage of present perfect. For “outer” circle varieties, the three Asian varieties (Filipino, Hong Kong, Singaporean) used the present perfect far less than the Indian and Kenyan varieties, with Indian English using it more than British English. The authors also discuss the role of proximity, with similar usage existing between Canadian and American, Australian and New Zealand, and Filipino, Hong Kong, and Singaporean varieties.

### Comments

Yao and Collins’s usage of multiple corpora from different English varieties served to paint a clearer picture in understanding regional variations in usage of a specific structure. Major findings carry important implications on the impact that geographical location, immigration patterns, acceptance of formal speech, and colonization can have on stylistic patterns.

Representing a *sociolinguistic* perspective, Ogoanah (2011) examined the usage of the particle *as in* in Nigerian English (NE) and how it served as a pragmatic marker distinct to that variety of English. Analyzing a constructed corpus of recorded interactions between and interviews with NE speakers, Ogoanah posited that the *as in* particle was employed to convey emphasis or to stress the importance of an upcoming statement (e.g., “He didn’t lecture me, *as in* we were *learning* from each other,” p. 209). Taking a *political/sociopolitical* interest, Alvaro (2013) investigated the usage of phrases such as “harmonious society” within a specialized corpus constructed from a state-run Chinese newspaper. By analyzing contexts in which the collocated phrases appeared, Alvaro highlighted the specialized usage of this phrase within Chinese Communist Party media outlets. This analysis, which the author linked to historical usage of English within the Chinese media, provided a picture of how the media help to maintain a specific kind of environment. Within a more *culturally* oriented domain, analyses of mass media outlets have allowed for an understanding of the ideologies and norms present within different English varieties. One area that has received attention is radio broad-

casts within the different nations of Africa. For example, Makalela (2013) constructed a corpus of Black South African English (BSAE) radio broadcasts to examine speech tokens specific to this variety. For six weeks, radio broadcasts (e.g., talk shows, interviews, call-in programs) were recorded and transcribed. This resulted in a corpus of approximately 209,000 words, which were analyzed utilizing a keyword list of BSAE features. Makalela identified four specific features dominant across radio programs, indicating that BSAE was becoming more stabilized within Black South African society.

Taken together the different studies listed here show just a small fraction of the ways in which corpora can be utilized to inform WE research across the various groupings of interest. This tool's importance in WE inquiry can be seen in the sheer number of studies that incorporate such an approach. To fully appreciate Third World varieties of English, as argued by Kachru (1976), we need to understand them.

## Case Studies and Ethnographic Inquiry

Further, WE research has made use of methodological approaches common to sociolinguistic inquiry, specifically the use of case studies, with both participant observation and interviews serving as vital data collection tools. Such an approach has typically been employed to provide a richer understanding of the sociolinguistic realities and cultural environments at play in which different WE varieties are spoken. It has also served to highlight the power of English and its impact on daily living for individuals and communities (e.g., Rudwick, 2008). Chew (2013), for example, utilized an ethnographic approach to investigate how Singlish (Singaporean English) is often used in Islamic classrooms in Singapore to promote a warmer and more comfortable environment for students. She followed six Islamic classes through the use of interviews, at-home and at-school observations, and video-recorded interactions. Chew concluded that Singlish usage within these classes served to resist the imperialism of standard English, promoting camaraderie and fellowship.

An area of inquiry that has also received attention is the growing number of international call centers that draw upon the population of outer circle countries to serve inner circle customers. Forey (2013) drew on a series of interviews with call center employees to examine the social impact that employment within this industry had on Indian women. While there were significant negative drawbacks to call center employment for women (e.g., additional work beyond household expectations, decreased social activity due to night shifts), there were economic, social, and professional benefits not

previously available. Key from this study is that the globalization of English has moved beyond the development of multiple varieties, showing the ability to reach the core of society.

Whether committing to a fully ethnographic study (as in Chew, 2013) or making use of tools common to such inquiry (such as the use of interviews in Forey, 2013), there is much that can be revealed when considering the socio-linguistic, sociopolitical, and cultural undercurrents of WE development. Such tools are key to understanding the core of how these three interests of WE interact with the language itself.

## Survey Usage

Surveys have been utilized by WE researchers to gain insights into daily practice, language attitudes, and the environment of language usage. Some researchers have taken information from large-scale, government-administered surveys (e.g., Wei & Su, 2015), while others have created and distributed their own. Surveys have helped to inform not only researchers but also instructors about the variety that exists across English users worldwide. In Hashim and Leitner's (2011) pilot study, the effect of language contact was investigated using a survey given to Malay English speakers, with a focus on various loan words. The survey correlated learners' ethnicity and age group, with their familiarity with specific lexical items from the public domain. The results revealed that certain loan words were more well known to some ethnic or religious groups than to others. Specifically, Chinese words were more familiar to Malays of Chinese descent. Surveys can also serve as a tool to investigate how different English varieties are viewed across communities. For example, Bernaisch (2012) investigated the attitudes that 169 Sri Lankan English speakers had towards Sri Lankan, Indian, British, and American English. Findings suggested that while all four varieties were viewed positively, respondents showed a preference for the British variety, followed by their own.

## Acoustic Analyses and Matched Guise

Two additional tools serve as informative approaches toward understanding both linguistic differences between varieties of English, and how these differences may relate to user attitudes. While acoustic analysis underscores how one variety differs in sound from another, the matched-guise technique allows an understanding of how these acoustic differences may impact perception.

An example of acoustic analysis is Maxwell and Fletcher (2010), a study that examined the production of Indian English (IE) diphthongs by L1 speakers of Hindi and Punjabi. The authors' goal was to evaluate the differences within IE that may stem from L1 influence, thus providing a better understanding of some of the variety in the phonological features of IE users. Based on the elicited production of ten words, their analysis revealed that L1 Punjabi speakers tended to demonstrate a wider range of phonetic realizations and were less likely to produce the target diphthongs. Beyond this more contrastive analysis approach, researchers have also utilized auditory analysis to look at how accents can inform understandings of attitudes or ideologies about language or language norms, often drawing from spoken corpora. Acoustic analysis has also been utilized as a tool to examine language shift (O'Sullivan, 2013; Roeder, 2012).

As a means to address the relation between production variety across different Englishes and language attitude, many studies have employed some form of a *matched-guise* design, in which the same speaker adopts different dialects (or even different languages) and participants respond by rating perceptions of the speech across various dimensions, such as friendliness or intelligence (see Lambert, Hodgson, Gardner, & Fillenbaum, 1960). Such studies provide researchers with clearer insight into cultural and sociolinguistic practices of speakers of different varieties. Cavallaro, Ng, and Seilhamer (2014) utilized the matched-guise technique, in addition to interviews, to investigate Singaporean English speakers' attitudes toward Singapore Colloquial English (SCE). Two-hundred fifty-nine participants from a Singapore university listened to six guises, three by a male speaker and three by a female one and rated the speakers across ten traits using a Likert scale. The guises consisted of Standardized Singapore English, SCE, and Basilectal SCE. Results showed that while most Singaporean speakers indicated a somewhat negative view of SCE, the interviews revealed that SCE is viewed more positively to express solidarity, especially within private domains.

## The Methodologies of EIL and ELF

As mentioned above, the distinction between EIL and ELF has at times lacked clarity among scholars (Galloway & Rose, 2015). This is no more evident than in Jenkins' highly cited 2000 book *The Phonology of English as an International Language*. Though at the time adopting the term EIL, Jenkins has since strongly advocated her work as representative of an ELF perspective



(Low, 2015). This alternation between EIL and ELF has led to Jenkins' work being highly cited as representative of both. Of particular interest, though, is that the methodology employed and the research questions addressed are characteristic of both EIL and ELF inquiry. As both streams are defined by similar ideologies (Galloway & Rose, 2015), with a focus on native/nonnative and nonnative/nonnative speaker communication (Low, 2015), it is not surprising that both incorporate similar methodological tools of inquiry. While primarily corpus-based, EIL/ELF inquiry has additionally made use of ethnographic tools such as observations and interviews, while also using various survey- and questionnaire-based approaches to generate data. Following a trend to incorporate a primarily qualitative approach, the above tools have often worked in harmony, allowing for a richer, more in-depth process of data collection (Friedman, 2012).

## Corpus-Based Inquiry

EIL/ELF inquiry has made substantial usage of corpora to address a variety of themes, including, but not limited to, Jenkins' (2000) focus on pronunciation features that hinder mutual intelligibility, as well as studies targeting lexicogrammar (e.g., Seidlhofer, 2004), pragmatics (e.g., Cogo & Dewey, 2012), and negotiation strategies (e.g., Matsumoto, 2010). Key to the EIL/ELF agenda is a focus on how such usage impacts interaction, rather than the more descriptive approach described for WE previously. Importantly, the findings of such inquiry have served as the background for many pedagogical volumes on EIL (e.g., Matsuda, 2012, 2017) and ELF (e.g., Bayyurt & Akcan, 2015).

Corpus-based inquiry has tended to follow one of two approaches. First, scholars have taken it upon themselves to generate their own corpus data, an approach exemplified in Matsumoto (2010), who recorded paired interactions between ELF speakers in a university dormitory dining hall (see Sample Study 32.1). Within such an approach, data may be collected both naturally (as done in Jenkins, 2000) and through semi-experimental means (as in Matsumoto, 2010). A second approach involves analyzing an already established corpus, such as the English as a Lingua Franca in Academic Settings (ELFA) corpus (ELFA, 2008). Recent topics of inquiry using the ELFA corpus have included idiom usage during international student-professor interactions (Franceschi, 2013) and the introduction of and negotiation of specialized terms in academic settings (Gotti, 2014).



### Sample Study 32.2

Matsumoto, Y. (2010). Successful ELF communications and implications for ELT: Sequential analysis of ELF pronunciation negotiation strategies. *Modern Language Journal*, 95, 97–114.

#### Research Background

On the basis that 80% of English speakers globally are nonnative, such speakers are more likely to encounter a fellow nonnative speaker than native. While a “great wall” (p. 97) may be perceived to exist between native and nonnative speakers, such a wall is likely to be diminished when two nonnative (or ELF) users interact. In such interactions, ELF users are likely to adopt various accommodation strategies to negotiate mutual intelligibility with their interlocutor. For the purposes of her study, Matsumoto focuses on pronunciation, as this area had received limited attention in existing ELF scholarship.

#### Research Problem/Gap

Matsumoto investigated how ELF users competently used English within a graduate student dormitory, targeting repair sequences used that allowed for the types of negotiations needed to ensure mutual understanding and pronunciation intelligibility. She poses the research question “How do ELF speakers succeed in negotiation meaning when their pronunciation causes a breakdown in understanding?” (p. 101)

#### Research Method

- Qualitative study of ELF.
- Six ELF-speaking graduate students at a US university (various linguistic backgrounds) living in the same dormitory.
- Data consisted of pair interactions in the dormitory dining hall and follow-up, semi-structured interviews. Data collection occurred over three months, with each interaction ranging from 25 to 29 minutes and interviews between 4 and 9 minutes. Interactions were video recorded, capturing both verbal and nonverbal behavior.
- Matsumoto examined all the interaction recordings, identifying and transcribing specific sequences that involved negotiations of meaning affected by pronunciation. As a means of data triangulation, Matsumoto considered information gathered from interviews with each speaker.

#### Key Results

- Matsumoto reported on three dyads. Yuka and Pham provided evidence of successful repair initiation, with Pham willing and able to accommodate her pronunciation in such a way as to make her utterances more intelligible to Yuka.
- Agus and Shuji demonstrated how negotiation practices are an acquirable skill, as they required more time to negotiate meaning during pronunciation breakdowns. Additionally, Shuji was more likely to provide contextual clues than to adjust his pronunciation.
- Hijung and Fang chose not to pursue repair, from which Matsumoto surmised that for some ELF speakers flow of conversation may be preferable to initiating repair.

### Comments

Considering the specified domain of contact Matsumoto was interested in, access to corpora of such interactions was likely nonexistent, necessitating a need to generate her own. Key is that her interactive data occurred in an everyday context, with Matsumoto exercising little influence over what was said and in what way. By focusing on ELF usage in a naturalistic setting, Matsumoto was able to show how different ELF users employ different linguistic strategies to overcome communicative breakdowns, or, in line with an overarching EIL/ELF objective, to attain mutual intelligibility.

The use of corpora in EIL/ELF inquiry has proven advantageous as it provides scholars with readily available data which can be analyzed through different approaches (e.g., discourse analysis, conversation analysis) on a number of different levels (e.g., interactional, phonemic).

## Ethnographic Tools

While ethnography in the pure sense (i.e., long-term observation and documentation of a target community's social life) is more characteristic of WE than EIL/ELF inquiry, several of the tools that define it play a significant role in addressing topics of EIL/ELF interest. Commonly used as a pair, observation and interviews allow insight into topics that are beyond the reach of a corpus-based approach. For example, Cogo (2012) examined the link between ELF usage in a business setting and super-diversity (i.e., contact between "a heightened diversity of communities, cultural backgrounds, and sociolinguistic resources," p. 288), via a case study in which she interviewed two users, individually and as a pair. This was after having spent time observing the office and gathering as much information as possible through a number of different sources (e.g., freelance staff, client e-mails, office pictures). The data revealed many characteristics of this work environment, including how while Spanish served as the medium for most internal communication, English was used extensively for communicating externally. Other themes of note included how staff put domain knowledge ahead of native speaker-like speech and a tension between the English needed for specific IT activities (domain relevant) and that which is used in general business activities (speak English well). Highlighting the importance of generating as detailed an analysis as possible, additional data were collected and analyzed from audio-recordings of spoken synchronous online interactions between the directors and clients (ultimately developing into a corpus).

While Cogo (2012) carried out a workplace-based study using observation and interviews, much attention has also been paid to academic settings.

Baker (2009) investigated intercultural communication in a Thai university setting, observing a number of L2 English interactions and conducting follow-up interviews. He gained access through his position as a visiting lecturer at the school. While the interactions allowed him to see how intercultural communication developed, the interviews provided an opportunity to delve into each interlocutor's experiences with foreign language usage, attitudes towards other languages and cultures, and how they viewed the relationship between language learning and use. A similar approach was taken in De Costa (2012), over a longer period of time (one year). Focusing on an Indonesian immigrant student studying in a high school in Singapore, through the use of in- and out-of-school observations and interviews conducted with the student and her teachers, De Costa explored the usage of ELF, arguing the need to understand language acquisition from a more dynamic perspective in a globalized world, specifically targeting the sociolinguistic realities in which many contemporary L2 English learners exist, previously overlooked by mainstream second language acquisition research (Firth & Wagner, 1997, 2007).

Though corpus-based approaches allow for an in-depth analysis of the language used within a specific environment, they do not allow for the depth of investigation ethnographic tools such as observations and interviews provide. This is not to value one over the other, as they can not only work in tandem (as in Cogo, 2012), but can also address distinct research questions. Ethnographic tools are less likely to address the pronunciation interests of Jenkins (2000), while corpora do not allow for an understanding of how ELF users view their place in society (as in Baker, 2009).

## Surveys and Questionnaires

Two final tools to consider, effective for data collection across a range of topics, are surveys and questionnaires. While such tools may not provide the depth of linguistic analysis that corpus-based approaches afford, or allow for the same level of ethnographic detail, they make for a quick, focused investigation, and potentially allow for a large sample size. A key example of how efficiently this method can produce data can be seen in Timmins' (2002) study in which he used a pair of questionnaires (delivery method not provided) to collect both student and teacher views on the use of native speaker norms in English education. Timmins collected 400 student responses from 14 countries and 180 teacher responses from 45 countries. The questionnaires used both scaled responses and open-ended items, and Timmins, again showing overlap between methodological tools, supported his questionnaire results

by conducting 15 interviews with students. On the basis of questionnaire results, Timmins argued that L2 English learners had a greater desire for native norms than did teachers, though this difference was stronger for pronunciation than for grammar.

Not all survey/questionnaire-based studies need to have a sample the size of Timmins (2002;  $N = 580$ ) to be effective. Much may depend on both the target population, and practicality considerations as well. For example, Galloway (2013) used a pre/post questionnaire to investigate the impact that a content-based class on Global Englishes would have on the attitudes of L1 Japanese speakers of English. An open and closed questionnaire was administered to 52 participants, half of whom attended the Global Englishes course (the other half took a course on tourism). Though large-scale changes were not found across the two questionnaires (learners still showed a preference for native English varieties), the Global Englishes group began to show a greater awareness of Global Englishes usage and an increase in their confidence as nonnative speakers. Moving beyond the classroom, Ingvarsdóttir and Arnbjörnsdóttir (2013) collected 238 responses to a survey investigating how Icelandic academics view writing academic papers in English. The survey featured all closed responses, though interviews were conducted to develop more well-rounded data. The responses revealed that while the academics felt relatively confident in their English knowledge, they still required assistance when composing academic papers, a concern considering the publication pressures they faced.

The three studies above demonstrate how the use of surveys and questionnaires can allow for relatively large collections of data. However, the researchers in all three studies recognized the limitations of such an approach, as they all chose to include subsequent interviews with a subset of respondents as a means to gather a more in-depth personal view of the larger data set. One further note is that surveys and questionnaires are optimal approaches to incorporate a quantitative view on EIL/ELF inquiry. This is the primary approach taken in all three of the above studies, descriptive for Timmins (2002) and inferential for Galloway (2013) and Ingvarsdóttir and Arnbjörnsdóttir (2013).

## Future Directions

To date, there has been little work done to outline the major methods that have been employed within World Englishes research (see De Costa et al., [forthcoming](#)), and this chapter serves as an introduction to some of the major

methodologies and practices used within WE, EIL, and ELF. In the future, it is expected that the methods that have been discussed in this chapter will change and evolve as do the means by which humans interact. For example, with the rise of new technologies and social media, an important area of inquiry that is gaining traction within World Englishes research is the effects of language contact on different dialects of English. Moving forward, the number of studies drawing from online platforms and social media postings may soon become a major source of understanding global English usage.

As discussed, while the three key streams of World Englishes overlap in the methodologies researchers employ, they differ in the research questions they address. WE inquiry focuses on understanding, cataloguing, and legitimizing different varieties of English adopted and utilized by different cultures and subcultures, as well as examining the role that English plays in social realms. EIL and ELF research has been more concerned with the phenomena present within multilingual contact. As such, while all three streams place a heavy emphasis on corpus- and ethnographic-based inquiry, along with survey- and questionnaire-based approaches, the history and research interests of WE have led to the incorporation of a wider range of tools, such as the use of matched-guise tasks. As EIL and ELF research continues to grow, and new themes of interest arise, there is little doubt that their tools will similarly expand.

As World Englishes research continues to develop, it is important that researchers adopt important and clearly defined approaches to inquiry. While the methods discussed here share much in common, their effectiveness, as is always the case with empirical research, comes down to how appropriately they address the specific questions being asked. And it is for this primary reason of illustrating the breadth of research inquiry on how English is used today that we have elected to distinguish the three streams on WE, EIL, and ELF in this chapter.

## Resources for Further Reading

Galloway, N., & Rose, H. (2015). *Introducing global Englishes*. New York: Routledge.

Designed as a university-level textbook, it provides an accessible introduction to the historical development and global spread of the English language (i.e., WE), along with current conceptualizations (i.e., ELF) and trends for the future. While considering WE and ELF as separate constructs, the authors

employ *global Englishes* as an umbrella term based on similarities between research paradigms. Specifically, they put forth that both share a pluricentric notion of English variation, a focus on usage by nonnative English speakers, and a focus on ownership independent of native English norms and have implications for pedagogy (p. xii). Moving from a description of the historical development and global spread of English (Chaps. 1 and 2), Galloway and Rose consider the advantages and disadvantages of English as a global language (Chap. 3) and variation across Kachru's three concentric circles (Chaps. 4–6), before moving to a discussion of ELF and global attitudes towards such usage (Chaps. 7 and 8). They conclude by highlighting both pedagogical considerations (Chap. 9) and potential future trends (Chap. 10). Full of tables and figures to aid in understanding, each chapter is accompanied by discussion questions, debate topics, presentation and writing prompts, and suggested readings. This is in addition to their companion website (<http://www.routledgetextbooks.com/textbooks/9780415835329/>).

Matsuda, A. (Ed.). (2017). *Preparing teachers to teach English as an international language*. Bristol, UK: Multilingual Matters.

Employing EIL as an overarching term encompassing WE and ELF, Matsuda's volume argues the need to reconceptualize English language teaching (ELT) to meet the diverse uses and users of English globally. She here views EIL as existing within multilingual contact between speakers who possess their own variety of English they are most familiar with. Primarily targeting ELT teacher educators, the depth of knowledge presented is additionally beneficial for both ELT teachers (in both second and foreign language contexts) and World Englishes researchers with an interest in pedagogy. Her volume includes chapters from well-established scholars (e.g., Bayyurt, Sifakis, Galloway, Hino, Rose) who draw upon theories and research in WE, ELF, and EIL "to illustrate diverse approaches to prepare teachers who can meet the diverse needs of English learners in international contexts today" (p. xvii). Split into six sections, this volume considers theoretical frameworks of EIL (Chaps. 1 and 2), provides overviews of both existing teacher preparation programs (Chaps. 3 and 4, 12–14) and courses dedicated to the teaching of EIL (Chaps. 5–8), and how a focus on EIL can inform other ELT topics (Chaps. 9–11). The volume concludes with a unit focused on example lessons, activities, and tasks to be utilized in EIL teacher preparation (Chap. 15).

Saraceni, M. (2015). *World Englishes: A critical analysis*. London: Bloomsbury Publishing.

In the beginning of the volume, Saraceni states that the field of WE is facing what he calls a “crisis.” He states that the book has a twofold objective, which is to present traditional views held in the WE field as well as highlighting limitations to WE research thus far. The book is divided into four parts: history, language, ideology, and pedagogy.

The first two chapters of the book outline the history of the English language along with the rise of “New Englishes,” until the subsequent term world Englishes (Chaps. 2 and 3), with the use of writings of previous and current scholars on the subject. Saraceni then moves to examining the current WE paradigm and presenting what he views as important limitations to previous work in WE and poses important questions that must be grappled with in the future (Chaps. 4 and 5). In the chapters on ideology, Saraceni outlines concepts of linguistic imperialism and resistance to it, although he observes that the WE response of attempting to recognize English varieties is a complex process that requires more work (Chap. 6). In the final section (Chap. 7), Saraceni points to the inadequacy of the traditional WE paradigms to address notions of hybridity and plurality (Chap. 7).

While there is not a specific focus on research methods within the book, the points raised and questions posed by Mario Saraceni can serve as a springboard into research that addresses them.

Seargeant, P. (2012). *Exploring world Englishes: Language in a global context*. New York: Routledge.

Seargeant explains that this book is focused on presenting the “practical issues” of WEs along with connecting them with major theories. It starts by discussing multiple issues within the paradigm, and provides a historical overview of English in the globalized context. Multiple examples are pulled from research, but there is limited discussion on the research methods. It also includes a chapter on policies and interventions that practitioners have attempted to put into place. Discussion of the kinds of research that have been carried out primarily informs discussions of findings and their impact. There is a discussion on corpus work and how it impacted the early views on the paradigm of WE.

In the second part of the book (e.g., Chap. 12), Seargeant includes a discussion on the major theoretical frameworks that inform research within WEs, but argues that WEs do not have “a fully methodologically based coherence” (p. 161), but that the common theories provide a unity of purpose for researchers within this area. He states:



Over the last two or three decades as the discipline has developed, researchers have drawn on a variety of different theoretical models, but threading throughout them is this focus on the way that variation in English is ultimately tied to political and cultural identity, and the conviction that studying the subject is able to expose the nature and significance of this relationship. (p. 161)

## Note

1. ICE stands for the International Corpus of English; more information can be found at <http://www.ucl.ac.uk/english-usage/projects/ice.htm>

## References

- Alvaro, J. J. (2013). Political discourse in China's English language press. *World Englishes*, 32, 147–168.
- Appiah, K. A. (2006). *Cosmopolitanism: Ethics in a world of strangers (issues of our time)*. New York: W.W. Norton & Company, Inc.
- Baker, W. (2009). The cultures of English as a lingua franca. *TESOL Quarterly*, 43, 567–592.
- Baker, W. (2015). *Culture and identity through English as a lingua franca*. Berlin, Germany: De Gruyter Mouton.
- Bayyurt, Y., & Akcan, S. (Eds.). (2015). *Current perspectives on pedagogy for English as a lingua franca*. Berlin: De Gruyter Mouton.
- Bernaisch, T. (2012). Attitudes towards Englishes in Sri Lanka. *World Englishes*, 31(3), 279–291.
- Bolton, K. (2005). Where WE stands: Approaches, issues, and debate in world Englishes. *World Englishes*, 24, 69–83.
- Bolton, K. (2012). World Englishes and Asian Englishes: A survey of the field. In A. Kirkpatrick & R. Sussex (Eds.), *English as an international language in Asia: Implications for language education* (pp. 13–26). New York: Springer.
- Bolton, K. (2018). World Englishes and second language acquisition. *World Englishes*, 37(1), 5–18.
- Bruthiaux, P. (2003). Squaring the circles: Issues in modeling English worldwide. *International Review of Applied Linguistics*, 13, 159–178.
- Cavallaro, F., Ng, B. C., & Seilhamer, M. F. (2014). Singapore colloquial English: Issues of prestige and identity. *World Englishes*, 33, 378–397.
- Chew, P. G.-L. (2013). The use of Singlish in the teaching of Islam. *World Englishes*, 32, 380–394.
- Cogo, A. (2012). ELF and super-diversity: A case study of ELF multilingual practices from a business context. *Journal of English as a Lingua Franca*, 1, 287–313.



- Cogo, A., & Dewey, M. (2012). *Analysing English as a lingua franca*. London: Continuum.
- De Costa, P. I. (2012). Constructing SLA differently: The value of ELF and language ideology in an ASEAN case study. *International Journal of Applied Linguistics*, 22, 205–224.
- De Costa, P. I., & Crowther, D. (2018). SLA and World Englishes: Establishing a dialogue and common reserve. *World Englishes*, 37.
- De Costa, P. I., Crowther, D., & Maloney, J. (Eds.). (Forthcoming). *World Englishes inquiry: Research methodology and topical understanding*. New York: Routledge.
- ELFA. (2008). *The corpus of ELF in academic settings*. Director: Anna Mauranen. Retrieved May 24, 2016, from <http://www.helsinki.fi/elfa/elfacorpus>
- Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *Modern Language Journal*, 81, 285–300.
- Firth, A., & Wagner, J. (2007). Second/foreign learning as a social accomplishment: Elaborations on a reconceptualized SLA. *Modern Language Journal*, 91(S1), 800–819.
- Forey, G. (2013). The impact of call centre employment on women in India. *World Englishes*, 32, 503–520.
- Franceschi, V. (2013). Figurative language and ELF: Idiomaticity in cross-cultural interaction in university settings. *Journal of English as a Lingua Franca*, 2, 75–99.
- Friedman, D. A. (2012). How to collect and analyze qualitative data. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 180–200). Oxford: Blackwell.
- Galloway, N. (2013). Global Englishes and English language teaching (ELT) – Bridging the gap between theory and practice in a Japanese context. *System*, 41, 786–803.
- Galloway, N., & Rose, H. (2015). *Introducing global Englishes*. New York: Routledge.
- Gotti, M. (2014). Explanatory strategies in university courses taught in ELF. *Journal of English as a Lingua Franca*, 3, 337–361.
- Hashim, A., & Leitner, G. (2011). Contact expressions in contemporary Malaysian English. *World Englishes*, 30, 551–568.
- Ingvarsdóttir, H., & Arnbjörnsdóttir, B. (2013). ELF and academic writing: A perspective from the expanding circle. *Journal of English as a Lingua Franca*, 2, 123–145.
- Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.
- Jenkins, J. (2015). *Global Englishes* (3rd ed.). New York: Routledge.
- Kachru, B. B. (1976). Models of English for the third world: White man's linguistic burden or language pragmatics? *TESOL Quarterly*, 10, 221–239.
- Kachru, B. B. (1991). Models of English for the third world: White man's linguistic burden or language pragmatics. In A. Brown (Ed.), *Teaching English pronunciation: A book of readings* (pp. 31–52). New York: Routledge.

- Kirkpatrick, A. (2010). *The Routledge handbook of World Englishes*. New York: Routledge.
- Kumaravadivelu, B. (2015, October 9). *Building bridges, enabling crossings: Challenges facing World Englishes in a global society*. Plenary speech presented at the 21st Conference of the International Association of WE, Istanbul, Turkey.
- Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology*, 60, 44–51.
- Low, E.-L. (2015). *Pronunciation for English as an international language: From research to practice*. New York: Routledge.
- MacKenzie, I. (2011). *Intercultural negotiations*. New York: Routledge.
- MacKenzie, I. (2014). *English as a lingua franca: Theorizing and teaching English*. New York: Routledge.
- Makalela, L. (2013). Black South African English on the radio. *World Englishes*, 32, 93–107.
- Matsuda, A. (Ed.). (2012). *Principles and practices of teaching English as an international language*. Bristol, UK: Multilingual Matters.
- Matsuda, A. (Ed.). (2017). *Preparing teachers to teach English as an international language*. Bristol, UK: Multilingual Matters.
- Matsumoto, Y. (2010). Successful ELF communications and implications for ELT: Sequential analysis of ELF pronunciation negotiation strategies. *Modern Language Journal*, 95, 97–114.
- Maxwell, O., & Fletcher, J. (2010). The acoustic characteristics of diphthongs in Indian English. *World Englishes*, 29, 27–44.
- O’Sullivan, J. (2013). Advanced Dublin English in Irish radio advertising. *World Englishes*, 32, 358–376.
- Ogoanah, F. N. (2011). The pragmatic roles of *as* in Nigerian English usage. *World Englishes*, 30, 200–210.
- Roeder, R. (2012). The Canadian shift in two Ontario cities. *World Englishes*, 31, 478–492.
- Rudwick, S. (2008). “Coconuts” and “oreos”: English-speaking Zulu people in a South African township. *World Englishes*, 27, 101–116.
- Saldaña, J. (2011). *Fundamentals of qualitative research*. Oxford, UK: Oxford University Press.
- Saraceni, M. (2015). *World Englishes: A critical analysis*. London: Bloomsbury.
- Sergeant, P. (2012). *Exploring World Englishes: Language in a global context*. New York: Routledge.
- Seidlhofer, B. (2004). Research perspectives on teaching ELF. *Annual Review of Applied Linguistics*, 24, 209–239.
- Seidlhofer, B. (2011). *Understanding ELF*. Oxford: Oxford University Press.
- Timmins, I. (2002). Native-speaker norms and international English: A classroom view. *ELT Journal*, 56, 240–249.

- Trudgill, P. (2002). *Sociolinguistic variation and change*. Edinburgh: Edinburgh University Press.
- Wei, R., & Su, J. (2015). Surveying the English language across China. *World Englishes*, 34, 175–189.
- Yao, X., & Collins, P. (2012). The present perfect in world Englishes. *World Englishes*, 31, 386–403.



# 33

## Heritage, Community, and Indigenous Languages

Shereen Bhalla and Terrence G. Wiley

### Introduction

This chapter provides a discussion on research priorities and methodologies as they relate to research activities that fall broadly under the labels *heritage*, *community*, and *indigenous* languages. In attempting to cover such a broad topic, there are several necessary cautions and delimitations. First, research that falls under these labels is not characterized by any single method or grouping of methodology. It often focuses broadly on characteristics of languages or their speakers, learners, families, or communities. In order to delimit the topic, this chapter will focus more on research orientations and methodologies oriented to learners and users rather than on linguistic characteristics of languages identified under these labels. As will be noted, the *heritage* label, as well as other labels, which could also be used, is sometimes contested (Wiley, 2014). While noting disagreement in the use of labels, there has been a growing body of research under the heritage label as well as related work which falls under community, indigenous, or Native American language—or more traditionally—work that falls under the domains of language maintenance or shift.

In approaching research questions related to heritage speakers and learners, one inevitably encounters issues involving identity, language status, and power

---

S. Bhalla (✉) • T. G. Wiley

Center for Applied Linguistics, Washington, DC, USA

e-mail: [Shereenbhalla@gmail.com](mailto:Shereenbhalla@gmail.com); [twiley@asu.edu](mailto:twiley@asu.edu)

differentials. In the United States, the term *heritage language* has been largely used to refer to immigrant languages, and although it is sometimes used to include indigenous languages, its application to these languages and their speakers remains controversial. Outside the United States, the term *community languages* is often preferred (Wiley, 2014). As Baker and Jones (1998) noted, both speakers of Navajo and Spanish-speaking Latinos in the United States have been classified as heritage speakers, while it was confined to educational contexts. One criticism leveled by Baker and Jones is that this term is often associated with “powerful majority languages, [as] it points more to the past and less to the future, to traditions rather than to the contemporary ... [and] with ancient cultures, past traditions and more ‘primitive times’” (p. 509).

Nevertheless, Van Deusen-Scholl (2003) notes a preference for the term to connote strong cultural and family ties to a language, and Hornberger and Wang (2008) have also noted its use in stressing family and ancestral connections. Wiley (2014) argues that the elasticity of the term involves questions of the politics of identity involving who may be a legitimate heritage language learner, whether “outsiders” to a heritage language group will be encouraged to learn it, and whether proficiency or ethnic identity is considered more important in defining it. Valdés (1997) has noted that, in dual language immersion programs in the United States designed to promote Spanish and English, some in the Spanish-speaking community have been concerned about the power differential between the languages, and the concern among some Spanish-speaking parents about giving their language “away casually to the children of the powerful” (p. 393).

In this chapter, we briefly discuss the evolution of research on heritage and community languages with an emphasis on methodologies focused on learners and communities. We note methodological heritage language trends over time, with emphasis on dissertations published through the online ProQuest database from 1977 to May 1, 2016, and research articles published in the *Heritage Language Journal* from 2003 to April 1, 2016. We close with two sample studies that exemplify the issues raised in this chapter.

## Core Historical Background and Current Issues

### Fishman’s Legacy and the Evolution of Methodologies

In considering the core issues in heritage language-community language (HL-CL) research methodologies, it is necessary to consider the work and legacy of Joshua Fishman and his colleagues, more than a half century ago, in their groundbreaking work, *Language Loyalty in the United States*, which provided

a significant foundation for Fishman's subsequent body of work on HL-CL maintenance and shift. This volume also serves as a methodological reference point and inspiration for newer generations of scholarship interested in heritage, community, Native American languages, and revitalization. In revisiting this seminal work, it is also instructive to reflect on how linguistic diversity and language loyalty were conceptualized by Fishman and his colleagues in 1965 in the light of the recent emphases on hybridity and super-diversity.

In understanding the context of Fishman's pioneering interest in language loyalties among immigrant populations, it is important to note that the US government had just reformed its previously discriminatory immigration policies by ending more than four decades of racialized quotas, and it was addressing several language-related civil rights concerns through the initiatives of the Johnson administration (November 22, 1963 – January 20, 1969). Over the next few years, it also debated the use of bilingual education for language minorities in the United States and authorized the use of federal funds to promote linguistic accommodations. Fishman (1996a) concluded the volume with ambitious recommendations for promoting language maintenance efforts among heritage speakers and communities.

In *Language Loyalty in the United States*, Fishman (1996a) also made many assessments as well as predictions. He noted that in the United States much more attention had been paid to the process of Americanization than to the process of language self-maintenance among those retaining languages other than English. For the purposes of this chapter, the question remains: What are the research questions and methodologies that help to illuminate that process? *Language Loyalty in the United States* provided an initial model for answering this question. Based on the data and technologies available at that time, it focused on: (1) understanding the historical context of language diversity, immigration, and language maintenance through studies of specific language groups; (2) assessing the context of the relationship between language maintenance and nationality through demographic analysis; (3) analyzing heritage and community language use in broad social domains through community and institutional surveys by concentrating on, for example, the roles of ethnic language presses and foreign language broadcasting, ethnic language schools, churches, and institutions, as well as organizational and community dynamics as they related to language maintenance; and (4) demonstrating the importance of case study research on specific ethnolinguistic communities with particular reference to the historical and growing importance of Spanish speakers in the United States. The most prevalent methodologies concentrated on demographic analyses of the US Census and public directories, which identified newspapers and other public media. These were augmented by surveys involving language use and attitudes.

Additionally, Fishman stated that the “U.S. Census data on mother tongue must be the starting point for most studies on language maintenance in the United States” (1996, p. 419). As noted, throughout much of US history up to 1965, US immigration policies had skewed the composition of the US population in favor of European immigrants. The emphasis on European immigration also had implications for the types of questions and data collection by the US Census. As Fishman (1966) noted, studies on US data up to the 1960s had been biased in favor of European immigrant languages:

The Bureau of the Census has most consistently reported mother tongue data pertaining to the major European languages. Prior to 1960, Arabic was the only non-European language for which separate figures were reported. In 1960, Chinese and Japanese were added. Other non-European languages are presumably subsumed under the “all other (languages)” heading, so that these are not separately identifiable. (The Bureau’s plans to issue a report giving details on smaller languages, including those of Asia and Africa, have not as yet been implemented owing to budgetary limitations.)

In some cases the Bureau has also followed the custom of subsuming numerically “minor” European languages under more frequently reported languages to which they are linguistically or culturally related. In other cases, numerically “minor” languages have been indistinguishably included under the “all other” category. (p. 420)

As of 2010, the decennial US Census has eliminated asking language background questions. To find demographic information on languages among the US population, one must know and use the American Community Survey (ACS), which is based on a sample of the national population. Despite limitations of the survey, considerable self-reported information related to language background and even literacy and biliteracy for some languages can be gleaned from ACS data (see de Klerk & Wiley, 2008).

Building on *Language Loyalty in the United States*, Fishman continued his own work with other major publications (e.g., Fishman, 1999) addressing reversing language shift and language revitalization, in which there was growing interest, particularly for indigenous languages. Much of Fishman’s work had been cast in an immigrant languages paradigm; thus, the work of Hinton and Hale (2013) and McCarty (2011)—to name a few—focusing on indigenous languages has provided an important contrast. For many years, what now falls under the domain of HL-CL research was subsumed under other labels, particularly those of bilingual education and foreign language education. As federally supported bilingual education increasingly came

under attack during the 1990s, the heritage label tended to be used as alternative to *bilingual* education, which had become maligned in some political debates over language minority policy (García, Zakharia, & Otcu, 2013). This political shift from the use of the term from bilingual to heritage, however, also had implications for research priorities and methodologies. It shifted some of the focus from more contentious politicized debates focused on the efficacy of bilingual education, including transitional bilingual education, which from a language maintenance standpoint is a weak model (Baker & Wright, 2017), to ways in which maintenance could be encouraged.

Similarly, within the Canadian context heritage and community language programs have also been influenced by political debates and inconsistent funding. Prior to 1990, the Canadian Government provided financial support to community-based language and cultural programs through grants distributed per capita (CLA, 2011). Funds from these grants were used to subsidize the costs that community-based schools incurred, such as classroom rental space, instructional resources, and teaching materials. In 1991, the Canadian Department of Multiculturalism and Citizenship further developed their policy on heritage language with the goal of “assist[ing] Canadians to preserve, enhance and share their cultures, languages and ethnocultural group identities” (Dewing, 2009, p. 5). However, when the Government of Canada terminated national funding to community-based language programs, several provincial governments began disseminating various degrees of funding to linguistic and cultural communities throughout Canada (CLA, 2011). This also impacted the national heritage language curriculum to guide programs and provide reference to teachers on how to teach or what must be taught.

Like the United States, Canada’s labeling of heritage programs also saw a political shift from *heritage* to *international*. It has been documented that due to the challenges surrounding the *heritage* label, it has restricted program participants and mandated that they have ties to a specific language (Cummins, 2014). Therefore, the use of *international* became a more widely accepted label to represent students of several backgrounds, while allowing them to connect to other parts of the world through second language acquisition (Cummins, 2014).

## **Evolving Attempts to Establish Research Priorities for the Field**

In 1999, the Center for Applied Linguistics (CAL) and the National Foreign Language Center (NFLC) convened the first national conference on HLs in the United States. The conference included various stakeholders, including



researchers. Following the conference, a steering committee consisting of the faculty from different departments at the University of California at Los Angeles along with a working group of prominent scholars from a number of universities and centers was established. The working group focused on identifying research priorities for the emerging field, and a conference was held at UCLA in September of 2000 to “identify broad areas of research in heritage language education and within these areas to define key researchable questions that might be political, sociological, psychological, or linguistic in nature” (UCLA Steering Committee, 2000, p. 476). The findings of the conference were published as a report that was organized into seven broad areas: “the heritage speaker, the family, the community, a language-specific focus, policies, programs and assessment” (UCLA Steering Committee, 2000, p. 476). Among the recommendations were that:

Comprehensive surveys were, therefore, seen as essential to assess needs, resources, problems, and attitudes, to establish current and potential heritage language use, and to explore options for promoting heritage language education at the national, community and personal levels. [UCLA Steering Committee, p. 479]

At the programmatic level, the report recommended “case studies, models, and portraits demonstrating exemplary activities” (UCLA Steering Committee, p. 479). Research was also encouraged to focus on documenting family histories and “problems and strategies for maintain[ing] the use of a heritage language” as well as “reasons why a particular community has maintained its language, relevant policies, and examples of effective programs” (p. 479). These foci imply the need for case studies, survey and evaluation research, as well as policy analysis. Research questions focused on the individual, community, and family also implied the need for ethnographic and qualitative methodologies.

### **Current Trends in Dissertations and HL-CL Articles in the *Heritage Language Journal***

As noted, *Language Loyalty in the United States* (Fishman, 1996b) provided a model of what research questions and methodologies illuminate the process of language maintenance and revitalization. To understand what methodologies are employed in HL-CL learning, we embarked on a study of research trends. Therefore, we set forth to determine what methodological trends remained prevalent over time and what research methods were documenting demo-

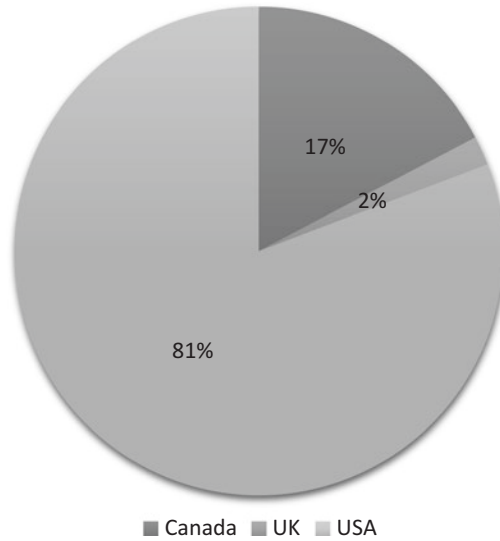
graphics, community dynamics and relations, and immigration as they relate to language maintenance.

To determine what research methods are prevalent in heritage and community language research, we conducted an advanced search in both the ProQuest online database of dissertations and theses and the *Heritage Language Journal (HLJ)*, an electronic journal. For the ProQuest search, we explored published dissertation abstracts from 1977 to May 1, 2016 that contained the words *heritage language* in the title. Additionally, we examined all volumes of the *HLJ* between 2003 and April 2016. This search led us to 156 dissertations in the last 19 years and 72 published studies in the *HLJ* in the past 13 years. It is important to note that several of the articles in the *HLJ* were not actual research studies but instead book reviews, theoretical essays, or pedagogical articles and thereby were not included in our analysis. In our analysis, we focused on dissertations from three English-speaking countries (the United Kingdom, the United States, and Canada) to determine where heritage language research is being published, what methods are being used, and what languages are prevalent in HL-CL research. The use of the heritage label has not always been accepted as the preferred label in all English-dominant countries and Australia; for example, a binational conference held in 2001 in Melbourne at the University of Victoria focused on heritage-community languages because of the preference to use “community in Australia” (see *International Journal of Bilingual Education and Bilingualism*, Vol. 8, Issue 2/3). The heritage label has primarily been mostly used in the United States and Canada although with a different connotation as previously noted.

### Data Collection Sites and the Labeling of HL and CL

The first step in our analysis of methodological trends in HL-CL was to determine where research was being published. While the *HLJ* has a global editorial board, we were intrigued in determining where recent graduates submitted their HL-CL dissertation studies. As Fig. 33.1 demonstrates, out of the 156 dissertations published in the last five years, 27 of them come from students in Canadian universities, with the majority of the dissertations submitted in US universities (80%) and only a small percentage of published work in HL-CL from the United Kingdom.

Therefore, it becomes clear that a majority of the HL-CL work has been published in American universities. We attribute this to several different factors, including an increase in immigration with residents who speak languages other than English at home (Rumbaut & Massey, 2014); greater interest among second- and third-generation Americans in maintaining and develop-



**Fig. 33.1** Heritage language dissertations between 2011 and 2016 by country

ing their heritage languages (Van Deusen-Scholl, 2003); rising trends in bilingual and dual language educational programs (Arias & Wiley, 2013); as well as augmented interest in attaining proficiency in the “critical” languages that are less-commonly taught (Bale, 2014). According to the Modern Language Association (2015), the above factors have all contributed to language offerings in less-commonly taught languages such as Arabic, Chinese, Hindi, Indonesian, Korean, Russian, and Turkish.

Based on the Canadian program label changes from “heritage” to “international,” we can conclude that this was an influencing factor in light of the number of published dissertations from Canada despite the plurilingualism of the country. As mentioned earlier, this categorical change from *heritage* to *international* limited program participants to students with ties to that particular language (Cummins, 2014). Similarly, we found that in the United Kingdom, the shift of classifying and labeling of “heritage” language programs to “complementary schools,” “supplementary schools,” and “community language schools” (Creese & Blackledge, 2010) may have had the same labeling impact of research being done within those programs.

## Research Methods

Once we had identified published HL-CL work, the next step was to categorize methodological trends prevalent in HL-CL dissertations published in the ProQuest database and articles in the *HLLJ*. Table 33.1 demonstrates the type

**Table 33.1** Type of data collected in HL-CL dissertations and articles

Methodology	Dissertations	Articles in the <i>Heritage Language Journal</i>
Quantitative	98	25
Qualitative	18	27
Mixed	31	6
Not stated	8	0

Source: ProQuest Dissertation Database and the *Heritage Language Journal*

of data collected (quantitative, qualitative, mixed methods, and not stated). As we examined the dissertation abstracts published in ProQuest, some of the authors did not identify the type of data collected or the methodology used, and we note this in our analysis.

Additionally, while the *HLJ* had almost equal numbers of articles in which quantitative and qualitative data were collected, the discrepancy between data methods was greater in the ProQuest dissertation abstracts. Out of the ProQuest dissertations, 98 of them utilized qualitative data, while only 18 of them collected qualitative data. However, there were far more mixed methodology dissertations (31) published compared to articles in the *HLJ* (6), which we can attribute to a larger collection of data in a dissertation that would eventually need to be whittled down for a research article. Through our analysis, we found that many of the dissertations and the articles in the *HLJ* employed more than one method in their data collection, which offered a triangulation of data preferred by researchers as results can be checked and measured from more than one distinct point or concept (Rothbauer, 2008). This further lends itself to researchers being able to determine “problems and strategies for maintain[ing] the use of a heritage language” (UCLA Steering Committee, 2000, p. 479).

Whereas the data collection types remain largely quantitative in nature, the methods employed provide a glimpse as to how the data was collected and what the foci of the studies were. In Table 33.2, we note the different trends in methodology and their popularity from dissertations and articles found in the *HLJ*. It is clear that methodologies such as observations, interviews, and questionnaires/surveys are commonly employed, as demonstrated in Table 33.2. This trend mirrors the need for qualitative methodologies that focus on the research priorities identified by the UCLA Steering Committee (2000) as they provide “reasons why a particular community has maintained its language, relevant policies, and examples of effective programs” (p. 479).

Almost half of the dissertations utilized observations of participants to collect data on HL-CL. Classroom and social observations allow the researcher to record events and witness language learning and use (Dörnyei, 2007). Observations provide the researcher multiple ways in which classrooms and other language learning sites become a myriad of options as they can be

**Table 33.2** Methodological trends in dissertations and articles

Methods	Dissertations published in ProQuest	<i>Heritage Language Journal</i> research articles
Observations	68	25
Interviews	14	35
Focus groups	14	0
Questionnaires	20	0
Surveys	19	45
Assessment (test/exam scores, diagnostic tool, etc.)	19	7
Text review	7	37
Document analysis	12	0
Interaction/conversation/verbal communication	8	0
Ethnography	4	6
Auto-ethnography	1	0

Source: ProQuest Dissertation Database and the *Heritage Language Journal*

structured or unstructured with the researcher being a participant or a non-participant as they examine the research site (Dörnyei, 2007). Ethnographic methods often consist of a multitude of methodologies such as interviews, focus groups, case studies, diary studies, research journals, and other reflective approaches to data collection (Dörnyei, 2007). As Lynch (2014) stated in his research on the foci of articles published in the *HLJ*, interviews were common for HL studies because they allowed researchers to study the “language of HL speakers as produced in interviews with researchers, in classroom conversations, in research-oriented tasks (e.g., sentence completion or cloze), and/or as observed behavior in grammaticality judgment tasks or perception experiments” (p. 224).

Questionnaires and surveys gauge language use and attitudes (Fishman, 1996b) and provide degrees of flexibility from open-ended to close-ended questionnaires and surveys giving the researcher more fluidity in their analysis. As stated by the UCLA Steering Committee, “case studies, models, and portraits demonstrating exemplary activities” (p. 479). Therefore, studies that employed several data collection methodologies provided the researcher a large degree of flexibility and lend themselves to different measures of the same concept for triangulation (Dörnyei, 2007).

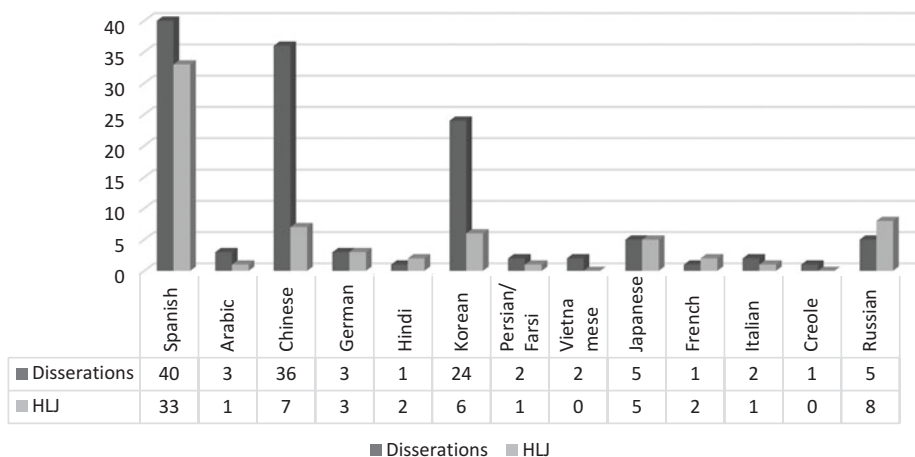
### **Heritage Languages Studied in Dissertations and Articles Found in the *Heritage Language Journal***

In Fig. 33.2, we see that Spanish, Chinese, and Korean were popular heritage languages of study. While Spanish HL-CL learning remains dominant, we can attribute this to the over two million 5–18-year-olds who live in homes

where languages other than Spanish are spoken (Fee, Rhodes, & Wiley, 2014; Wiley & Bhalla, 2017) as well as it being the most widely spoken non-English language (Rumbaut & Massey, 2013). However, the rise of Chinese and Korean is interesting to note as these are rarely offered for heritage language study in primary or secondary schools (Fee et al., 2014; Wiley & Bhalla, 2017)

However, despite the decline of federal support and the increase in state restrictions, much of the heritage and community language support has come from the respective local communities (You & Liu, 2011). As noted, both Chinese and Korean are among those that are classified as *critical* languages by the US government (Bale, 2014; You & Liu, 2011). Korea and China have been important trading partners of the United States, and they also are major sources of immigration to the United States; the American Community Survey estimates that there are over 2.5 million homes in which Chinese is spoken and over 1 million homes where Korean is the household language (U.S. Census Bureau, 2010; Wiley & Bhalla, 2017). Additionally, from our analysis of where the HL-CL research took place, we found that two of three HL-CL dissertations from the United Kingdom studied Chinese heritage language learning.

Both Hindi and Russian are also classified as *critical languages* by the US federal government, and the rising numbers in HL-CL studies are indicated in Fig. 33.2. Although Hindi is the official language of India, it may also serve as more of a cultural point of reference for Indian Americans and less of an



**Fig. 33.2** Language(s) studied by articles in the *Heritage Language Journal* and dissertations. Note: Some of the articles published in the *Heritage Language Journal* and dissertations published in the ProQuest Database contain the examination of one more heritage language

actual heritage language. As explained by Gambhir (2008), a heritage learner of Hindi may come from a family where Hindi or another Indian language is spoken in the home. Like Chinese and Korean, Russian is also seen as a language important for national security (Kagan, 2010). However, what distinguishes Russian as a HL-CL apart from other languages is the multiple waves of immigration, with many Russian heritage learners who are first or 1.5 generation despite the long-standing history of Soviet immigration going as far back as the twentieth century (Kagan, 2010). In the next section, we will discuss the challenges and issues that impact the future of HL-CL research.

## Challenges and Unresolved Issues

### The Importance of Definitions in HL-CL Research

Further complicating the issue of definitions used in HL-CL are issues of status differentials between purported languages and dialects. As noted by Wiley (2014), some purported HL learners may not have had direct family or community contact with their alleged heritage language. Among the dispersed/diasporized Chinese ethnic groups, for example, the issue of which variety of *Chinese* constitutes one's heritage language may be problematic, if it is assumed that they should have an only one heritage language. Consider the case of those who are so-called dialect speakers of varieties other than Mandarin. Thus, research studies on Chinese families and communities need to consider the multilingualism of the context because Mandarin may function as a *surrogate* heritage language, which can connect ethnic Chinese children with Chinese culture (Liu & Wiley, 2013; Wiley, 2008; Wiley et al., 2008). Similarly, Hindi might be selected as *heritage* language for those of Indian heritage who speak other languages. Even more complex cases involve those who live in highly multilingual contexts, where it does not make sense to try to identify a single *heritage* language. In addition to these concerns, Wiley (2014) has suggested that a number of factors need to be considered, such as the type of educational program, learner characteristics, the nature of the community, and its sociolinguistic context in relation to other communities.

Research focus on pedagogical considerations, where a target heritage language is identified for instructional purposes, however, has proposed “broad” and “narrow” definitions. In this regard, Carreira and Kagan (2011) note:

1. A broadly defined HL is part of that person's family or cultural heritage, the language may not have been spoken in the home, and the person has

no functional proficiency in the language and would most likely have to study that language as an L2 [second language] learner.

2. A narrowly defined HL “was first in the order of acquisition but was not completely acquired because of the individual’s switch to another dominant language” (Polinsky & Kagan, 2007, p. 369). Unlike broadly defined heritage speakers, who are like second language (L2) learners, linguistically speaking, narrowly defined heritage speakers bring to the classroom some measure of competence in the language (Carreira & Kagan, 2011, pp. 41–42).

Pedagogical definitions such as these have utility for research on learner assessment related program placement and, thereby implicitly place emphasis on prior schooling and literacy considerations. This also tends to place emphasis on the assessment of proficiencies in standardized varieties of language versus non-school varieties that are presumed to represent HL speakers with lower proficiency (Polinsky & Kagan, 2007). The question remains, if we are not to privilege only school varieties of languages, it is important to note that many HL-CLs are not formally taught (Wiley, 2014). A focus only on school-based norms of proficiency may ignore how many HLs are actually used and stigmatize their speakers.

### Sample Study 33.1

You, B., & Liu, N. (2011). Stakeholder views on the roles, challenges, and future prospects of Korean and Chinese heritage language-community language schools in the Phoenix metropolitan area: A comparative study. *Heritage Language Journal*, 8, 67–92.

#### Research Background

Chinese and Korean are languages that have been classified as “critical” by the United States and have large numbers of speakers in the country. This study examines how these heritage and community language schools and educational programs deal with challenges and future prospects.

#### Research Problems/Gaps

Due to the decline of federal support for public bilingual education, there has been an increase in heritage and community language programs by local communities. This study sets out to answer what are the roles, major challenges, and future prospects of Korean and Chinese HL-CL schools in the United States according to the perceptions of stakeholders?

#### Research Method

- *Type of research*: This is a comparative study which examines stakeholders’ perspectives on Korean and Chinese HL-CL schools.
- *Setting and participants*: This study took place in the Phoenix metropolitan area and participants were Korean and Chinese teachers and parents from five Korean and five Chinese HL-CL schools.



- *Instruments/techniques*: Data was collected through surveys and in-depth interviews with school administrators, teachers, and parents.
- *Data analysis*: The survey data were coded and analyzed using the statistical data software SPSS. Interview data was coded, categorized, and theme searched using the qualitative data analysis software NVIVO 7.

### Key Results

These findings of this study show that both Korean and Chinese stakeholders viewed the HL-CL schools and their education as very important not only for maintaining children's HL-CL, but also to help them build a positive sense of cultural and ethnic identity.

### Comments

This mixed-methods study demonstrates how issues impacting heritage language programs are not limited to one language, but can impact speakers of many different heritage languages. In this comparative study examining the role of stakeholders, it is noted that community resources and support can assist with positive language development and maintenance.

## Sample Study 33.2

Kondo-Brown, K. (2009). Heritage background, motivation, and reading ability of upper-level postsecondary students of Chinese, Japanese, and Korean. *Reading in a Foreign Language, 21*(2), 179–197.

### Research Background

The association between L2 motivation and the L2 success is often attributed to several different motivational constructs. This study examines to what degree extrinsic and intrinsic motivation constructs and L2 achievement are related.

### Research Problems/Gaps

This study investigated how learner beliefs, affective variables, and Japanese L2 reading ability are associated with *kanji* learning variables, *kanji* strategies, and *kanji* inferencing ability. The research question that guided this study was: among L2 learners of advanced college-level Japanese, to what degree are affective factors related to students' demonstrated reading comprehension and *kanji* knowledge?

### Research Method

- *Setting and participants*: The participants were 43 English first language (L1) learners of advanced Japanese at the University of Hawai'i at Manoa. Twenty-four of the participants were female and 19 were male, with 24 of the participants having a connection to Japanese heritage.
- *Instruments/techniques*: There were several instruments used in collecting data from the study participants, which included a reading comprehension test, a *kanji* knowledge test, affective subscales, self-determination motivation subscales, Japanese language learning beliefs subscales, and Japanese L2 reading motivation subscales.
- *Data analysis*: All quantitative data was analyzed using descriptive statistics. Different correlations between proficiency measures, affective variables, Japanese reading-specific affective variable were drawn.

**Key Results**

Findings from this study demonstrate that affective variables have direct and indirect associations with the development of L2 reading ability of these students. Participants who had been studying the language for six years demonstrated noticeable differences in reading comprehension ability and *kanji* knowledge.

**Comments**

This quantitative study examines how motivational constructs impact language learning for both heritage speakers and non-heritage learners of Japanese. Findings from this study lend themselves to pedagogical implications, which enhance heritage and foreign language learning.

## Resources for Further Reading

Jansen, L. (2007). Heritage language reading in the university: A study of students' experiences, strategies, and preferences. *Heritage Language Journal*, 5, 98–116.

This study examines the reading experiences, strategies, and curriculum preferences of 124 university-level heritage language students enrolled in four different heritage language programs at UCLA—Korean, Russian, Thai, and Vietnamese. Study participants were surveyed about their exposure to print, early and current reading experiences, strategy use, learning goals, and preferences. Findings from this survey revealed that most students were interested in achieving university-level academic reading proficiency, despite reporting that they spend little time reading in their HL despite the availability of print in many of their homes.

Nik, G. (2012). Hindi heritage language learner's performance during OPIs: Characteristics and pedagogical implications. *Heritage Language Journal*, 9, 18–36.

This study focuses on Hindi heritage language learners' oral performance doing practice oral proficiency interviews (OPI). The focus of this paper is on grammar patterns and discourse strengths and deficiencies with data collected by interviews between the Hindi learners. Findings from this study concentrate on pedagogical implications on teaching strategies that highlight the heritage learners' versus foreign language learners' language abilities in Hindi.

O'Rourke, P., & Zhou, Q. (2016). Heritage and second language learners: Different perspectives on language learning. *International Journal of Bilingual Education and Bilingualism*, 7, 1–10.

Attitudes of HLLs and second language learners were surveyed about their experiences in classroom language learning as well their overall language learning experience in this article. Study participants were high school students who participated in the STARTALK summer language study program in 2014 and 2015. Findings from this study found that compared to second language learners, HLLs were less motivated to study their or another language and were less likely to believe that learning a new language would help them academically and professionally.

Sekerina, I. (2013). A psychometric approach to heritage language studies. *Heritage Language Journal, 10*, 203–210.

In this article, Sekerina outlines why assessment tools that identify the weaknesses and strengths of HLLs through their HL progression are needed. The lack of standards for HLLs that are unique is often the main issue of HL research, and Sekerina notes that for HLLs age of arrival is more of a critical factor than age of acquisition. Findings from this article call for HLL research that uses more sophisticated psychometric measures other than ANOVAs and *t*-tests.

Thompson, G. L. (2014). Understanding the heritage language student: Proficiency and placement. *Journal of Hispanic Higher Education, 14*, 82–96.

Focusing on the fast-growing number of Spanish HL in both the university and college settings, this article outlines the need for placement exams that correctly measure language ability, easy to administer to large groups of students, and are simple to evaluate. Thompson investigates the implementation of a placement exam at a large public university, assesses the results, and recommends ways to further develop heritage placement exams for Spanish HL students.

Xiao, Y., & Wong, K. F. (2014). Exploring heritage language anxiety: A study of Chinese heritage language learners. *Modern Language Journal, 98*, 589–611.

The authors in this paper recommend a set of connections between HL studies and theory construction noting the benefits that this population of study offers for linguistics research. In particular, this article focuses on several grammatical phenomena from the standpoint of their representation in HLLs, includ-

ing case, aspect, and other interface phenomena as well as how HL speakers influence current debates about language acquisition under different conditions.

## References

- Arias, M. B., & Wiley, T. G. (2013). Language policy and teacher preparation: The implications of a restrictive language policy on teacher preparation. *Applied Linguistics Review*, 4, 83–104.
- Baker, C., & Jones, S. P. (1998). *Encyclopedia of bilingual education and bilingualism*. Clevedon, UK: Multilingual Matters.
- Baker, C., & Wright, W. (2017). *Foundations of bilingual education and bilingualism*. Clevedon, UK: Multilingual Matters.
- Bale, J. (2014). Heritage language education and the “national interest”. *Review of Research in Education*, 38, 166–188.
- Canada Languages Association (CLA). (2011). *Overview*. Retrieved from <http://www.canadianlanguages.ca>
- Carreira, M., & Kagan, O. (2011). The results of the National Heritage Language Survey: Implications for teaching, curriculum design, and professional development. *Foreign Language Annals*, 44, 40–64.
- Creese, A., & Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching? *Modern Language Journal*, 94, 103–115.
- Cummins, J. (2014). Mainstreaming plurilingualism: Restructuring heritage language provision in Canadian schools. In P. Trifonas & T. Aravossitas (Eds.), *Rethinking heritage language education* (pp. 1–19). New York: Routledge.
- Dewing, M. (2009). *Canadian multiculturalism*. Ottawa: Library of Parliament.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: Oxford University Press.
- Fee, M., Rhodes, N. C., & Wiley, T. G. (2014). Demographic realities, challenges, and opportunities. In T. G. Wiley, J. K. Peyton, D. Christian, S. K. Moore, & N. Liu (Eds.), *Handbook on heritage, community, and native American language education in the United States: Research, policy and practice* (pp. 6–18). London: Routledge.
- Fishman, J. A. (1996a). Appendix A: Methodological notes. In J. A. Fishman, V. C. Nahirny, J. E. Hofman, & R. Hayden (Eds.), *Language loyalty in the United States: The maintenance and perpetuation of non-English mother tongues by American ethnic and religious groups* (pp. 414–423). London: Mouton & Co.
- Fishman, J. A. (1966b). Appendix B. Language maintenance and language shift as a field of inquiry. In J. A. Fishman, V. C. Nahirny, J. E. Hofman, & R. Hayden (Eds.), *Language loyalty in the United States: The maintenance and perpetuation of Non-English mother tongues by American ethnic and religious groups* (pp. 424–458). London: Mouton & Co.

- Fishman, J. A. (1999). *Handbook of language and ethnic identity*. New York: Oxford University Press.
- Gambhir, V. (2008). The rich tapestry of heritage learners of Hindi. *South Asia Language Pedagogy and Technology*, 1, 5–23.
- García, O., Zakharia, Z., & Otcu, B. (2013). Introduction. In O. García, Z. Zakharia, & B. Otcu (Eds.), *Bilingual community education and multilingualism: Beyond heritage languages in a global city* (pp. 3–42). Bristol: Multilingual Matters.
- Hinton, L., & Hale, K. (2013). *The green book of language revitalization in practice*. Bingley, England: Emerald Group Publishing.
- Hornberger, N. H., & Wang, S. C. (2008). Who are our heritage language learners? Identity and biliteracy in heritage language education in the United States. In D. M. Brinton, O. Kagan, & S. Bauckus (Eds.), *Heritage language education: A new field emerging* (pp. 3–38). New York: Routledge.
- Kagan, O. (2010). Russian heritage language speakers in the U.S.: A profile. *Russian Language Journal*, 60, 213–228.
- de Klerk, G., & Wiley, T. G. (2008). Language policy: Using the American community survey to investigate bilingualism and biliteracy among immigrant communities. *Journal of Southeast Asian American Education and Advancement*, 3, 68–78. <https://doi.org/10.7771/2153-8999.1112>
- Liu, N., & Wiley, T. G. (2013). Attitudes toward Mandarin, dialects, and language diversity among Chinese immigrants and international students in the United States. In M. Zhou (Ed.), *Selected symposium papers on language planning in Modern China: One hundred years*. Beijing, China: Commercial Press.
- Lynch, A. (2014). The first decade of the *Heritage Language Journal*: A retrospective view of research on heritage languages. *Heritage Language Journal*, 11(3), 224–242.
- McCarty, T. L. (2011). *Ethnography and language policy*. New York: Routledge.
- Modern Language Association. (2015). *Highlights of the MLA's 2013 survey of enrollments in languages other than English*. Modern Language Association.
- Polinsky, M., & Kagan, O. (2007). Heritage languages: In the “wild” and in the classroom. *Languages and Linguistics Compass*, 1, 368–395.
- Rothbauer, P. (2008). Triangulation. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (pp. 892–894). Thousand Oaks, CA: Sage.
- Rumbaut, R. G., & Massey, D. S. (2013). Immigration and language diversity in the United States. *Daedalus*, 142(3), 141–154.
- U.S. Census Bureau. (2010). *B16001 Language Spoken at Home by Ability to Speak English for the Population 5 Years and Over* [data table]. American Community.
- UCLA Steering Committee. (2000). Heritage language research priorities conference report. *Bilingual Research Journal*, 24, 475–484.
- Valdes, G. (1997). Dual-language immersion programs: A cautionary note concerning the education of language-minority students. *Harvard Educational Review*, 67(3), 391–430.
- Van Deusen-Scholl, N. (2003). Toward a definition of heritage language: Sociopolitical and pedagogical considerations. *Journal of Language, Identity, and Education*, 2, 211–230.

- Wiley, T. G. (2008). Dialect speakers as heritage language learners: A Chinese case study. In D. M. Brinton, O. Kagan, & S. Bauckus (Eds.), *Heritage language: A new field emerging* (pp. 91–106). New York: Routledge.
- Wiley, T. G. (2014). The problem of defining heritage and community languages and their speakers: On the utility and limitations of definitional constructs. In T. G. Wiley, J. K. Peyton, D. Christian, S. K. Moore, & N. Liu (Eds.), *Handbook on heritage, community, and native American language education in the United States: Research, policy and practice* (pp. 19–26). London: Routledge.
- Wiley, T. G., & Bhalla, S. (2017). The demographics of heritage and community languages in the United States. In O. Kagan, M. Carreira, & C. Chik (Eds.), *A handbook on heritage language education: From innovation to program building*. New York, NY: Routledge.
- Wiley, T. G., De Klerk, G., Li, M.-Y., Liu, N., Teng, Y., & Yang, P. (2008). Language attitudes toward Chinese “dialects” among Chinese immigrants and international students. In A. He, & Y. Xiao (Eds.), *Chinese as a heritage language in the United States* (pp. 67–87). Monograph, National Foreign Language Resource Center, University of Hawaii at Manoa. Honolulu, HI: University of Hawaii Press.
- You, B., & Liu, N. (2011). Stakeholder views on the roles, challenges, and future prospects of Korean and Chinese heritage language-Community language schools in Phoenix: A comparative study. *Heritage Language Journal*, 8, 259–284.



# 34

## Translation and Interpreting

Claudia V. Angelelli

### Introduction

This chapter is about research issues, traditions, and methods in translation and interpreting (T&I). T&I enable readers, speakers, and signers from different linguistic backgrounds to access information in a different language than their own and to communicate with each other. While the term “translation” refers to the manipulation of written, video, or electronic text by humans or machines, “interpreting” is used for the communication of spoken or sign languages in real time, face-to-face or remotely. Interpreting involves human interaction during a communicative event across languages. Translation involves the delivery of information that may or may not require human interaction in real time.

Even when the origins of translation can be traced to the times of Cicero, and interpreting has always been referred to as “the world’s second oldest profession,” the scholarly study of T&I, as well as the development of translation and interpreting studies (TIS) as a field of inquiry in its own right, together with the resulting empirical and theoretical research produced in the field, are viewed as fairly recent phenomena (Angelelli & Baer, 2015). The growth of this scholarly field is influenced, in part, by a more direct engagement with related disciplines (e.g., bi/multilingualism, cognitive psychology, cultural

---

C. V. Angelelli (✉)

Department of Languages & Intercultural Studies, School of Management & Languages, Heriot-Watt University, Edinburgh, UK  
e-mail: [C.Angelelli@hw.ac.uk](mailto:C.Angelelli@hw.ac.uk)

studies, educational linguistics, linguistic anthropology, psycholinguistics, and sociology of language, to name just a few). Evidence of this growth is the number of scholarly journals, the interest of publishing companies in this area of study, the appearance of new doctoral programs, and the increase of master's offerings in TIS in various parts of the world. In addition to the increasing interest in academic curriculum developments, the supply of short courses and seminars, webinars, blended and online training opportunities offered by individuals and companies, with varying degrees of expertise and success, has also raised questions about professionalization and quality in T&I. These issues, as well as additional ones that are discussed later in this chapter, are discussed in the two sample studies that appear at the close of this chapter.

The growth of interest in TIS is also due to recent developments in technologies that call for interdisciplinary collaboration (e.g., expertise in TIS coupled with expertise in mathematics and computer science to develop software applications in T&I), the increase in the volume of digital communicative events across languages and the resulting need for T&I, as well as the more pressing social, political, and economic issues that result in geographic displacement and migration (Angelelli, 2011, 2012).

## Current and Core Issues

Initially, most discussions in T&I have centered on issues of cognitive and linguistic abilities, and the focus was primarily on literary translation and conference interpreting (Gile, 2002). Most of these initial discussions concerned students' selection, competences, their performance at the tasks at hand, and "research related to T&I pedagogy relied almost entirely on anecdotal accounts, case studies at best, detailing how translation/interpreting was taught in a particular institution or program" (Colina & Angelelli, 2015, p. 110).

During the 1990s, studies focused less on classroom and more on naturalistic settings. Consequently, research topics included more types of T&I, beyond literary translation or conference interpreting (e.g., technical, scientific, legal, or educational translation, stressing processes in addition to products, and machine-assisted translation). Studies in interpreting that had initially been experimental in nature and mostly about conference interpreting (e.g., Gile, 2002) have broadened their scope to include other settings [e.g., legal (Angelelli, 2015; Berk-Seligson, 2002; Hale, 2004; Morris, 2010), medical (e.g., Angelelli, 2004a; Metzger, 1999; Meyer, 2012), educational (e.g., Angelelli, 2016; Valdés, Chavez & Angelelli, 2000), and social work (e.g., Tipton, 2010)] and used a variety of approaches and methods, from



ethnography of communication (Angelelli, 2004a), focus groups (Angelelli, 2007), observations (Bolden, 2000), case studies (e.g., Karoly, 2014; Kuo & Nakamura, 2005) to experimental (e.g., Dimitrova, 2005) and survey-based studies (e.g. Angelelli, 2004b; Li, 2000).

Interestingly, despite the importance of T&I in cross-linguistic/cultural communication, these constructs sometimes appear to be subordinated or subsumed in larger ones. For translation this occurs with the term “language industry,” and translation gets overshadowed in software localization or project management. And the same appears to be true for the construct of interpreting, as discussions on interpreting processes and products appear to be subsumed in other issues (e.g., access to services on the part of linguistic minorities or technological developments).

### Sample Study 34.1

Zimányi, K. (2013). “Somebody has to be in charge of a session”: On the control of communication in interpreter-mediated mental health encounters. *TIS Translation and Interpreting Studies*, 8, 94–111.

#### Research Background

A study on mental health interpreting, with a specific focus on control and how participants perceive control issues in interpreted-mediated encounters. The larger study, of which this research is a part, sheds light on how narratives—specifically the stories told by clients—are co-constructed with the interpreter and therefore how the interpreter’s presence and renditions influences the narratives.

#### Research Problems/Gaps

This study explores how the participants of interpreter-mediated encounters in mental health settings (specifically occupational therapy, psychology, and mental health nursing) take control of the communication flow.

#### Research Method

- *Setting and participants:* From two areas of mental health, occupational therapy and mental health nursing, 11 mental health professionals (including 4 mental health nurses, 1 occupational therapist, 2 psychologists, and 4 therapists and psychotherapists) who have worked with interpreters took part in this study, as well as 12 interpreters (with experience in mental health settings) representing the following minority-language groups in Ireland: Bosnian, Chinese, Czech, Italian, Polish, Romanian, Spanish, and Sudanese.
- *Instruments/techniques:* A series of semi-structured interviews (approximate length between 21 and 84 minutes with an average of 45) about interpreted narratives in mental health settings in Dublin, Ireland. Protocol included: (1) the perception of mental health among respondents; (2) the cultural and situational significance of narratives; (3) the interpreter’s familiarity with the evolving narrative and with the participants in the interpreter-mediated encounter; (4) modes of interpreting; (5) interpreting narratives in mental healthcare; and (6) general issues on interpreting in mental healthcare settings.
- *Data analysis:* The thematic coding process was facilitated by the use of NVivo.

**Key Results**

In the interpreter-mediated encounter, when it comes to individual participants, there is not one single interlocutor who dominates. Respondents indicated that although the client's story/narrative may not always flow or may be restricted at times, clients may temporarily take control. The two providers involved (mental health professional and interpreter) share the communication flow with the interpreter being responsible for the transferring of meaning between the two interlocutors who do not share a language.

**Comment**

This work is important as it sheds light on an area of healthcare interpreting that is currently under-researched, that is, mental health with a specific focus on occupational therapy and nursing. In addition, as results appear to contradict the assumption that healthcare providers dominate the talk in monolingual and bilingual encounters, it opens an opportunity for further research across healthcare settings as well as languages.

In translation studies, initial discussions on the construct of translation focused on notions of equivalence and faithfulness to the source text (the original). In those the concerns were more about the product than about the process, and the agency/role of the translator was not considered an issue that deserved attention. The concern on faithfulness to the source text and on finding equivalency was questioned as researchers were thinking of adaptation rather than translation (e.g., Bastin, 1998). This shift beyond equivalent response gave way to the functionalist approach to translation (Nord, 1997). No longer preoccupied by the text equivalency but rather by its function, this new approach continues to be in vogue and is complemented by reader response theory.

In addition, more contemporary discussions in literary translation and theories of translation center around the power of translators (Bandía, 2008), their role (Inghilleri, 2005), and visibility (Venuti, 1995) as they are considered agents of change, manipulating text and playing a key role in foreignizing or domesticating texts/cultures for target readership. In Interpreting Studies, rather than focusing on the interpreter's recognition or prestige, discussions focused on interpreter's participation in the interaction as a result of exercising her agency (Angelelli, 2004a, 2004b; Hsieh, 2016) rather than the interpreter being perceived as a mere channel or language encoder/decoder. In its inception, interpreting was conceptualized as an exchange between two speakers who did not share a language with an invisible interpreter facilitating it. This model:

- portrays an invisible interpreter who is a mere conduit or channel between two speakers who do not share a common language;
- assumes no interaction between interpreter and speakers, and no direct interaction between speakers except through the interpreter;

- pretends that interpreting can happen in a social vacuum;
- equates the interpreter to a language modem;
- presupposes that in any given utterance there is only one meaning, which is independent of the parties to the interaction (interpreter included) rather than socially co-constructed.

Since the 1990s we have witnessed a shift in the perception of the interpreter's role, from a language conduit to an essential partner in a cross-cultural conversation or co-constructor in the interaction (Roy, 2002; Wadensjö, 1998) to a co-participant with agency (Angelelli, 2004a, 2004b; Davidson, 2000). Through ethnography of communication, participant and nonparticipant observations, focus groups, interviews, discourse analysis, survey-based studies, and corpora of transcribed interactions, research has shed light on interpreters' role. Nevertheless, the conceptualization of the interpreter as a conduit is still prevalent, especially among practitioners and trainers, both in sign and spoken languages. This is evident in various publications about ethics.

Various professional associations of translators and interpreters (e.g., The American Translators Association) have produced codes of ethics in which they describe (or prescribe) the role of the translator/interpreter. The California Healthcare Interpreting Association was the first and only one to base their code on empirical research (see Angelelli et al., 2007). It recognizes that the interpreter has multiple roles as a party in the patient/clinician interaction in addition to the conduit role portrayed in earlier writings. These roles are (CHIA, 2002, pp. 35–42):

- Message converter which equates the conduit model discussed on page 5.
- Message clarifier “includes gaining more information from a speaker to explain a message or concept in an alternate or more easily understood manner to facilitate communication” (ibid., p. 36).
- Cultural clarifier involves confirming and providing cultural information, particularly about cultural health beliefs (ibid., p. 37).
- The fourth role, patient advocate (ibid., pp. 39–42), prohibited for the court interpreter, may be required in the healthcare setting to support the health and well-being of the patient.

Unlike statements in codes of ethics from professional associations that tend to be prescriptive in nature, empirical research shows that both translators (Baker, 2006; Inghilleri, 2005; Tonkin & Frank, 2010; Tyulenev, 2015) and interpreters (Angelelli, 2004a; Berk-Seligson 2002; Davidson, 2000) impact processes and products. It is not a coincidence that as translation and

interpreting become the target of more empirical research, what in its inception was characterized as a craft, or an art, anecdotal or experiential-based knowledge has become an evidence-based practice and a field of inquiry in its own right.

Technological developments also constitute a core issue in T&I. The arrival of new technology and software applications dealing with communication has impacted several aspects of T&I. For interpreting, multimodal and remote interpreting is an example. By using different platforms of video/audio communication, users of languages of limited diffusion or from remote areas or sign-language users can access services in a more efficient way. When there is no interpreter on site, by connecting to this service, hospitals, schools, government offices, and courts of law can now offer language services to users of non-societal languages. For translation, in addition to machine translation or CAT tools, social media (e.g., Facebook, Instagram, and Google Translate) provide another example of how technology has impacted translation. Social media has enabled crowdsourcing, and volunteer and collaborative translation (Jimenez-Crespo, 2015).

Additionally, technology has facilitated data collection from virtual communities of translators, editors, and project managers. This data on issues that translators face as they work (captured via the use of eye trackers and screen sharing) are helping the industry develop better products for computer-assisted translation tools (e.g., memory management software).

## Challenges and Controversial Issues

Within TIS, several challenges and controversial issues include the need to link theory/research and practice (as previously discussed), translators' and interpreters' education, and their status and professionalization. These three are highlighted as examples and discussed separately. They are, however, definitely interconnected.

Similar to many other professions (Sela-Sheffy & Shlesinger, 2011), translation/interpreting was initially perceived as a practice, an art or a craft. It was also thought of as a by-product of bilingualism. Throughout time we have witnessed different degrees of tension among three groups in this field: individuals who practice the profession (practitioners) and who may or may not hold a degree in the field; individuals who teach translation or interpreting (teachers) and who may or may not practice the profession or hold a degree in the field; and individuals who focus on translation, interpreting, or both as objects of study (researchers) and who may or may not practice or teach translation and/or interpreting. Although several individuals share member-

ship in the three groups and more in two (practice and teaching), this is not common practice (Gile, 1995).

*Par excellence* a site of contact between theory/research and practice is the education of future translation/interpreting professionals as well as their testing and certification. Currently, given the limited opportunities available to develop a fine cadre of T&I teachers, coupled with the importance placed on practical experience T&I classrooms are generally led by practitioners (who may or may not be versed in pedagogy) or language teachers (who may or may not be versed in translation and interpreting). T&I instructors rely on their practice and their experiences to teach incoming students. Although sharing experiences and anecdotes with students can account for some aspects of teaching, designing curriculum around practical experiences as a pedagogical approach is risky. Learning by trial and error or assuming “one size fits all” is not sound pedagogy. Experiential information, albeit interesting, is generally limited to one person’s individual experience, personal opinions, or anecdotes. T&I classrooms, like any other classroom, function better when tasks and activities are conducted in a student-centered learning environment where student-learning outcomes are attainable, content progresses in order of complexity, and issues are discussed providing the right amount of scaffolding. This once again is better achieved when T&I practitioners who are responsible for T&I instruction can pair with applied linguists or translation and interpreter researchers and engage in team teaching and problem solving through, for example, action research (Nicodemus & Swabey, 2015). An empirically based education would allow future T&I professionals to learn about the specifics of the industry/setting in which they work and gain awareness of the nature of situated practices (Angelelli, 2007).

Therefore, an education that integrates research and practice would differ from traditional models based on training. Not only would students understand how their practice is grounded, but also their assessment would be valid, reliable and more inclusive of competencies beyond traditional ones. Currently most measurement instruments and scoring procedures in T&I focus on cognitive (analytical skills, information processing, memory) and linguistic skills (language proficiency and specific terminology). Although both of these are essential subcomponents of T&I competence, they do not constitute the whole (Angelelli & Jacobson, 2009). As research has demonstrated, other skills, for example, interpersonal or social ones (Angelelli, 2004a, 2004b; Gavioli, 2012; Inghilleri, 2005), are as crucial as cognitive and linguistic skills but are seldom taught and almost never measured. This means that constructs such as neutrality, objectivity, and invisibility are assumed but are not tested. A more profound dialogue between theory and practice would result in replicating the reality of translators and interpreters at work during assessment.

Ethical, interpersonal, and social issues (such as problem solving, teamwork or time management for translators or alignment, affect, trust, and respect) should be accounted for in assessment of students of T&I rather than taken for granted or simply ignored. We cannot afford not to test what is either an essential behavior in good performance or an absolute inappropriate behavior that would render a performance unacceptable. Instead of neglecting or taking for granted social and interpersonal skills, programs would be testing them side by side with cognitive and linguistic ones. In so doing, testing becomes more integrative of all the dimensions present in any T&I assignment. This encompassing approach to testing would provide a more thorough and precise view of the candidates' abilities. This can only happen as a result of a meaningful dialogue between theory and practice. In sum, research produced during the last two decades has been groundbreaking in expanding our knowledge on translators' and interpreters' roles, their complexity, and responsibility, and the processes embedded in their practice.

In terms of professionalization, the status of translators and interpreters has been described as insecure, marginalized, and ambivalent for some time (Hammond, 1994). In fact, to some researchers (e.g., Sela-Sheffy & Shlesinger, 2011), translators and interpreters constitute an interesting case of an under-scrutinized professional group. Some studies have been conducted in the area of status and field—for example, legal translators and interpreters in Spain (Monzó, 2009), the struggles in the professionalization of an emerging sector such as the in-house translators in Hong Kong (Chan, 2009)—as well as in role and identity of translators (Meylaerts, 2010), interpreters (Morris, 2010), and bilingual youth interpreting for their families and immediate communities (Angelelli, 2016).

In addition, within the continuum of ad hoc interpreting, the case of bilingual youth who have interpreted for their families and immediate communities has been the focus of various studies. Research on circumstantial bilinguals who become young interpreters for their families and communities contributes to our understanding of the life experiences of individuals who begin to interpret early in their lives (Valdés & Angelelli, 2003). This broader view, which commonly refers to language mediation (Antonini, 2011), (child) language brokering, and nonprofessional translation and interpreting (Ervin & Meyer, 2016), has shed light on the challenges faced when students are non-mainstream elective bilinguals (Valdés & Figueroa, 1994).

## Limitations and Future Directions

Translation and interpreting are, by nature, interdisciplinary. TIS has benefited from the synergies with related fields of inquiry such as bilingualism, cognitive psychology, sociolinguistics, sociology of language, among others as new frameworks and research methods from these related fields began to be adopted or adapted for conducting research in T&I (see Angelelli & Baer, 2015). Calls for more interdisciplinary work have been made for quite some time. Interestingly, many of these calls have not been addressed, so we still encounter limitations in how research is conducted. When studies in T&I do not have a strong base on research methods, for example, we see confusion generated by making claims that cannot be sustained, by not grounding research on underlying conceptual or theoretical frameworks, ignoring previous research because it may be produced outside the field of T&I practice, and quite frequently not differentiating between descriptive and prescriptive approaches.

When it comes to research methods, interdisciplinarity is crucial, as expertise in a field of study is essential to address a phenomenon, as is mastery in conducting research. Many times, individuals working as translators and interpreters, as well as researchers on these topics, bring insightful knowledge to research groups working on, for example, artificial intelligence, multimodal communication, or software development. Pairing researchers and practitioners, as in action research, is not uncommon, especially in T&I pedagogy. As a result of interdisciplinary research conducted over several decades, a more comprehensive and nuanced picture of the contexts, participants, and processes involved in as well as the products resulting from T&I has emerged.

By nature T&I is intrinsically related to applied linguistics, cognitive psychology, cognitive translatology, and bi/multilingualism. While previous studies in the intersection of T&I and bilingualism (Malakoff & Hakuta, 1991; Muñoz Martin, 2013; Shreve, 1997; Valdés & Angelelli, 2003) have shed light on some cognitive, linguistic, and educational processes in translators and interpreters, several issues remain to be studied. This is especially true for the distinction between elective and circumstantial bilinguals (Valdés & Figueroa, 1994). The relationship between language and identity as well as linguistic behaviors exhibited by professional and nonprofessional interpreters and translators needs to be further explored. Specifically studying how circumstantial bilinguals who brokered communication for their families when they were young grow into translators and interpreters or are critical



users of translation and interpreting (Angelelli, 2016) could provide us with a fuller picture of the needs of students populating T&I classrooms as well as the T&I end-user perspective and experience.

In addition, although TIS has taken a sociological turn for some time now (Angelelli, 2012; Sela-Sheffy & Shlesinger, 2011; Wolf & Fukari, 2007), several issues remain to be explored. As social agents, translators and interpreters are in a unique position to produce, reproduce and impact cultural production. We still need to understand how, when, and why they engage in these practices and what their level of awareness is about such engagement. Specifically in translation studies, when we consider the impetus of reader response and reception theory, we still need to explore the effects of readers' subjectivity (Chan, 2009) in translated texts (beyond literary ones). We also need to understand how, when, and why readers move across cultural and linguistic borders and what the interaction is among publishing policies, publishing houses, readers, authors, and translators over translated and non-translated work.

In predicting future directions and further research, given the pace of current technological developments and the social media penetration rate, we could hypothesize that future research agendas will call to study the views and concepts of ethics and responsibility in research, especially while studying translation of digital texts and interactions, the post-editing of online collaborative work, and the collection of digital data. These new technologies and software developments allow researchers (1) to collaborate and conduct research in areas of the world that deal with the protection of human subjects differently, (2) to look at corpora resulting from human interactions that were not initially collected for the purpose of research and sharing, and (3) to take part in a variety of fora which are open to the public. These three are just a few examples of issues that call for researchers to rethink their roles and responsibilities as they engage with various areas of the population in an unprecedented way.

### Sample Study 34.2

Mellinger, C. D., & Shreve, G. M. (2016). Match evaluation and over-editing in a translation memory environment. In R. Muñoz Martín (Ed.), *Reembedding translation process research* (pp. 131–148). Philadelphia and Amsterdam: John Benjamins.

#### Research Background

This translation process research study investigates professional translator behavior when using computer-assisted translation tools. Professional translators often use specialized software programs that can store their work in paired segments or units (source and target texts) called translation memories. These



tools allow the previously created translations to be reused and leveraged; however, this translation process is inherently different from unaided translation. Translation process research provides insight into the impact of these tools on translator behavior.

### Research Problems/Gaps

This study investigates how professional translators behave when presented with translation proposed by translation memory systems.

### Research Method

- *Type of research:* The study examines de-identified data derived from an experimental task conducted as part of a doctoral dissertation (Mellinger, 2014).
- *Setting and participants:* In the initial study, 9 Spanish-English translators with 4–7 years of professional experience completed an experimental task online.
- *Instruments/techniques:* Mellinger (2014) used TransCenter, a keystroke logging software, to present participants with a Spanish source text that had been segmented at the sentential level and to record their behavior. The participants either translated the Spanish source segment from scratch or edited a suggested English translation.
- *Data analysis:* The data were qualitatively analyzed to better understand professional translation and editing behavior when working with a translation memory.

### Key Results

From the qualitative analysis of experimental data, the researchers observe non-obligatory lexical and syntactic changes. The authors posit that this behavior is indicative of a mismatch between the participant's idea of what constitutes an adequate translation and the proposed translation, resulting in a tendency to overedit. Moreover, the preferential changes and persistence during the translation task are argued to be the result of a change in the translation paradigm when performing computer-assisted translation. The authors conclude with a call for additional research on editing behavior when interpreting translation process data.

### Comments

This study shows how translation process research allows for greater insight on translator behavior and how translation technologies influence the translation task. It also opens new areas of study by calling for research on editing and revision in the translation process.

## Resources for Further Reading

Angelelli, C. V., & Baer, B. J. (Eds.). (2015). *Researching translation and interpreting*. New York: Routledge.

This book offers a broad and systematic mapping of the research in the fields of TIS (including both spoken and sign languages). It explores the general features of a post-positivist approach to research in translation/

interpreting-oriented phenomena. It focuses on (1) the theoretical concepts (e.g., agency and role, bilingualism, cognitive processes, collaboration and volunteer work, fictional representation, gender, pedagogy, power and conflict, professionalism, identity and status, and reader response) framing research in TIS and explores these concepts by tracing how they have travelled from other disciplines and have been adopted or adapted in TIS and (2) the methodologies and methods used in T&I.

Saldanha, G., & O'Brien, S. (2014). *Research methodologies in translation studies*. New York: Routledge.

Focusing on empirical research in translation studies, this book offers a comprehensive review of current research methods used in translation. Studies are grouped into four categories: Studies oriented to product, process, participants, and contexts, and each of these categories constitutes a chapter. The final chapter offers insights to produce a research report by discussing structuring and framing the content as well as giving tips and suggestions on how to report the data.

## References

- Angelelli, C. V. (2004a). *Medical interpreting and cross-cultural communication*. London: Cambridge University Press.
- Angelelli, C. V. (2004b). *Re-visiting the interpreter's role*. Amsterdam/Philadelphia: John Benjamins.
- Angelelli, C. V. (2007). Validating professional standards and codes: Challenges and opportunities. *INTERPRETING: International Journal of Research and Practice in Interpreting*, 8, 175–193.
- Angelelli, C. V. (Guest Ed.). (2011). Translators and interpreters: Geographic displacement and linguistic consequences. Special issue of *The International Journal of the Sociology of Language*, 207.
- Angelelli, C. V. (2012). The sociological turn in translation and interpreting studies. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 7, 125–128.
- Angelelli, C. V. (2015). Justice for all? Issues faced by linguistic minorities and border patrol agents during interpreted arraignment interviews. *MonTi: Monografías de Traducción e Interpretación*, 7, 181–205.
- Angelelli, C. V. (2016). Looking back: A study of (ad-hoc) family interpreters. *European Journal of Applied Linguistics*, 4, 5–31.
- Angelelli, C. V., Agger-Gupta, N., Green, C. E., & Okahara, L. (2007). The California standards for healthcare interpreters: Ethical principles, protocols and

- guidance on roles and intervention. In C. Wadensjö, B. E. Dimitrova, & A.-L. Nilsson (Eds.), *The Critical Link 4* (pp. 167–177). Amsterdam and Philadelphia: John Benjamins.
- Angelelli, C. V., & Jacobson, H. E. (2009). *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice*. Amsterdam: John Benjamins Publishing.
- Antonini, R. (2011). Natural translator and interpreter. In Y. Gambier & L. Van Doorslaer (Eds.), *Handbook of translation studies* (Vol. 2, pp. 102–104). Amsterdam, Netherland: John Benjamins.
- Baker, M. (2006). *Translation and conflict. A narrative account*. New York and London: Routledge.
- Bandía, P. (2008). *Translation as reparation: Writing and translation in postcolonial Africa*. St. Jerome: Manchester.
- Bastin, G. (1998). *Traducir o Adaptar. Estudio de la adaptación puntual y global de obras didácticas*. Univesidad Central de Venezuela. Consejo de Desarrollo Científico y Humanístico.
- Berk-Seligson, S. (2002). *The bilingual courtroom: Court interpreters in the judicial process*. Chicago: Chicago University Press.
- Bolden, G. (2000). Towards understanding practices of medical interpreting: Interpreters' involvement in history taking. *Discourse Studies*, 2, 387–419.
- Chan, A. (2009). Perceived benefits of translator certification to stakeholders in the translation profession. A survey of vendor managers. *Across Languages and Cultures*, 11, 93–113.
- CHIA. (2002). *California standards for healthcare interpreters*. Retrieved July 2016, from [www.chiaonline.org](http://www.chiaonline.org)
- Colina, S., & Angelelli, C. V. (2015). Translation and interpreting pedagogy. In C. V. Angelelli & B. J. Baer (Eds.), *Researching translation and interpreting studies* (pp. 108–117). London and New York: Routledge.
- Davidson, B. (2000). The interpreter as an institutional gatekeeper: The sociolinguistic role of interpreters in Spanish-English medical discourse. *Journal of Sociolinguistics*, 4, 379–405.
- Dimitrova, B. (2005). *Expertise and explicitation in the translation process*. Amsterdam and Philadelphia: John Benjamins Publishing Company (Benjamins Translation Library, 64).
- Ervin, F., & Meyer, B. (Guest Eds.). (2016). Non-professional interpreting and translation: translational cultures in focus. Special issue of *European Journal of Applied Linguistics*, 4.
- Gavioli, L. (2012). Minimal responses in interpreter mediated medical talk. In C. Baraldi & L. Gavioli (Eds.), *Coordinating participation in dialogue interpreting* (pp. 201–228). Amsterdam and Philadelphia: John Benjamins.
- Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam and Philadelphia: John Benjamins.

- Gile, D. (2002). Conference and simultaneous interpreting. In M. Baker (Ed.), *Routledge Encyclopedia of translation studies* (pp. 40–45). London: Routledge.
- Hale, S. (2004). *The discourse of court interpreting*. Amsterdam and Philadelphia: John Benjamins.
- Hammond, D. L. (Ed.). (1994). *Professional issues for translators and interpreters* (Vol. VII, American Translators Association Monograph Series). Amsterdam and Philadelphia: John Benjamins Publishing Company.
- Hsieh, E. (2016). *Bilingual health communication: Working with interpreters in cross-cultural care*. New York: Routledge.
- Inghilleri, M. (Guest Ed.). (2005). Bourdieu and the sociology of translation and interpreting. Special issue of *The Translator*, 11.
- Jimenez-Crespo, M. (2015). Collaborative and volunteer translation and interpreting. In C. V. Angelelli & B. J. Baer (Eds.), *Researching translation and interpreting studies* (pp. 58–70). London and New York: Routledge.
- Karoly, A. (2014). Translation in foreign language teaching: A case study from a functional perspective. *Linguistics and Education*, 25, 90–107.
- Kuo, S., & Nakamura, M. (2005). Translation or transformation? A case study of language and ideology in the Taiwanese press. *Discourse and Society*, 16, 393–417.
- Li, D. (2000). Tailoring translation programs to social needs: A survey of professional translators. *Target. International Journal of Translation Studies*, 12, 127–149.
- Malakoff, M., & Hakuta, K. (1991). Translation skills and metalinguistic awareness in bilinguals. In E. Byalstock (Ed.), *Language and processing in bilingual children* (pp. 141–166). Cambridge: Cambridge University Press.
- Mellinger, C. D. (2014). *Computer-assisted translation: An empirical investigation of cognitive effort*. Unpublished Ph.D. dissertation, Kent State University, Kent, OH. Retrieved from <http://bit.ly/1ybBY7W>
- Metzger, M. (1999). *Sign language interpreting: Deconstructing the myth of neutrality*. Washington, DC: Gallaudet University Press.
- Meyer, B. (2012). Ad-hoc interpreting for Partially-Language-Proficient patients: Participation in multilingual constellations. In C. Baraldi & L. Gavioli (Eds.), *Coordinating participation in dialogue interpreting* (pp. 99–114). Amsterdam and Philadelphia: John Benjamins.
- Meylaerts, R. (2010). Habitus and self-image of native literary author-translators in diglossic societies. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 5, 1–19.
- Monzó, E. (2009). Legal and translational occupations in Spain. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 4, 135–154.
- Morris, R. (2010). Images of the court interpreter: Professional identity, role definition and self-image. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 5, 20–40.

- Muñoz Martin, R. (2013). More than a way with words: The interface between cognitive linguistics and cognitive translology. In A. Rojo & I. Ibrarretxe-Antuñano (Eds.), *Cognitive linguistics and translation: Advances in some theoretical models and applications* (pp. 75–96). Berlin: De Gruyter.
- Nicodemus, B., & Swabey, L. (2015). Action research. In C. V. Angelelli & B. J. Baer (Eds.), *Researching translation and interpreting studies* (pp. 157–167). London and New York: Routledge.
- Nord, C. (1997). *Translating as a purposeful activity: Functionalist approaches explained*. Manchester: St. Jerome.
- Roy, C. (2002). The problem with definitions, descriptions and the role of metaphors of interpreters. In F. Pöchhacker & M. Shlesinger (Eds.), *The interpreting studies reader* (pp. 344–353). London and New York: Routledge.
- Sela-Sheffy, R., & Shlesinger, M. (Eds.). (2011). *Identity and status in the translational professions* (Vol. 32). Amsterdam and Philadelphia: John Benjamins Publishing.
- Shreve, G. M. (1997). Cognition and the evolution of translation competence. In J. H. Danks, G. M. Shreeve, S. B. Fountain, & M. K. McBeath (Eds.), *Cognitive processes in translation and interpreting* (pp. 120–136). Thousand Oaks, CA: Sage.
- Tipton, R. (2010). On trust: Relationships of trust in interpreter-mediated social work encounters. In M. Baker, M. Olohan, & M. C. Pérez (Eds.), *Text and context: Essays on translation and interpreting in Honour of Ian Mason* (pp. 188–208). Manchester: St. Jerome.
- Tonkin, H., & Frank, M. E. (Eds.). (2010). *The translator as mediator of cultures* (Vol. 3). Amsterdam and Philadelphia: John Benjamins Publishing.
- Tyulenev, S. (2015). Agency and role. In C. V. Angelelli & B. J. Baer (Eds.), *Researching translation and interpreting studies* (pp. 17–31). London and New York: Routledge.
- Valdés, G., & Angelelli, C. V. (2003). Interpreters, interpreting and the study of bilingualism. *Annual Review of Applied Linguistics*, 23, 58–78.
- Valdés, G., Chavez, C., & Angelelli, C. V. (2000). Bilingualism from another perspective: The case of young interpreters from immigrant communities. In A. Roca (Ed.), *Research on Spanish in the United States: Linguistic issues and challenges* (pp. 42–81). Somerville, MA: Cascadilla Press.
- Valdés, G., & Figueroa, R. (1994). *Bilingualism and testing: A special case for bias*. Norwood, NJ: Ablex.
- Venuti, L. (1995). *The translator's invisibility- A history of translation*. London and New York: Routledge.
- Wadensjö, C. (1998). *Interpreting as interaction*. New York: Addison Wesley Longman Inc.
- Wolf, M., & Fukari, A. (Eds.). (2007). *Constructing a sociology of translation*. Amsterdam and Philadelphia: John Benjamins Publishing Company.



# 35

## Identity

Ron Darvin

### Introduction

Identity is a fundamental construct in applied linguistics as it is inextricably intertwined with language. Our sense of ourselves, our subjectivity, is constructed in language (Weedon, 1987), and at the same time, the performance of identities shapes how linguistic resources are deployed (De Fina, 2016). When people speak, they not only exchange information but also reorganize “a sense of who they are and how they relate to the world” (Norton, 2013, p. 4). Drawing on these poststructuralist perspectives, Norton (2013) defines identity as “the way a person understands his or her relationship to the world, how that relationship is constructed across time and space, and how the person understands possibilities for the future” (p. 4). Subject positions are continually negotiated through language (Baxter, 2016) and also with great ambivalence; hence identity, while central to applied linguistics, is also a very slippery term. Identity ascriptions and affiliations cover a wide range of categories: ethnic, racial, national, gender, social class, language, sexual, religious, age, lifestyle (Block, 2012). As a person moves across and participates in diverse spaces, his or her sense of self and relation to the world continually shift. In this perpetual state of flux, identity is dynamic, multiple, and often-times contradictory (Norton, 2013).

---

R. Darvin (✉)

Language & Literacy Education Department, University of British Columbia,  
Vancouver, BC, Canada

e-mail: [ron.darvin@ubc.ca](mailto:ron.darvin@ubc.ca)

Key to researching identity are the concepts of indexicality, positionality, and intersectionality. As language is used to perform personal identities or characteristics, individuals can adopt linguistic styles or expressions that may *index* geographical origin, gender, age, or class and demonstrate membership in specific social groupings or identity categories. When these individuals speak with an interlocutor, they bring to the particular situation their “history as a subjective being” (Davies & Harré, 1990, p. 48), as agents who are able to *position* themselves (Bucholtz & Hall, 2005). What one says in interaction, however, also positions others, making them subjects of this discursive process. This positionality draws attention to how language as a social practice is subject to conditions of power and how identity is a site of struggle as people seek to claim a legitimate position within these conditions. Intersectionality, on the other hand, recognizes that while people can be positioned because of an inscription of their identity (e.g., being “Black,” “woman,” or “gay”), these dimensions actually interrelate or *intersect*. The lived experiences of a working-class Black lesbian will be markedly different from those of a middle-class White heterosexual man, and identity researchers need to recognize that there are variations and inequalities within these dimensions (Block & Corona, 2014).

In the past two decades, there has been a surge of identity research in applied linguistics, with monographs in abundance (e.g., Block, 2014; Clark, 2009; De Costa, 2016; Higgins, 2011; Kamada, 2010; Kramsch, 2009; Menard-Warwick, 2009; Mercer & Williams, 2014; Norton, 2013). The 2015 issue of *Annual Review of Applied Linguistics* edited by Alison Mackey (2015) was dedicated to identity, and assembled a variety of articles that discussed this construct as it intersected with translanguaging (Creese & Blackledge), investment (Darvin & Norton), transnationalism (Duff), written discourse (Matsuda), migrants and migration (Wodak & Boukala), and many others. Preece’s (2016) *Routledge Handbook of Language and Identity* brought together various perspectives informing language and identity studies, categories and dimensions of identity, and the key issues that surround them. These issues are also addressed in the two sample studies that appear at the end of this chapter.

## Current and Core Issues

While there is a wide range of terms for identity, including “self,” “position,” “role,” “subjectivity,” “subject,” and “agent,” most recent research on identity has been consistently marked by a social constructionist paradigm that pays attention to the micro-level of interaction and meaning making. The contexts



of identity construction are many: from everyday conversation to institutional settings, commodified contexts, and digital environments (Benwell & Stokoe, 2016), and these provide the backdrop for diverse approaches. In the past three decades, there has been much debate about how best to describe and interpret the relationship between language and identity. This variety of perspectives has paved the way for a diversity of applied linguistic approaches that range from the ethnomethodological (Benwell & Stokoe, 2016; Harklau, 2005) to the critical (Norton, 2013; Pennycook, 2001). Researchers of identity adopt, modify, and design methodologies that fit their specific purposes, and this section discusses some of the most popular methodologies and their corresponding issues.

## Case Study

As a research approach, case study has generally provided a contextualized profile of a person (e.g., a language learner or teacher) or a group (e.g., a class or a community of practice). Case studies analyze linguistic, cultural, and social phenomena to understand individuals' experiences and development within specific contexts. Longitudinal case-based approaches to analyzing L2 development, for instance, reveal how data based on larger numbers of learners can obscure individual variation over time and variation of developmental patterns and processes across individuals (e.g., Larsen-Freeman, 2006). There are some case studies, as part of mixed-methods research designs, that analyze both quantitatively measured linguistic dimensions of learning/use and qualitatively described sociocultural dimensions (e.g., Duff et al., 2013; Kinginger, 2008). A great number of recently published case studies in applied linguistics though have foregrounded personal and sociocultural aspects of lived experiences without detailed linguistic descriptions (see Darvin, 2017 for an example). This emerging body of research examines the shifting identities of language learners of different life stages, using diverse repertoires, while constrained by norms, ideologies, and policies (Duff, 2014).

While the case study is able to provide a better understanding of the context and processes of a phenomenon, it has often been caught in a methodological crossfire because of a number of issues. These include questions of what one can generalize on the basis of an individual case and how such a case contains a bias toward verification or a tendency to confirm the researcher's preconceived ideas (Flyvbjerg, 2011). Flyvbjerg (2011) challenges these "misunderstandings" by pointing out that even scientific conclusions of Newton, Einstein, and Darwin were based on carefully chosen cases and that



the case study contains no greater bias than other methods of inquiry that also involve subjective judgments of the researcher. The number of cases and the length of time a study is conducted may be debatable, but Duff (2014) argues these decisions can be made based on the research topic and the number of candidates who agree to participate. By making sound methodological decisions, identity researchers, through the case study, are able to cast a light on the complexity and contradictions of individuals and to find generalizable connections.

## Narrative Inquiry

Canagarajah (1996) asserts that narratives are able to represent knowledge “from the bottom up” (p. 327) and have the powerful potential to represent in a comprehensive and more open-ended way the identities of research participants. As a systematic study and interpretation of stories of life experiences, narrative inquiry has become a popular research approach in applied linguistics (Barkhuizen, 2008, 2010; Bell, 2002; Pavlenko, 2002, 2007). One specific approach to narrative inquiry is highlighting the “identity work” research participants engage in “as they construct selves within specific institutional, organizational, discursive and local cultural contexts” (Chase, 2005, p. 658). In a study that examined researcher identity, Norton and Early (2011) draw on Bamberg’s (2006) conception of big stories that focus on life histories and small stories that are centered on small talk and regular conversation. They reflect that while small stories-in-interaction may not create a coherent sense of self, they can highlight the diverse identities that one performs in everyday interactive practices.

One of the challenges of narrative researchers is to understand what is “story worthy” in the participant’s social setting and to orient to the particularity of his or her story and voice (Chase, 2005). Pavlenko (2007) identifies a number of issues of narrative research that are linked to content and thematic analysis. The lack of both a theoretical premise and a set of established procedures, she argues, obscures how conceptual categories are identified and linked to specific instances. The focus on what recurs in the text can also lead to the oversight of important events that need not have happened repeatedly and of the gaps and absences in a text that are themselves meaningful. The most problematic issue however, Pavlenko points out, is the lack of attention to how narratives are constructed through carefully chosen language. In this case, storytellers use words that position them in specific ways and reflect their own interpretive stances.

## Conversational Analysis

Benwell and Stokoe (2016) assert that while conversational analysts don't usually begin with identity as a theoretical tool, they recognize that social life is grounded in interaction and constituted by talk. Identities are made relevant by conversational participants and are constructed in sequence. If ethnomethodologists and conversational analysts were to adopt a theory on identity therefore, it would be an "indexical, context bound theory, in which identity is understood as an oriented-to, recipient-designed accomplishment of interaction" (p. 68). Research thus has to begin with an analysis of identity categories that is not based on what analysts take to be relevant but on what people do and say as they deploy categories. CA collects naturally occurring data where participants proceed with their daily routines. This means, however, that there is low standardization and comparability on the level of the interaction (Kasper & Wagner, 2014).

In a study of how the identity of "ESL student" is constructed in routine classroom interactions between students and the teacher, Talmy (2009) demonstrated how CA can be productive in critical research on second language education. By grounding theories of language socialization and cultural production in the participants' discursive practices, he asserts that CA can provide an analytic frame that shows how power is achieved in interaction before any a priori conclusion. By examining language at work, the researcher of identity can investigate how biases and assumptions of race, gender, class, and sexual orientation are "instantiated, resisted, accommodated, reproduced, and/or transformed" (p. 206) in daily routines and activities. For Kasper and Wagner (2014), one of the greatest advantages of CA is that it offers researchers a coherent, integrated theory and methodology of interactions to investigate language-related real-life problems. Longitudinal CA research, for instance, helps examine language learning of individuals over time and traces changes in their use of lexical and grammatical forms. Social problem-oriented CA examines how power imbalances are constructed through interactional practices in various settings like courtrooms or hospitals, positioning speakers by virtue of their situated identities. Kasper and Wagner point out that while CA in its earlier stages was limited by an entrenched monolingualism, more recent research has rectified this through studies that focus on the practices of multilingual interaction (He, 2013; Higgins, 2009; Neville & Wagner, 2011; Li, 2011).

## Critical Discourse Analysis

Predicated on the idea that discursive and social processes are woven together, critical discourse analysis (CDA) offers identity researchers a framework to examine how identity constructions are circumscribed by power relations and ideology. By critiquing unequal social practices, CDA demonstrates how the social positions of individuals allow them differential access to linguistic, cultural, and economic capital and varying degrees of recognition. It draws attention to contextual constraints and affordances that influence the projected and ascribed identities of people (Zotzmann & O'Regan, 2016). Cheng (2013) points out that a great number of CDA studies in the past two decades have examined media and political discourses and that CDA would benefit from exploring more diverse text types. With the rise of digital media, there has also been a greater need for multimodality analytical tools that take into account new genres and formats to examine super-diverse and super-vernacular texts (Lin, 2014).

Recognizing that discourses about social groups are reproduced by structures of power, Escamilla (2013) uses CDA to examine how Japanese mainstream news media constructs the identity of *gaikokujin* or resident foreign nationals, particularly ethnically Japanese Brazilians who back-migrated to Japan. By drawing from a corpus of 36 articles from regional and national news platforms, he analyzes how lexical choices contribute to the othering and deauthentication of these foreign nationals. In this case, CDA enables the examination of how public discourse can construct minority identity and reinforce dominant cultural assumptions (Belcher & Nelson, 2013). Because identity is itself a discursive phenomenon, CDA uses language as a way to understand how identities are formed, represented, and enacted in unequal ways. Although it does not have a homogeneous theoretical framework or a set of fixed methodological tools, its commitment to a critique of inequitable social practices is geared toward contributing to greater transformative action (Zotzmann & O'Regan, 2016).

## Challenges and Controversial Issues

One fundamental challenge in identity research is that no matter how meticulously it is executed and articulated across spatiotemporal scales, it will always be partial (Block, 2010). Researchers will need to continually grapple with how many interviews, stories, and artifacts can sufficiently represent an individual's identity. Another issue is that identity is not only language mediated

but semiotically mediated. Blommaert (2005) defines identity as “particular forms of semiotic potential, organised in a repertoire” (p. 207), and so any analysis that is limited to the linguistic will be partial and will miss other semiotic resources that comprise communication: gaze, posture, gestures, dress, and so on. (Block, 2010). Through the affordances of the digital, individuals have discovered new, constantly evolving, multimodal means of self-representation, shaping new identities and imaginations (Darvin, 2016, 2018). These affordances complexify self-identifications especially as people textualize themselves in social media to project a self of their choosing, while anticipating comments and reactions (Davies & Merchant, 2009; Stornaiuolo, Higgs, & Hull, 2013; Weber & Mitchell, 2008). How researchers are able to integrate the semiotic affordances of the visual, aural, spatial, and gestural, in their analysis of identities thus becomes increasingly significant in the digital age.

In digital environments, methodological approaches to identity often involve graphically rendered texts, interactions in communities of interest, and multiple layers of data from participant interviews (Thorne, Sauro, & Smith, 2015). Thorne and Black (2011) outline three interrelated dynamics within digitally mediated environments that help construct functional selves:

indexical linkages to macro-level categories (such as nation state affiliation, cultural/linguistic/ethnic affiliations), 2) functionally defined subject positions (such as student, youth, author, editor, expert, and novice, among others), and 3) fluid shifts in language choice, stance, and style that enable participants to personalize, make relevant, and move forward a variety of social actions. (p. 259)

To examine these dynamics, Stornaiuolo, Higgs, and Hull (2013) propose a mix of qualitative and quantitative data in studies of learners’ authoring process across online/offline spaces, multiple languages, and semiotic systems. Because authorship has become more distributed and interactive, they argue that researchers need to understand the networked logics of this process to guide their methodological practices. Because the abundance of sites leads to a multiplicity of data and because networked spaces are hybrid spaces, Stornaiuolo and Hall (2014) posit that methodological approaches have to be multidimensional as well. Tracing the movements of people, texts, and ideas in cross-contextual meaning is a methodological challenge in networked contexts, and this asserts the need to trace “resonances” or the “intertextual echoing of ideas across spaces, people, and texts” (p. 28).

### Sample Study 35.1

Norton Peirce, B. (1995). Social identity, investment, and language learning. *TESOL Quarterly*, 29, 9–31.

#### Research Background

This classic study was conducted at a time when conceptions of the good language learner assumed idealized and homogenous conditions of learning and understood learners dichotomously as simply being motivated/unmotivated or introverted/extroverted. To address this gap, Norton conducted this research to examine how language learning was a socially and historically constructed process and how learners engage differently with their learning in various contexts.

#### Research Problems/Gaps

The research questions were divided into two parts:

1. "How are the opportunities for immigrant women in Canada to practice ESL socially structured outside the classroom? How do immigrant women respond to and act upon these social structures to create, use or resist opportunities to practice English? To what extent should their actions be understood with reference to their investment in English and their changing social identities across time and space?" (p. 13)
2. "How can an enhanced understanding of natural learning and social identity inform SLA theory, in general, as well as ESL pedagogy for immigrant women in Canada?" (p. 14)

#### Research Method

- *Type of research*: Longitudinal case study.
- *Setting and participants*: Ontario, Canada, in 1991. Five women who recently immigrated from Vietnam, Poland, the former Czechoslovakia, and Peru.
- *Instruments/techniques*: Diary study, interviews, participant observation, pre-study and post-study questionnaires.
- *Data analysis*: Organization of data according to each participant, cross-collation based on different sites of language use (home, workplace, school), analysis of each participant with regard to their social position, categorization of the data according to the production of gender and ethnicity.

#### Key Results

Through this study, Norton conceptualized a theory of identity in language learning that foregrounds the conditions of power in contexts where learning takes place. Drawing on poststructuralist notions of subjectivity, she asserts that identity is multiple and always a site of struggle, as learners negotiate asymmetrical relations of power. How learners are able to invest in their learning is shaped by how they position themselves and are positioned by target language speakers.

#### Comments

When Norton introduced the constructs of identity and investment, her research was considered pathbreaking, a significant component of what has become the sociocultural turn of language learning research. With its in-depth analysis of the language learning experiences of five women, this research has demonstrated how the number of participants in a case study can be limited but still be able to draw generalizable and productive conclusions. The case study in identity research has the powerful capacity to demonstrate the complexities and contradictions of individuals.

## Future Directions

As the digital becomes more embedded in everyday life, researchers are confronted with a more complex array of potential sites and sources of data. Tying together disparate insights from diverse digital practices to understand identity production becomes more complex. Hine (2015) talks about the need for an “ethnography of an embedded, embodied, everyday Internet” (p. 56) where researchers are able to move between face-to-face to mediated forms of interaction, challenging the notion of conventionally bounded field sites and allowing a multi-sited field to emerge. This connective, itinerant, or networked ethnography requires an openness to exploring connections as they present themselves. Leander (2008) defines connective ethnography as a methodological approach “that considers connections and relations as normative social practices and Internet social spaces as connected to other social spaces” (p. 37). Recognizing the connectedness of these spaces, Büscher and Urry (2009) have devised mobile methods that allow researchers to examine these online and offline practices through the frame of movement. In this case, data collection takes into account the bodily travel of people and virtual travel across networks of mediated communications, as people are connected in interactions face-to-face and via mediated communications.

Hine (2015) points out that a number of issues arise when a field site extends to the online realm. While ethnographers emphasize immersion in a setting as a means of knowledge generation, an ethnography of the internet raises questions about how to define prolonged immersion and how to determine the boundaries of limitless online space. New issues of privacy and confidentiality also arise when researchers gain access to the social media accounts of participants. Researchers become perpetually present, and participants need to decide which social media activities they want to provide access to (Baker, 2013; Eynon, Fry, & Schroeder, 2008). Not only does this require continual negotiation through informed consent but it also involves participants actively managing their privacy settings. At the same time, for ethnographers to access and observe social media activities of participants in sites like Twitter or Facebook, they need to make themselves present by adding or friending them. Researchers become more visible, and this requires them to make informed decisions about how to manage their own online identities (for a more detailed discussion on the ethics of online research, see Ethics of Online Research Methods in Applied Linguistics (special issue). 2016. *Applied Linguistics Review*).

As participants chat with other online users, Leander (2008) also raises questions regarding the nature of these interactions. In Lam’s (2004) study of Chinese immigrant youth and their online literacies, for instance, the researcher

regularly browsed web pages that participants were constructing, used screen capture to document their online activities, and recorded dialogues in the forums of chat rooms. Because she did not have consent from all the interlocutors of these chat rooms, the key ethical issue that arises concerns the extent to which these interactions could be regarded as data. Another issue is when research projects involve the production of identity texts that become permanent and perpetually present artifacts. When they are made available online, multimedia self-presentations of participants can also fix representations of identity and influence their lives in complex, consequential ways (Nelson, Hull, & Roche-Smith, 2008).

While the researcher uses various strategies to produce data and record aspects of research settings, a reflexive attention to the choices made is key to a richer analysis. While it may not lead to clear answers, its “muddiness and ambiguity,” Hine (2015) argues, enable an autoethnographic stance that delivers an account from the self that has navigated the online terrain. While there has been criticism of autoethnography as self-indulgent and narcissistic, it can offer an embodied and emotional account of media engagement, where the researcher is able to make sense of the influences that shape and constrain online practices. For Marsh (2013), auto-cyberethnography enables researchers not only to describe their own experiences in online environments but also investigate the development of their digital practices over time. This perspective allows the individualized nature of digital engagement and identity construction to be explored with greater depth, compared to observing participants and asking them for retrospective accounts.

### Sample Study 35.2

Lee, C. (2014). Language choice and self-presentation in social media: The case of university students in Hong Kong. In P. Seargeant & C. Tagg (Eds.), *Language of social media* (pp. 91–111). New York: Palgrave Macmillan.

#### Research Background

Recognizing how identities shift across time and space, this study focuses on the ways bilingual undergraduate students in Hong Kong used multiple linguistic resources on social media sites like Facebook to perform multiple identities. Acknowledging how the technological affordances of social media provide a wide range of multimodal references, the researcher was interested in how the construction of profiles in particular contributes to a “presentational” culture where the “self” is created for multiple audiences.

#### Research Problems/Gaps

How do social media spaces allow users to project or extend existing aspects of identity? What is the relationship between identity performance and language choice in social media?

### Research Method

- *Type of research*: Techno-biography, defined as “a life story in relation to technologies” (p. 94).
- *Setting and participants*: Twenty participants, who shared a similar linguistic repertoire of speaking Cantonese fluently, having knowledge of standard written Chinese, and English.
- *Instruments/techniques*: Techno-biographic interviews, where the researcher asked participants about their current technology practices, ways of participation, routines, history of using technology in different phases and domains of life, and cross-generational comparisons. The study also involved the observation of their online language use and an online questionnaire survey.
- *Data analysis*: Using field notes and data collected during the two phases, a profile for each participant was created, and transcripts were coded using ATLAS.ti to identify emerging patterns. By analyzing the data, the researcher then examined how participants constructed their identities through various social media activities such as creating an online profile, providing status updates, and posting photos.

### Key Results

One’s existing and imagined identities shape language choices in social media. Perceived audiences and purposes, together with lived experiences, produce a great diversity of digital practices, debunking the notion of a homogenous digital native generation. Through a techno-biographic approach, the researcher is able to examine more closely the situated nature of social media and how aspects of one’s identity are developed over time.

### Comments

To construct a techno-biography, Lee describes her method of data collection as adopting “an ethnographic style” (p. 96) as it pays attention to details about people’s online lives. As research of digital practices continues to evolve, the parameters of Hine’s “ethnography of the internet” are still to be defined. In this study, each interview began with around 30 minutes of the participants’ online activities being recorded using Camtasia, followed by each participant describing what they were doing in the screen recording. In this sense, the method is like an auto-cyberethnography that offers an embodied and emotional account of media engagement, through what is also a narrative, a story of one’s digital life.

## Resources for Further Reading

De Costa, P. I. (2016). *The power of identity and ideology in language learning: Designer immigrants learning English in Singapore*. New York: Springer.

Apart from providing a theoretically rich analysis of how language learning can be reconceptualized in the age of globalization, this book discusses the critical ethnographic case study as a powerful method for identity research, and the ethical issues that circumscribe this approach.



Mackey, A. (Ed.). (2015). Identity. *Annual Review of Applied Linguistics*, 35.

This issue brings together prominent and emerging identity scholars who engage with diverse research methods and demonstrate how identity continues to be a powerful construct in applied linguistics.

Norton, B. (2013). *Identity and language learning: Extending the conversation* (2nd ed.). Bristol: Multilingual Matters.

Republished with an introduction that discusses methodological innovations and future directions, this canonical text is required reading for those who intend to pursue identity research.

Preece, S. (Ed.). (2016). *The Routledge handbook of language and identity*. Oxon: Routledge.

This comprehensive volume provides various theoretical perspectives and methodologies relevant to identity research, including conversation analysis, linguistic ethnography, and critical discourse analysis.

## References

- Baker, S. (2013). Conceptualising the use of Facebook in ethnographic research: As tool, as data and as context. *Ethnography and Education*, 8, 131–145.
- Bamberg, M. (2006). Stories: Big or small: Why do we care? *Narrative Inquiry*, 16, 139–147.
- Barkhuizen, G. (2008). A narrative approach to exploring context in language teaching. *English Language Teaching Journal*, 62, 231–239.
- Barkhuizen, G. (2010). An extended positioning analysis of a pre-service teacher's better life small story. *Applied Linguistics*, 31, 282–300.
- Baxter, J. (2016). Positioning language and identity. In S. Preece (Ed.), *The Routledge handbook of language and identity* (pp. 34–49). Oxon: Routledge.
- Belcher, D., & Nelson, G. (2013). Why intercultural rhetoric needs critical and corpus-based approaches: An introduction. In D. Belcher & G. Nelson (Eds.), *Critical and corpus-based approaches to intercultural rhetoric* (pp. 1–6). Ann Arbor: University of Michigan Press.
- Bell, J. S. (2002). Narrative inquiry: More than just telling stories. *TESOL Quarterly*, 36, 207–213.
- Benwell, B., & Stokoe, E. (2016). Ethnomethodological and conversational analytic approaches to identity. In S. Preece (Ed.), *Routledge handbook of language and identity* (pp. 66–82). Oxon: Routledge.

- Block, D. (2010). Researching language and identity. In B. Paltridge & A. Phakti (Eds.), *Continuum companion to research methods in applied linguistics* (pp. 337–349). London: Continuum.
- Block, D. (2012). Class and SLA: Making connections. *Language Teaching Research*, 16(2), 188–205.
- Block, D. (2014). *Social class and applied linguistics*. Oxon: Routledge.
- Block, D., & Corona, V. (2014). Exploring class-based intersectionality. *Language, Culture and Curriculum*, 27, 27–42.
- Blommaert, J. (2005). *Discourse: A critical introduction*. Cambridge, UK: Cambridge University Press.
- Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7, 585–614.
- Büscher, M., & Urry, J. (2009). Mobile methods and the empirical. *European Journal of Social Theory*, 12(1), 99–116.
- Canagarajah, S. (1996). From critical research practice to critical research reporting. *TESOL Quarterly*, 30, 321–331.
- Chase, S. (2005). Narrative inquiry: Multiple lenses, approaches, voices. In N. Denzin & Y. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 651–680). Thousand Oaks, CA: Sage.
- Cheng, W. (2013). Corpus-based linguistic approaches to critical discourse analysis. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Chichester: John Wiley.
- Clark, J. B. (2009). *Multilingualism, citizenship, and identity: Voices of youth and symbolic investments in an urban, globalized world*. London: Bloomsbury.
- Darvin, R. (2016). Language and identity in the digital age. In S. Preece (Ed.), *Routledge handbook of language and identity* (pp. 523–540). Oxon: Routledge.
- Darvin, R. (2017). Social class and the inequality of English speakers in a globalized world. *Journal of English as a Lingua Franca*, 6(2), 287–211.
- Darvin, R. (2018). Social class and the unequal digital literacies of youth. *Language and Literacy*, 20(3), 26–45.
- Davies, B., & Harré, R. (1990). Positioning: The discursive production of selves. *Journal for the Theory of Social Behaviour*, 20, 43–63.
- Davies, J., & Merchant, G. (2009). *Web 2.0 for schools: Learning and social participation*. New York: Peter Lang.
- De Costa, P. I. (2016). *The power of identity and ideology in language learning: Designer immigrants learning English in Singapore*. Switzerland: Springer.
- De Fina, A. (2016). Linguistic practices and transnational identities. In S. Preece (Ed.), *The Routledge handbook of language and identity* (pp. 163–178). Oxon: Routledge.
- Duff, P., Anderson, T., Ilnyckyj, R., Van Gaya, E., Wang, R., & Yates, E. (2013). *Learning Chinese: Linguistic, sociocultural, and narrative perspectives*. Berlin, Germany: De Gruyter.
- Duff, P. A. (2014). Case study research on language learning and use. *Annual Review of Applied Linguistics*, 34, 233–255.

- Escamilla, R. (2013). Discriminatory discursive strategies used by the Japanese mainstream news media in constructing the identity of resident foreign nationals: A critical discourse analysis-based examination. In D. Belcher & G. Nelson (Eds.), *Critical and corpus-based approaches to intercultural rhetoric* (pp. 72–96). Ann Arbor: University of Michigan Press.
- Eynon, R., Fry, J., & Schroeder, R. (2008). The ethics of internet research. In N. Fielding, R. Lee, & G. Blank (Eds.), *The Sage handbook of online research methods* (pp. 23–41). Thousand Oaks, CA: Sage.
- Flyvbjerg, B. (2011). Case study. In N. Denzin & Y. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 301–316). Thousand Oaks, CA: Sage.
- Harklau, L. (2005). Ethnography and ethnographic research on second language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 179–193). Mahwah, NJ: Lawrence Erlbaum.
- He, A. W. (2013). The wor(l)d is a collage: Multi-performance by Chinese heritage language speakers. *Modern Language Journal*, 97, 304–317.
- Higgins, C. (2009). “Are you Hindu?” Resisting membership categorization through language alternation. In H. Nguyen & G. Kasper (Eds.), *Talk-in-interaction: Multilingual perspectives* (pp. 111–136). Honolulu: University of Hawai’i.
- Higgins, C. (Ed.). (2011). *Identity formation in globalizing contexts: Language learning in the new millennium*. Berlin: Walter de Gruyter.
- Hine, C. (2015). *Ethnography for the internet: Embedded, embodied and everyday*. London: Bloomsbury.
- Kamada, L. (2010). *Hybrid identities and adolescent girls*. Bristol, UK: Multilingual Matters.
- Kasper, G., & Wagner, J. (2014). Conversation analysis in applied linguistics. *Annual Review of Applied Linguistics*, 34, 171–212.
- Kinginger, C. (2008). Language learning in study abroad: Case studies of Americans in France. *Modern Language Journal*, 92, 1–124.
- Kramsch, C. J. (2009). *The multilingual subject: What foreign language learners say about their experience and why it matters*. Oxford: Oxford University Press.
- Lam, W. S. E. (2004). Second language socialization in a bilingual chat room: Global and local considerations. *Language Learning & Technology*, 8, 44–65.
- Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27, 590–619.
- Leander, K. (2008). Toward a connective ethnography of online/offline literacy networks. In J. Coiro, M. Knobel, C. Lankshear, & D. Leu (Eds.), *Handbook of research on new literacies* (pp. 33–65). New York: Lawrence Erlbaum.
- Li, W. (2011). Multilinguality, multimodality, and multicompetence: Code- and modeswitching by minority ethnic children in complementary schools. *Modern Language Journal*, 95, 370–384.
- Lin, A. (2014). Critical discourse analysis in applied linguistics: A methodological review. *Annual Review of Applied Linguistics*, 34, 213–232.
- Mackey, A. (Ed.). (2015). Identity. *Annual Review of Applied Linguistics*, 35.

- Marsh, J. (2013). Researching young children's literacy practices in online virtual worlds: Cyber-ethnography and multi-method approaches. In P. Albers, T. Holbrook, & A. S. Flint (Eds.), *New literacy research methods* (pp. 195–209). New York: Routledge.
- Menard-Warwick, J. (2009). *Gendered identities and immigrant language learning*. Bristol, UK: Multilingual Matters.
- Mercer, S., & Williams, M. (Eds.). (2014). *Multiple perspectives on the self in SLA*. Bristol, UK: Multilingual Matters.
- Nelson, M. E., Hull, G. A., & Roche-Smith, J. (2008). Challenges of multimedia self-presentation taking, and mistaking the show on the road. *Written Communication, 25*, 415–440.
- Neville, M., & Wagner, J. (2011). Language choice and participation: Two practices for switching languages in institutional interaction. In G. Pallotti & J. Wagner (Eds.), *L2 learning as social practice: Conversation-analytic perspectives* (pp. 211–235). Honolulu: University of Hawai'i.
- Norton, B. (2013). *Identity and language learning: Extending the conversation* (2nd ed.). Bristol, UK: Multilingual Matters.
- Norton, B., & Early, M. (2011). Researcher identity, narrative inquiry, and language teaching research. *TESOL Quarterly, 45*, 415–439.
- Pavlenko, A. (2002). Narrative study: Whose story is it, anyway? *TESOL Quarterly, 36*, 213–218.
- Pavlenko, A. (2007). Autobiographic narratives as data in applied linguistics. *Applied Linguistics, 28*, 163–188.
- Pennycook, A. (2001). *Critical applied linguistics: A critical introduction*. Mahwah, NJ: Lawrence Erlbaum.
- Stornaiuolo, A., Higgs, J., & Hull, G. (2013). Social media as authorship: Methods for studying literacies and communities online. In P. Albers, T. Holbrook, & A. S. Flint (Eds.), *New literacy research methods* (pp. 224–237). New York: Routledge.
- Stornaiuolo, A., & Hall, M. (2014). Tracing resonance: Qualitative research in a networked world. In G. Gudmundsdottir & K. Vasbø (Eds.), *Methodological challenges when exploring digital learning spaces in education* (pp. 29–43). Rotterdam: Sense.
- Talmy, S. (2009). Resisting ESL: Categories and sequence in a critically “motivated” Analysis of classroom interaction. In H. Nguyen & G. Kasper (Eds.), *Talk-in interaction: Multilingual perspectives* (pp. 181–213). Honolulu: University of Hawai'i.
- Thorne, S., Sauro, S., & Smith, B. (2015). Technologies, identities, and expressive activity. *Annual Review of Applied Linguistics, 35*, 215–233.
- Thorne, S. L., & Black, R. W. (2011). Identity and interaction in Internet-mediated contexts. In C. Higgins (Ed.), *Identity formation in globalizing contexts: Language learning in the new millennium* (pp. 257–278). Berlin: de Gruyter.

- Weber, S., & Mitchell, C. (2008). Imaging, keyboarding, and posting identities: Young people and new media technologies. In D. Buckingham (Ed.), *Youth, identity, and digital media* (pp. 25–47). Cambridge, MA: MIT Press.
- Weedon, C. (1987). *Feminist practice and poststructuralist theory*. Oxford, UK: Basil Blackwell.
- Zotzmann, K., & O'Regan, J. (2016). Critical discourse analysis and identity. In S. Preece (Ed.), *Routledge handbook of language and identity* (pp. 113–127). Oxon: Routledge.



# 36

## Gesture Research

Gale Stam and Kimberly Buescher

### Introduction

Research in second language (L2) acquisition examining the speech and gesture of L2 learners and teachers has shown that both of their gestures are important in L2 teaching and learning (see Stam, 2013, for a review). In this chapter we provide an overview of this research, discuss the various research designs and methods used and their challenges, and present suggestions for future research.

Although the teaching of kinesics and cultural emblems in the L2 classroom has been advocated since the 1980s (e.g., Pennycook, 1985; von Raffler, 1980; Wylie, 1985) on the grounds that being competent in an L2 involves the use of the entire body, it was not until the 1990s that gestures in L2 acquisition and teaching were studied empirically. These studies grew out of the groundbreaking work of David McNeill (1985, 1992) and Adam Kendon (1972, 1980), who recognized that language involved more than speech alone. According to McNeill and Kendon, language includes co-speech gestures—the spontaneous movements of the hands that accompany speech—and these gestures need to be taken into account when individuals are speaking their first language (L1) or their L2.

---

G. Stam (✉)

National Louis University, Skokie, IL, USA

e-mail: [gstam@nl.edu](mailto:gstam@nl.edu)

K. Buescher

University of Massachusetts - Boston, Boston, MA, USA

e-mail: [kimberly.buescher@umb.edu](mailto:kimberly.buescher@umb.edu)

Kendon and McNeill came to their realizations about speech and gesture from different backgrounds. Kendon, an anthropologist, while doing a frame-by-frame analysis of discourse from an interaction perspective, realized that gesture and speech patterned similarly and that gestures were not haphazard, but organized into hierarchical units. These units (Kendon, 1980; McNeill, 1992) include the gesture phrase (the entire movement from start to finish) and gesture phases (the smaller movements that make up the gesture phrase), and an understanding of them is necessary when conducting or evaluating any type of research that involves gesture especially in L2 acquisition and teaching.

Gesture phases consist of the *preparation* (the movement of the hand to a position for the stroke to be executed, optional), the *stroke* (the part of the movement that expresses meaning, obligatory), *holds* (both pre-stroke and post-stroke maintaining of the position, optional), and *retraction* or *return* (movement back to rest position, optional).

McNeill, a psycholinguist, realizing that speech and gesture were related while giving a speech at a conference in France and the translator was both speaking and gesturing, has focused on the relationship between gesture, speech, and thought. He claims that gesture provides us an enhanced window onto the mind (1992). The model he proposes for verbal thought (McNeill, 2005) is “an interactive model..., an ‘imagery-language dialectic’ in which thought, language, and gesture develop over time and influence each other and in which the static and dynamic aspects of language are combined” (Stam, 2013, p. 4).

## Types of Gestures

It is important to note that co-speech gestures are one type of gesture and that gestures can be differentiated based on their relationship with speech (McNeill, 1992). These different types of gestures have ramifications for how research is conducted in L2 acquisition and teaching.

### Co-speech Gestures

Co-speech gestures are synchronous with and only occur with speech. They perform similar pragmatic functions as speech and tend to occur with elements that are high in communicative dynamism (McNeill, 1992): new, contrastive, or focused information. These gestures can be categorized according to their semiotic properties (Stam, 2013): their degree of iconicity (concreteness), metaphoricity (abstractness), deixis (concrete and abstract pointing in space and time), temporal highlighting (beats—quick movements of the

hands that accompany repairs, prosody, introduction of new material), and social interactivity (addressing an interlocutor).

As Stam and McCafferty (2008) have pointed out, co-speech gestures are multifunctional. Some of these functions are for self (e.g., lexical searches and retrievals, lightening of cognitive load, and organization of spatial information), while others are for the interlocutor (e.g., providing information that is not present in speech but is in the speaker's mind, retaining turns, and indicating involvement in the conversation).

## Speech-Linked Gestures

Speech-linked gestures, like co-speech gestures, occur with speech. However, their timing is different. They are not synchronous with speech. Rather they occur with speech pauses and fill a gap in the sentence. For example, he went [gesture showing someone climbing].

## Emblems

Emblems are culturally specific gestures that can occur with or without speech. They are conventionalized gestures that are known to members of a cultural group, and every culture has them. Many times the same form exists in several cultures, but the meaning is different. Examples of emblems are the okay sign or the thumbs up sign in English. Because emblems are conventionalized and learned gestures, they can be and often are taught in language classrooms.

## Pantomimes

Pantomimes occur without speech. They involve the use of the entire body to depict actions, objects, or entire stories. Pantomimes are the type of gestures often used by young children in telling a story, by mimes such as Marcel Marceau, or in word guessing games where speech is forbidden, such as charades.

## Signs

Signs occur without speech. They are codified movements of the hands that have linguistic properties (morphology, syntax, phonology) and form the basis of sign languages such as American Sign Language or alternate sign languages that develop under specific circumstances where speaking is not possible or prohibited (see Stam, 2013).



## SLA and Gesture Studies: Core Issues

Second language and gesture studies have been conducted both in and out of the classroom. They have examined teachers' gestures as well as learners' gestures using naturalistic and controlled experimental data. The main topics they have investigated include classroom interaction, the function and perception of teachers' gestures, emblems, thinking for speaking, referent marking, bilingual development, the use of gestures to facilitate learning, and learners' gestures and their functions—for self and for other (see Stam, 2013).

Studies on emblems, culturally specific codified gestures, in second language acquisition (SLA; e.g., Coburn, 1998; Jungheim, 2008; Mohan & Helmer, 1988) are based on the work of Morris, Collett, Marsh, and O'Shaughnessy (1979). These studies have concentrated on assessing learners' ability to recognize and interpret L2 emblems. To do this, learners are shown videos with the emblems being performed and are then asked to fill out questionnaires or respond orally in the case of children (Mohan & Helmer, 1988) about what they thought the emblems meant.

The majority of the other studies have been based on two main theoretical frameworks: McNeillian and conversation analytic (CA). Although these frameworks are different, they are not mutually exclusive, and some studies combine them (e.g., Smotrova & Lantolf, 2013; van Compernelle & Williams, 2011). However, they bring different assumptions, perspectives, and methods to the research.

As mentioned previously, the McNeillian (1992, 2015) psycholinguistic perspective views speech and co-speech gesture as integral parts of language, a single-integrated system expressing two aspects of thought—the verbal and the imagistic. Speech and co-speech gesture are seen as developing together from a growth point. They complement each other, and together indicate the totality of what a speaker is thinking. This means that sometimes the speech and co-speech gesture represent the same entities, and sometimes the gestures indicate another aspect of thought that is in the speaker's mind but is not expressed through speech. Fundamental for research using this perspective is looking at the interaction of co-speech gestures and speech. This involves identifying the gesture stroke of the co-speech gesture in relation to speech through both regular and slow motion video with sound on.

The CA framework is a sociolinguistic perspective. "The central goal of conversational analytic research is the description and explication of the competences that ordinary speakers use and rely on in participating in intelligible, socially organized interaction" (Heritage & Atkinson, 1984, p. 1). Its primary

focus of analysis is sequences in conversation, turns, and repairs as they relate to specific speech acts (e.g., requests, invitations, offers, interviews, teaching sequences) in naturally occurring data (see Atkinson & Heritage, 1984), and it has well-developed transcription conventions. CA examines gaze, head movements, and body posture in relation to sequences and sees gestures as occurring with lexical affiliates (one word) or the rhythm of the language spoken (Schegloff, 1984) only in relation to turns. Because of its focus on sequences, it also considers gestures to include speech-linked gestures, those gestures that fill speech gaps (Olsher, 2008).

The methods that researchers use in their studies on SLA and gesture are based on their research questions and theoretical framework. For example, studies done within a McNeillian framework (e.g., Brown, 2008, 2015; Brown & Gullberg, 2008; Choi & Lantolf, 2008; Gullberg, 2008, 2009; Laurent, Nicoladis, & Marentette, 2015; Stam, 2006, 2010, 2015; Yoshioka, 2008) that investigate shifts in thinking for speaking (Slobin, 1991) from L1 to L2, L2 learners' use of referents, and bilingual language development tend to have clearly defined research questions about what is being investigated and methods. They also tend to be quantitative and qualitative (for an example, see Sample Study 36.1). They involve the use of a stimulus (a video or book) and videotaping of participants' narratives. All phases of gestures are coded in regular and slow motion, paying particular attention to relevant gestures for the study, and examination of the timing of the stroke of the gesture in relation to speech is paramount for answering their research questions.

### Sample Study 36.1

Stam, G. (2015). Changes in thinking for speaking: A longitudinal case study. *Modern Language Journal, 99*(S1), 83–99.

#### Research Background

The researcher, taking the perspective that speech and gesture are a single-integrated system (McNeill, 1992, 2005) and that there are cross-linguistic typological differences in the expression of motion events (Slobin, 1991), built upon a previous longitudinal study—1997 to 2006 (Stam, 2010)—to investigate whether a Mexican Spanish-speaking English language learner's thinking for speaking patterns about motion further changed linguistically and gesturally in her L1 (Spanish) and L2 (English) over five additional years (2011).

#### Research Problems/Gaps

The author sought to investigate how the participant's thinking for speaking patterns about motion changed linguistically and gesturally over a 14-year period and how it compared with monolingual Spanish speakers and native English speakers. This study fills one of the gaps in SLA research, the lack of long-term longitudinal research.

### Research Method

- *Type of research*: Quasi-experimental McNeillian longitudinal study of speech and gesture in expression of motion events (thinking for speaking).
- *Setting and participants*: Adult L1 Mexican Spanish learner of L2 English. Five monolingual Spanish speakers and five native English speakers (data from McNeill Lab, Center for Gesture and Speech, at the University of Chicago).
- *Instruments/techniques*: Audio-video recording of participants' narrations of the cartoon *Canary Row* (Freleng, 1950).
- *Data analysis*: Linguistic and gestural analysis of learner's cartoon narration in Spanish and English in 1997, 2006, 2011 and comparison with native speakers' data. Gesture and speech synchrony: where stroke of path and manner gestures occurred.

### Key Results

The participant's speech and gesture in the expression of motion in her L2 English changed over the 14 years under study, becoming more native speaker-like in the expression of path and manner; however, she retained some L1 Spanish features especially in terms of boundary crossing, indicating a resistance of the learner's thinking for speaking to shift completely from L1 to L2.

### Comments

This study provides one example of research on L2 thinking for speaking and what gestures tell us about learners' conceptualizations from a McNeillian perspective. This approach uses a McNeillian transcription and coding scheme (McNeill, 1992) and synchrony of the stroke of co-speech gesture with motion event speech elements to investigate how a L1 Spanish-speaking L2 English learner's expression of motion events changes longitudinally. The researcher's research questions guided her data collection methods and analysis.

Freleng, F. (Director). (1950). *Canary Row* [Animated Film]. New York: Time Warner.

Studies within a McNeillian framework that investigate whether gestures can facilitate learning are generally experimental in nature (e.g., Hirata, Kelly, Huang, & Manansala, 2014; Morett & Chang, 2015; Tellier, 2008) and follow standard experimental protocol with a control group, pre- and post-tests, different groups with different training conditions (one including gesture), and comparison of the various groups' results. These studies have investigated whether gestures facilitate the learning of vocabulary, perception of Japanese vowels, and Mandarin Chinese lexical tones, for example.

Those studies using a combination of McNeillian and Vygotskian frameworks (e.g., McCafferty, 1998; McCafferty & Ahmed, 2000; Peltier & McCafferty, 2010; Platt & Brooks, 2008) tend to be descriptive qualitative studies that use narrative, conversational, or classroom data to investigate

such topics as L2 private speech, zone of proximal development (ZPD), appropriation of metaphoric and beat gestures, self-regulation, and gestures of identity. Because of their exploratory nature, they may or may not have clearly defined research questions.

Classroom research studies also tend to be descriptive, but tend to be conducted within a CA framework. These studies investigate such topics as teachers' gestures during explanations and trouble talk, the functions of teachers' gestures, and learners' perceptions of them (e.g., Eskildsen & Wagner, 2013; Lazaraton, 2004; Sime, 2006). Again because these studies are exploratory in nature and are done within CA, they generally do not have clearly defined research questions and may interpret gesture as more than just co-speech gestures. Exceptions to CA studies are the ones by Smotrova and Lantolf (2013), which combines CA transcription with McNeillian and Vygotskian frameworks, and Tellier and Stam (2012), which is within a McNeillian framework and involves future French language teachers explaining words to native and non-native speakers of French outside of the classroom.

Other descriptive qualitative studies within a CA framework look at learners' gestures in terms of initiation of repairs, embodied completions, and intersubjectivity (e.g., Mori & Hayashi, 2006; Olsher, 2008; Seo & Koshik, 2010) during small group activities or tutoring sessions. These studies often do not have defined research questions and interpret gesture as more than just co-speech gestures. An exception to this is the study by Eskildsen and Wagner (2015) below.

## SLA and Gesture Studies: Challenges in Methods and Analysis

Because Gesture and SLA is a relatively new field, it presents a number of challenges in terms of methods and analysis.

### Methods

In terms of methods, the research varies widely in many areas, including the participants' L1 and L2, their L2 proficiency level or age, the types of data that are collected, the nature of the data collection, and the length of the study. As outlined above, although the methods are based on the theoretical perspective of a particular study and its research questions, there can be some overlap in the methods used from each of the different perspectives.

## Languages, Proficiency, and Age

Regarding languages, studies have been done with participants whose L1 is American Sign Language (ASL), Dutch, English, French, German, Indonesian, Italian, Japanese, Korean, Mandarin Chinese, Mexican Spanish, Swedish, Thai, Turkish, and Venezuelan Spanish and whose L2s have generally included Dutch, English, French, Italian, Japanese, Korean, Swahili, Swedish, and Spanish. There have also been several studies on bilinguals' use of gesture in the following language pairs: French-English, Hindi-English, Mandarin Chinese-Japanese, Spanish-English, and Turkish-English. Although speech and gesture have been investigated in a number of languages, it is not sufficient. More research is needed in a wider variety of L1s, L2s, and bilingual language pairs as well as participants who are fluent in more than two languages.

As for the proficiency levels and ages of participants in the research, the range includes monolinguals to bilinguals, beginners to near-native-like learners, and children aged four to adults. Although these ranges are quite inclusive, gaps remain for future research to address. For SLA to be able to address developmental questions, it is important to have a wider variety of ages and proficiency levels.

## Data Collection

The types of data collected by researchers interested in SLA and gesture include first and foremost audio-video recordings of the participants, which are then transcribed, so that gestures and speech can be captured and speech, gesture, and speech-gesture synchrony can be studied. Depending on their research questions, many researchers also collect proficiency measures for their participants through self-ratings or standardized assessments. Other assessments include auditory tests and gesture interpretation tests which are generally created for the purpose of a particular study. Interviews are also sometimes conducted, mainly for participants to reflect on their use of gesture or other nonverbal forms of communication. Questionnaires or online surveys are often used to collect background information on participants, to make sure that they did not realize that gesture was the focus of the study, or to ask participants for their perceptions on elements of the study such as their interest in the particular topic or their evaluations of gazing activities or use of gesture. In addition, some studies use eye tracking, near-infrared spectroscopy (NIRS), observation and field notes, collection of written texts and teaching

materials, informal conversations, stimulated recalls, rank ordering, Likert scales, or participant comments to inform their understanding.

For data collection, it is important to consider which research tasks, languages, groupings or conditions, and materials to use to answer particular research questions. These depend on the theoretical perspective for each study. Research tasks often include recording interactions of participants doing particular activities or narrating what they saw in a particular video or book to an interlocutor. From a CA perspective, researchers often watch video recordings and comment on particular aspects that arise in the data. The experimental studies generally include different groupings of participants who participate in different research conditions such as not seeing gestures, seeing gestures, repeating gestures, or seeing images instead of gestures and measuring and comparing the outcome on particular tasks. Groupings can include many participants or a whole classroom in addition to dyads, small groups, or even a single participant. Some studies investigate L1 gestures only, some L2 gestures only, but most look at both L1 and L2 gestures and the difference between them.

The most common materials used in data collection are video clips, shown to participants, who narrate what they viewed. The most common video clip used across studies is *Canary Row* (Freleng, 1950) with Sylvester and Tweety Bird which is used in thinking for speaking research due to the number of motion events in the clip. The advantage of this is that participants' narrations can be compared across languages or across proficiency level groups as they are narrating the same stimulus and it allows for replicability. Other videos that have been used include *Simpsons*, *Nine Months*, *Pink Panther*, *Pear Story*, and *Pingu* (a Swiss animated cartoon), a weather report, refusals and nonrefusals, classroom interactions, lectures, or a person performing emblems, putting away objects, or pronouncing particular words. Audio files, books such as Mayer's (1969) *Frog stories*, or printed cartoons have also been used.

Studies have ranged in length from one session to 14 years, although most are one session to a few weeks long. A common feature among the experimental studies is to counterbalance the tasks or languages used so that there is no ordering effect. Finally, in gesture studies, participants are generally only excluded if they do not gesture much.

## Analysis

Types of analysis also depend on the methods, theoretical perspective, and research questions. Researchers working from McNeillian, Vygotskian, or

CA perspectives each have specific ways in which they analyze their data. As most SLA gesture research involves collecting audio-video recordings in order to investigate gesture, analysis involves watching video data many times, often in slow motion for gesture-speech synchrony, and transcribing and/or coding both speech and gesture. Researchers of nonverbal behaviors in general code for blinking; dropping a hand; positions of head, torso, or brow; gaze shifts and the object of gaze; head or hand repetitions; and embodied completions.

### **Gesture Coding and Speech Transcription**

Several different software programs have been used for transcriptions and coding. These include ELAN, Final Cut Express, CHAT, AUSLAN, SPPAS, iMovie, MUMIN, MediaTagger, and Adobe Premiere Elements. How researchers code within a particular program is dependent on what is of importance to them and which coding scheme they are using (e.g., Ekman & Friesen, 1969; Kendon, 2004; McNeill, 1992).

Therefore, depending on what their analysis will be, researchers may code gestures for the following: illustrators, regulators, affect displays, self-adaptors, path, manner, path and manner, ground, iconics, metaphorics, deictics, beats, conventional, cohesive, Butterworth, interactive, proxemics, kinesics, haptics, artifactual communication, kinetographic, iconographic, baton, and those that perform self-regulatory functions. In addition, they may code for gesture phrases, phases, and space, imitations, catchments, meaning of gesture, viewpoint (character or observer), hand shape, movement, handedness, and the trajectory. Because not every researcher codes the same way or analyzes the same aspects of gesture, comparisons between studies are often difficult.

For analysis of speech, researchers generally code for pauses, the introduction and syntactic role of particular verbal elements, private speech, speech disfluencies, timing, language used, and particular speech units, such as I-phrases, utterances, clauses, or sentence units containing particular feature(s). Those researchers who are working within a CA framework follow CA transcription conventions (see Sample Study 36.2). In contrast, researchers who are interested in gesture-speech synchrony often code for the co-occurrence of gesture and the following: ground noun phrase, verb, satellite, references to people, lexical expressions, lexical searches, pauses, referring expressions, and inanimate introductions.

## Sample Study 36.2

Eskildsen, S. W., & Wagner, J. (2015). Embodied L2 construction learning. *Language Learning*, 65, 268–297.

### Research Background

As gestures are considered a “window onto cognition,” as a resource for inter-subjectivity, and as part of interactional development, Eskildsen and Wagner (2015) examined their role in each of these areas in L2 vocabulary teaching and learning of two English prepositions *under* and *across (from)* (Kendon, 1972; McNeill, 1992).

### Research Problems/Gaps

The authors sought to investigate how two English prepositions, *across (from)* and *under*, are accompanied by gestures over time in second language vocabulary learning for one L2 learner and how the gestures are reused later with the participant’s classmates in order to demonstrate understanding.

### Research Method

- *Type of research:* Conversation analytic investigation of speech and gesture
- *Setting and participant:* Adult L1 Mexican Spanish, L2 English learner; ESL program at community college in the United States
- *Instruments/techniques:* Audio-video recordings of classroom interactions
- *Data analysis:* Conversation analysis of the first six instances of participant’s use of *under* in 2.5 months and *across (from)* over a period of 3 years, 4 months

### Key Results

For *under*, both the teacher and the participant’s use of gesture to explain and understand this term, especially at the beginning, were crucial. Although the participant’s gestures retained essential features across time for *under*, his use of gesture changed over time to become more spontaneous and less hesitant. As for *across (from)*, the participant developed the ability to distinguish *across* from *across from* as seen through the change in his gestures. Their analyses indicate the gradual nature of L2 learning and the need for opportunities in the L2 in different situations in order for learners to appropriate environmental affordances.

### Comments

This study provides one example of research on speech and gesture from a conversation analytic perspective. This approach uses transcription, with particular transcription conventions, to investigate the sequential organization of participants’ use of speech and/or gesture. From this perspective, the researcher sets out to investigate a phenomenon without any preconceived ideas of what they will find.



## Types of Analysis

Those doing research from a quantitative perspective measure a wide variety of speech, gestural, and nonverbal elements. For speech, it often includes the number of clauses, referents, words or words per minute, the length of utterances, word types, word tokens, speech time, or introductions of referents. For gesture, it often includes the number of gestures, the length, the rate, the proportion of a particular type of gesture within the overall dataset, the dimension of gestures, or the gesture space used. For nonverbal behavior, the most common is eye fixation. Finally, most studies calculate interrater reliability.

Researchers interested in making comparisons most often investigate whether participants' gestures are more similar to L1 or L2 users. Other comparisons include participants' speech and gesture in the L1 compared to the L2 under different conditions, the difference in the gesture rate between the L1 and L2, and the differences between different groups or conditions, in eye gazing, or over time. ANOVAs and *t*-tests are the most common statistics that are run on SLA gesture data; however, others consist of descriptive statistics, the Kuder-Richardson formula 20, means, percentages, standard deviations, and chi-squares.

While the above is most often used by those doing experimental research, those doing CA describe or analyze teachers' gestures, students' use of artifacts in a particular setting, learners' use of gesture in small group activities, use of gesture for interpersonal cognition, types of gestures used to initiate repair sequences and embodied actions, and the functions of particular types of gestures. These studies do not generally have research questions and instead their focus is on describing students' or teachers' use of gestures in particular settings, which make them difficult to replicate.

Depending on the type of research done, some elements in the data are excluded and do not contribute to the analysis such as disfluencies, gestures with objects, repair or performance features, unintelligible sounds, self-referential clauses, gestures produced to elicit assistance from the interlocutor, or the use of a mix of the L1 and L2.

## Summary

There are several issues and challenges concerning the methods used in gesture and SLA research, such as what is meant by the term *gesture*. Some researchers are open to including any nonverbal behavior such as bodily movements or eye gaze, while others restrict their focus to only co-speech gestures.

Consequently, there is a need for common definitions to be used by researchers. The differences in the kinds of data that are collected can be problematic because they do not allow for comparison across studies. The variety in the way that SLA researchers analyze gestures also presents challenges for the field because claims by one researcher may contradict another's, but they may be looking at very different aspects of gesture. Although methods and analyses used in a study depend on the theoretical perspective one is taking and the research questions being asked, common definitions, research methods, and analyses could provide a way for the field to grow, develop, and make contributions to the body of knowledge on L2 acquisition.

## Future Directions

As a new field, gesture research in SLA has the potential to grow and expand our understanding of the L2 acquisition process and improve L2 teaching. However, for this area of research to make the contributions it is capable of, we need additional naturalistic and controlled research studies across more languages as well as access to studies that are being conducted in other countries. In addition, studies need to have clear research questions and clear definitions about what is being studied and analyzed so that they can be replicated. Also the classification system that is being used to code and analyze gesture needs to be clearly delineated. This does not mean that we need to study the same phenomenon or use the same theoretical perspective, but rather that we should approach the research in a more systematic way.

Furthermore, training in the coding and analysis of gesture is needed. Coding of gesture requires practice. Researchers need to learn how to see movement and differentiate between different types of movement and understand their relevance for research questions. Coding needs to be reviewed and questions about it addressed by gesture coding experts for grounding. This could be done by the sharing of data and by workshops on coding offered at professional organizations or institutions.

## Resources for Future Reading

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

This book by Adam Kendon is aimed at researchers in the fields of linguistics, communication, semiotics, anthropology, cognitive science, and gesture

study that are interested in research on gesture. It provides an in-depth overview of the history of gesture, what gesture is, how it is used in interaction, and different types of gesture.

McCafferty, S. G., & Stam, G. (Eds.). (2008). *Gesture: Second language acquisition and classroom research*. New York: Routledge.

This volume, edited by Steven McCafferty and Gale Stam, is the first one on gesture in L2 research and is suitable for researchers in the fields of anthropology, communication, linguistics, psychology, sociology, and language teaching. The five sections of the book provide solid evidence for why gesture should be considered in L2 research. Section 1 gives an overview of gesture studies and its application for L2 research, while the other sections have illustrative studies of L2 and gesture research in different contexts.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: The University of Chicago Press.

This book by David McNeill is for any researcher who is interested in learning more about McNeill's theory of the interconnectedness of speech, gesture, and thought. The book clearly explains what gestures are, how they are used in different contexts, and how they can be studied. It also provides an illustration of McNeill's coding scheme that has been used in many L2 gesture studies.

McNeill, D. (2005). *Gesture & thought*. Chicago: The University of Chicago Press.

This book further develops the language-imagery dialectic that McNeill proposed in *Hand and Mind*. It provides a clear description of the growth point, in Vygotskian terms a psychological predicate and a fundamental principle of McNeill's model. It also contains evidence from speech and gesture studies conducted by both McNeill and others that support his viewpoint and a chapter on gesture coding.

## References

- Atkinson, J. M., & Heritage, J. (Eds.). (1984). *Structures of social action*. New York: Cambridge University Press.
- Brown, A. (2008). Gesture viewpoint in Japanese and English: Cross-linguistic interactions between two languages in one speaker. *Gesture*, 8, 256–276.

- Brown, A. (2015). Universal development and L1–L2 convergence in bilingual construal of manner in speech and gesture in Mandarin, Japanese, and English. *Modern Language Journal*, 99(S1), 66–82.
- Brown, A., & Gullberg, M. (2008). Bidirectional crosslinguistic influence in L1–L2 encoding of manner in speech and gesture: A study of Japanese speakers of English. *Studies in Second Language Acquisition*, 30, 225–251.
- Choi, S., & Lantolf, J. P. (2008). Representation and embodiment of meaning in L2 communication: Motion events in the speech and gesture of advanced L2 Korean and L2 English speakers. *Studies in Second Language Acquisition*, 30, 191–224.
- Coburn, K. (1998). The nonverbal component of communicative competence. In S. Santi, I. Guaïtella, C. Cavé, & G. Konopczynski (Eds.), *Oralité et gestualité: Communication multimodale, interaction* (pp. 625–628). Paris: L'Harmattan.
- van Compernelle, R. A., & Williams, L. (2011). Thinking with your hands: Speech-gesture activity during an L2 awareness-raising task. *Language Awareness*, 20, 203–219.
- Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1, 49–98.
- Eskildsen, S. W., & Wagner, J. (2013). Recurring and shared gestures in the L2 classroom: Resources for teaching and learning. *European Journal of Applied Linguistics*, 1, 139–161.
- Eskildsen, S. W., & Wagner, J. (2015). Embodied L2 construction learning. *Language Learning*, 65, 268–297.
- Freleng, F. (Director). (1950). *Canary Row* [Animated Film]. New York: Time Warner.
- Gullberg, M. (2008). A helping hand? Gestures, L2 learners, and grammar. In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 185–210). New York: Routledge.
- Gullberg, M. (2009). Reconstructing verb meaning in a second language: How English speakers of L2 Dutch talk and gesture about placement. *Annual Review of Cognitive Linguistics*, 7, 221–245.
- Heritage, J., & Atkinson, J. M. (1984). Introduction. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 1–15). New York: Cambridge University Press.
- Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research*, 57, 2090–2101.
- Jungheim, N. O. (2008). Language learner and native speaker perceptions of Japanese refusal gestures portrayed in video. In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 157–182). New York: Routledge.
- Kendon, A. (1972). Some relationships between body motion and speech. In A. Siegman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177–210). Elmsford, NY: Pergamon Press.
- Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). The Hague: Mouton.

- Laurent, A., Nicoladis, E., & Marentette, P. (2015). The development of storytelling in two languages with words and gestures. *International Journal of Bilingualism, 19*, 56–74.
- Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher: A microanalytic inquiry. *Language Learning, 54*, 79–117.
- Mayer, M. (1969). *Frog, where are you?* New York: Dial Press.
- McCafferty, S. G. (1998). Nonverbal expression and L2 private speech. *Applied Linguistics, 19*, 73–96.
- McCafferty, S. G., & Ahmed, M. K. (2000). The appropriation of gestures of the abstract in L2 learners. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 199–218). Oxford: Oxford University Press.
- McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92*, 350–371.
- McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.
- McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.
- Mohan, B., & Helmer, S. (1988). Context and second language development: Preschooler's comprehension of gestures. *Applied Linguistics, 9*, 275–292.
- Mori, J., & Hayashi, M. (2006). The achievement of intersubjectivity through embodied completions: A study of interactions between first and second language speakers. *Applied Linguistics, 27*, 195–219.
- Morett, L. M., & Chang, L.-Y. (2015). Emphasizing sound and meaning: Pitch gestures to enhance Mandarin lexical tone acquisition. *Language, Cognition, and Neuroscience, 30*, 347–353.
- Morris, D., Collett, P., Marsh, P., & O'Shaughnessy, M. (1979). *Gestures, their origins and distribution*. London: Jonathan Cape.
- Olsher, D. (2008). Gesturally-enhanced repeats in repair turn: Communication strategy or cognitive language-learning tool? In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 109–130). New York: Routledge.
- Peltier, I. N., & McCafferty, S. G. (2010). Gesture and identity in the teaching and learning of Italian. *Mind, Culture, and Activity, 17*, 331–349.
- Pennycook, A. (1985). Actions speak louder than words: Paralinguage, communication and education. *TESOL Quarterly, 19*, 259–282.
- Platt, E., & Brooks, F. B. (2008). Embodiment as self-regulation in L2 task performance. In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 66–87). New York: Routledge.
- Schegloff, E. A. (1984). On some gestures' relation to talk. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 266–296). New York: Cambridge University Press.
- Seo, M.-S., & Koshik, I. (2010). A conversational analytic study of gestures that engender repair in ESL conversational tutoring. *Journal of Pragmatics, 42*, 2219–2239.

- Sime, D. (2006). What do learners make of teachers' gestures in the classroom? *International Review of Applied Linguistics*, 44, 211–230.
- Slobin, D. I. (1991). Learning to think for speaking: Native language, cognition, and rhetorical style. *Pragmatics*, 1, 7–26.
- Smotrova, T., & Lantolf, J. P. (2013). The function of gesture in lexically focused L2 instructional conversations. *Modern Language Journal*, 97, 397–416.
- Stam, G. (2006). Thinking for speaking about motion: L1 and L2 speech and gesture. *International Review of Applied Linguistics*, 44, 143–169.
- Stam, G. (2010). Can a L2 speaker's patterns of thinking for speaking change? In Z. Han & T. Cadierno (Eds.), *Linguistic relativity in SLA: Thinking for speaking* (pp. 59–83). Clevedon, UK: Multilingual Matters.
- Stam, G. (2013). Second language acquisition and gesture. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Oxford: Blackwell Publishing Ltd.
- Stam, G. (2015). Changes in thinking for speaking: A longitudinal case study. *Modern Language Journal*, 99(S1), 83–99.
- Stam, G., & McCafferty, S. G. (2008). Gesture studies and second language acquisition: A review. In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 3–24). New York: Routledge.
- Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, 8, 219–235.
- Tellier, M., & Stam, G. (2012). Stratégies verbales et gestuelles dans l'explication lexical d'un verbe d'action. In V. Rivière (Ed.), *Spécificités et diversité des interactions didactiques* (pp. 357–374). Paris: Riveneuve editions.
- von Raffler-Engel, W. (1980). Kinesics and paralinguistics: A neglected factor in second language research. *Canadian Modern Language Review*, 36(2), 225–237.
- Wylie, L. (1985). Language learning and communication. *The French Review*, 58, 777–785.
- Yoshioka, K. (2008). Linguistic and gesture introduction of Ground reference in L1 and L2 Narrative. In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 211–230). New York: Routledge.



# 37

## Language Policy and Planning

David Cassels Johnson and Crissa Stephens

### Introduction

This chapter examines the historical development of research methods in language policy and planning (LPP). Expanding upon Ricento (2000) and Johnson and Ricento (2013), we separate the history of the field into four eras. The LPP field began pragmatically, without fixed methods or an established disciplinary identity, out of calls for practical solutions to language “problems” in developing nations. Most of the first wave of LPP research focused on developing typologies, heuristics, and frameworks for corpus and status planning. During the second wave of LPP research, important questions of epistemology, method, and research goals arose and engendered debate about the social impact of language planning. As theories of LPP continued to emerge, the need to test, expand, and refine them continued. Ushering in a third wave, scholars employing critical (socio)linguistics and social theory began to question the proposed ideological neutrality of earlier frameworks and, instead, focused on how language policies were hegemonic instruments of social control. A fourth wave of LPP research—an “empirical turn”—has been shaped by increased empirical interest in the role of agents in LPP processes. This chapter concludes with an examination of current debates and innovative solutions, which are pushing scholars into new methodological territory where they must grapple more deeply with issues of research ethics, social activism,

---

D. C. Johnson (✉) • C. Stephens

College of Education, University of Iowa, Iowa City, Iowa, USA

e-mail: [david-c-johnson@uiowa.edu](mailto:david-c-johnson@uiowa.edu); [crissa.stephens@gmail.com](mailto:crissa.stephens@gmail.com)

and engagement. We argue that these methodological issues will help shape the future of LPP research and illustrate these issues through two sample studies that appear at the end of the chapter.

## Core Issues

Haugen (1959, p. 8) originally defined language planning as “the activity of preparing a normative orthography, grammar, and dictionary for the guidance of writers and speakers in a non-homogeneous speech community.” Kloss (1969) expanded on this definition, identifying what Haugen had described as corpus planning (i.e., planning the forms of a language), which Kloss differentiated from status planning (i.e., planning the functions of a language). Early developments in the field (Wave 1: 1960s–1970s) were engendered by (socio)linguists who were enlisted to help with language development for indigenous languages in developing nations (Fishman, Ferguson, & Das Gupta, 1968). Most of this early research focused on developing typologies, taxonomies, and models for language planning. For example, Haugen (1983) proposes four processes of language planning (Haugen, 1983, p. 273):

1. selecting a norm (deciding which language varieties should be used in which social contexts);
2. codification (the development of an explicit form of the language, usually written);
3. implementation (strategies for spreading the language form); and
4. elaboration (or updating the language variety continually in keeping with changes in the “modern world.”)

Early work in language planning was dominated by objectivist epistemologies and research methods, characterized by an emphasis on delimiting the subjectivities of researchers and an interest in excluding sociopolitical variables in the science of language planning, which ignored ideological and discursive aspects of policy. Some early LPP scholars were concerned with determining which languages in a multilingual (often postcolonial) environment were best suited for standardization, furthering the goals of national unification, the spread of literacy, and/or modernization (Ricento, 2000). For example, Tauli (1974, p. 51) made controversial claims about the utility of languages, which he claims “can be evaluated with objective scientific often quantitative methods ... Not all languages describe things equally effectively.”



This work treated issues of language planning as “problems” to be solved and saw language itself as relatively neutral, unimplicated in issues of identity, and unmired by the effects of social power structures.

These early activities engendered important edited volumes like *Language Problems of Developing Nations* (Fishman et al., 1968) and *Can Language Be Planned?* (Rubin & Jernudd, 1971). Most of the contributions were conceptual proposals, descriptive accounts of language planning projects, strategies for implementing language planning, and historical investigations of particular contexts and communities. Though methods in this early work were understandably inchoate and exploratory, this early work helped identify potential data sources and theories and conceptual frameworks to guide data collection, which influenced LPP research methods that would follow.

During the second wave (late 1970s–early 1980s), questions of the sociopolitical ramifications of language planning became more important. In what Ricento (2000) calls an “intermediary stage” of LPP research, researchers built upon the classic work discussed above with new critiques emerging that focused on the separation of language from its sociocultural and political contexts. Cobarrubias (1983) argues that corpus and acquisition planning are not “philosophically neutral” (p. 41) because status planning is inextricable from the functions of language in social institutions such as education and religion and the actions of language planners themselves are tied to ideological principles with social consequences for the populations they impact. Even more forcefully, in Ruiz’s (1984, p. 2) influential article on orientations in language planning, he declares that “[o]rientations are basic to language planning in that they delimit the ways we talk about language and language issues ... they help to delimit the range of acceptable attitudes toward language, and to make certain attitudes legitimate. In short, orientations determine what is thinkable about language in society.”

As LPP developed, the objectivism that characterized much of the early scholarship was called into question, a movement that was aligned with critical linguistics (Fowler, Hodge, Kress, & Trew, 1979), systemic functional linguistics (Halliday, 1985), and developments in sociolinguistics (Hymes, 1972). Thus, LPP research was influenced by a movement away from objectivist research that transcended individual disciplinary divisions and foregrounded power and inequality. This shift in research foci foreshadowed the turn toward poststructuralist and postmodern theoretical orientations that acknowledge the social, constructive role of language, and seek to interrogate its implication in social power structures.

## Challenges and Controversial Issues

Most of the challenging and controversial issues in LPP arose with the emergence of critical language policy (CLP) (Wave 3: 1990s–2000s). Tollefson's (1991) *Planning Language, Planning Inequality* helped usher in a third wave of LPP research, within which he lays out a withering critique of what he calls the neoclassical approach, and instead proposes the historical-structural approach. Like early language planning research, historical-structural analysis focuses on state policies and institutions but emphasizes historical events and structural mechanisms that engender social inequality. In subsequent publications, Tollefson (2002, 2006, 2013) outlines the “critical” in CLP research, which (1) is critical of neoclassical language planning research; (2) is influenced by critical theory; (3) emphasizes the relationships among language, power, and inequality, which are held to be central concepts for understanding language and society; and (4) entails social activism.

The connections between language (policy) and power are a central focus, although conceptualizations vary, and include state power, ideological power, and discursive power (Tollefson, 2015). Salient examples of each of these foci include Wiley and Wright's (2004) historical examination of how US federal educational policy has marginalized minority and indigenous language users (state power); Pan's (2011) analysis of how English language policy in China relies on ideological hegemony (ideological power); and Barakos' (2016) incorporation of governmentality (Foucault, 1991) to highlight how Welsh language policy gets recontextualized as essentialist ideologies within bilingual business practices that promote language as a neoliberal tool of consumerism (discursive power). Barakos' scholarship and many others (e.g., Cincotta-Segi, 2011; Johnson, 2013) reveal how the CLP framework can be extended to focus on micro-level interactions, particularly when combined with discourse analytic techniques.

The historical-structural approach, and CLP research more generally, opened the door for the proliferation of research on the role of ideology, identity, and discourse in language policy processes and has left a durable mark on LPP research methods. Yet, CLP has also been criticized for being overly deterministic and underestimating the role of human agency (Ricento & Hornberger, 1996), for not capturing language planning processes (Davis, 1999), and for diminishing the benevolent impact of many language planning initiatives (Fishman, 1994). Subsequent LPP researchers utilized, built upon, and critiqued CLP and tested the theories with empirical data collection in schools and communities. This fourth wave of LPP research—the

“empirical turn” (2000s onward)—has enjoyed increasing theoretical and methodological developments. Some have corroborated the tenets of CLP and demonstrated the power of language policy to reinforce existing social structures that marginalize minority languages and their users (e.g., Tollefson, 2013). On the other hand, others have demonstrated the power of individuals to interpret and appropriate language policy in creative and unpredictable ways (e.g., Menken & García, 2010). Furthermore, some research has shown how both local and national efforts to create more democratic language policy can increase educational and social opportunities, especially policies that promote multilingualism as a resource in schools and society (e.g., Hornberger & Johnson, 2007). Nevertheless, while a new sense of optimism has perhaps permeated this focus on agency, the power of policy as discourse must not be ignored, and other scholarship has attempted to strike this balance (e.g., Johnson & Johnson, 2015).

For the majority of the history of the field, scholars have relied upon their disciplinary expertise (in sociology, political science, law, linguistics, economics, etc.), yet the fourth wave of LPP research has also been characterized by the development of LPP-specific research methods. Many of these research methods help ameliorate the “perennial challenge” (Hult, 2010) in the field—making connections across and between the diverse layers and levels of policy texts, discourses, and practices (i.e., the macro to the micro). Ethnography of language policy has been influential in furthering both theory and method in the field (Johnson, 2009; McCarty, 2011), contributing insights into the role of language policies in structuring inequality and reinforcing dominant ideologies (Chimbutane, 2011; Johnson & Johnson, 2015; Pérez-Milans, 2015) and illuminating the ways that policies are created, interpreted, and appropriated in education (De Costa, 2010; Hopkins, 2012; Johnson, 2012; Mortimer, 2016). Ethnographic approaches that capture the socially situated nature of language policy on the ground have also revealed how the actions of language policy stakeholders can promote equality and counter dominant ideologies (Freeman, 1998; Hornberger, 1998; McCarty, Collins, & Hopson, 2011). Rather than focusing purely on the language of policy documents, work in this vein has characterized teachers as the final arbiters of educational language policy (Hornberger & Johnson, 2007; Menken & García, 2010; Ricento & Hornberger, 1996; Stephens & Johnson, 2015), while Johnson and Johnson (2015) expand on this to investigate the possibility that a number of individuals can be *language policy arbiters*. This research provides a counterpoint to a unidirectional conceptualization of power and discourse in language policy implementation.

Diverse discourse analytic techniques have proven particularly useful for making connections across language policy texts, discourses, practices, and contexts including critical discourse analysis (CDA) (Cincotta-Segi, 2009; Johnson, 2011), nexus analysis (Hult, 2010, 2015), scalar approaches that connect language policy activity as it develops across social spaces and time (Hult, 2010; Mortimer, 2016), narrative analysis to examine policy and ideology (King, 2013; Lee, 2009), intertextual LPP analysis (Johnson, 2015), linguistic landscape analysis (Hult, 2014; Shohamy & Gorter, 2009), and linguistic-anthropological approaches (Mortimer & Wortham, 2015), such as the use of speech chains to trace the situated meanings of policy language from official documents to various social and educational contexts (see Mortimer, 2013).

There is general agreement in the field that language policy should be conceptualized and studied as multiply leveled (or layered). Yet, recent research has critiqued monolithic and static depictions of the relationship between macro and micro because they do not account for the multiple constraints on social interaction, which can change over time. Furthermore, critics argue that research focusing on the micro tends to overestimate the individual's ability for novel and seminal action (Wortham & Reyes, 2015). Questioning and reconceptualizing the macro-micro dialectic is becoming an important feature within LPP research. While a multiply-layered understanding of context is implicit in the macro-micro distinction (especially when other "meso" layers are added), ideologies are multiply layered as well and can change. As Mortimer and Wortham (2015, p. 163) argue, "[I]nstead of connecting micro-level events to macro-level structures (e.g., connecting a classroom language practice to an official policy), we must explore heterogeneous domains and scales of social organization relevant to understanding meaningful social action." Therefore, as Hult (2010) has argued, within any discursive event, there are many potential sociolinguistic scales at work, and the analyst identifies how the unique configuration of semiotic resources is made relevant within the interaction. It is tempting to equate structure with macro-level social processes/systems and agency with micro-level human interactions, yet both macro and micro discourses, and both structure and agency, can emerge in a single discursive event and shape a single policy document. Nevertheless, the benefit of replacing levels/layers with "scales" is still being debated, especially in LPP research that must characterize physical contexts (as opposed to, or as well as, discursive contexts) and/or does not make use of discourse analytic techniques.

While ethnography and different forms of discourse analysis have been widely used in empirical studies, LPP research is characterized by a marked

multidisciplinarity as researchers leverage the methods from their disciplinary home to study LPP processes. Examples are too numerous to review in this chapter but include Grin (2003), who uses economic research methods to analyze the value of language acquisition and has helped established the economics of language policy (Grin & Vaillancourt, 2015); May (2008) and colleagues (e.g., Sonntag, 2009), who use political theory to explore LPP processes and discourses; and Kochenov and de Varennes (2015), who consider how laws and legal systems impact LPP.

## Limitations and Future Directions

The need for LPP conceptualizations that capture language policy as socially embedded, never neutral, and multiply-layered has spurred theoretical and methodological innovation in LPP (Hult, 2010; Johnson, 2011; Martin-Jones, 2015; Mortimer, 2013). Some researchers have framed this as a continuum of macro to micro influences on policy activity (Johnson, 2009), while others have sought to reframe this dichotomy (Hult, 2010; Karjalainen, 2016; Mortimer, 2016). Research has continued to focus on the power, influences, and origins of macro-level ideologies and policies (Barakos, 2016; Pan, 2011; Tollefson, 2015; Wiley & Wright, 2004), while other empirical research has illuminated policy activity in schools and communities (Amir & Musk, 2013; Bonacina-Pugh, 2012). Martin-Jones (2015) points to the current methodological trend of combining ethnographic data collection with discourse analysis to illuminate connections across multiple layers of policy activity (e.g., Shoba & Chimbutane, 2013).

LPP research methods have not lost their interdisciplinarity—nor their connections to their foundations in (socio)linguistics, sociology, economics, law, and political theory—but they have matured in their own right. In Hult and Johnson's (2015) volume *Research Methods in Language Policy and Planning*, for example, researchers representing varying disciplinary homes present LPP-specific research methods. Each of the chapters includes some theoretical background, ideas for developing research questions, and data collection/analysis techniques. In contrast to LPP studies in eras past, researchers now have a broader base of more developed, articulated methodological tools from which to choose.

Through the evolution of methods in the first three eras of LPP research, it is now well-established that language policies and plans have social implications and consequences. The trend toward ethnographic and discursive methods (that acknowledge the social role of language policies) has motivated

researchers to examine their own positionality in research and policy processes (Johnson, 2017; Lin, 2015; Perez-Milans, 2012). We see an emergence of “critical” ethnography (Heller, 2011; McCarty et al., 2011; Phyak, 2013) and increasing interest in activism, research ethics, and epistemology in LPP research (Canagarajah & Stanley, 2015; De Costa, 2014). Madison (2005) argues that within a critical ethnographic paradigm, researchers must always account for their own subjectivity “in relation to the Other” (p. 9) through dialogue with participants and making a meaningful difference in the worlds of both researchers and Others.

One of the areas in which researchers can openly articulate their own positions and subjectivities is in the written report. Ramanathan (2006, 2011) grapples with her own conflict in researching-texting practices and criticizes LPP scholarship that presents its textual products in hermetically sealed ways—“a modus operandi that often leaves little room for addressing uncertainties and tensions in the researching-texting process” (Ramanathan, 2011, p. 256). Instead, she emphasizes the development of meaning in texts as a process that should be rendered “porous, unstable, and changeable” (p. 247). Echoing Fishman’s (1994) critique, she emphasizes how crucial it is to interrogate the ideological aspects of research and “complexify the researcher’s voice” (p. 268). Similarly, Canagarajah and Stanley (2015, p. 41) argue that genre conventions in academic writing make it challenging to give voice to minority communities, yet it is essential that LPP scholars push back against positivistic writing genres that attempt to synthesize research findings into generalizable and monolithic “truths”: “Since the subjects exist in the report only through the voice of the researcher, there is a tendency for their complexity to be suppressed and their identity to be generalized (or essentialized).”

### Sample Study 37.1

Chimbutane, F. (2011). *Rethinking bilingual education in postcolonial contexts*. Bristol: Multilingual Matters.

#### Research Background

In Mozambique, Portugal enforced monolingual Portuguese education as a tool of cultural assimilation, concomitantly positioning African languages as inferior dialects that should be eliminated from schools. Acquisition of Portuguese was a necessary precondition for becoming an *assimilado*, a class status that ranked lower than white Europeans but higher than the large majority of indigenous Africans. Mozambique gained independence in 1975 and yet Portuguese was promptly declared the official language, the goal of which was to preserve national unity through a common lingua franca. Nevertheless, the 1990 constitution promoted the use of African languages and a 2003 educational language policy promoted bilingual education.

### Research Method

- Combines one-year ethnographic data collection (2007–2008) and discourse analytic methods.
- *Setting and participants*: Grade 4 and 5 students and teachers in two bilingual education schools in rural Mozambique.
- *Data collection*: Classroom observation, audio recordings, note taking, interviewing, questionnaires, and document analysis; also, the larger historical, sociopolitical, sociolinguistic and policy context in Mozambique in the spirit of Tollefson's (1991) historical-structural approach.
- *Data analysis*: Discourse analysis reveals how the colonial history of educational language policy plays out in the local policies and practices in the schools.

### Key Results

Chimbutane argues that medium-of-instruction policies help maintain the hegemonic status of Portuguese and the main purpose of bilingual education remains helping students transition into Portuguese. Nonetheless, he challenges the notion that transitional bilingual education *always* leads to cultural assimilation and loss of the L1, and especially in a context where pupils are surrounded by their mother tongue outside of school, these programs can strengthen the maintenance of low-status languages. While bilingual education might open ideological space for African languages, leading to social and cultural transformation, Portuguese is still viewed as *the* mechanism for social opportunity.

### Comments

Chimbutane leveraged his positionality as a former insider within the schools and his knowledge of both African languages—Changana and Chope—to provide a more robust ethnographic understanding. The combination of ethnographic data collection and sociohistorical and discourse analysis allows Chimbutane to examine the connections across multiple layers of LPP activity. This study is an important contribution to the understanding of lasting impact of colonial language policies in Africa.

### Sample Study 37.2

Fitzsimmons-Doolan, S. (2009). Is public discourse about language policy really public discourse about immigration? A corpus-based study. *Language Policy*, 8(4), 377–402.

### Background

The notion that language policies are influenced by language ideologies is a popular sentiment within the field. Fitzsimmons-Doolan's goal is to test this assumption by examining the overlap between public discourse about language policy and public discourse about immigration in newspapers. A series of English-focused language policies in Arizona—Proposition 203, which restricted access to bilingual education, and Proposition 103, which made English the official language—have stirred this debate.



### Research Method

- Corpus linguistics analysis of newspaper articles.
- Newspapers from Tucson and Phoenix are included because the former is a traditionally politically liberal city, while the latter is much more conservative.
- Corpus-based research uses computer software (in this case WordSmith Tools) to analyze large bodies of naturally occurring texts.
- Fitzsimmons-Doolan uses keyword analysis, which identifies the words in a text that indicate what it is about or, as she describes it, the “aboutness” of a corpus. Keywords are emblematic of the corpus.
- This created four corpora: (1) Tucson language policies, (2) Tucson immigration, (3) Phoenix language policies, and (4) Phoenix immigration.

### Key Findings

Results revealed that, of the top 20 keywords in each of the corpora, only 6 words (6%) overlapped and none of those words were especially related to policy. This finding challenges the argument that there is a link in public discourse about language policy and public discourse about immigration, as reflected in newspapers. While she did not find shared aboutness in the immigration and language policy corpora, a look at the collocations of non-keywords did reveal biases against Spanish, the naturalization of English as *a/the* language, and negative sentiments about immigrants and immigration.

### Comments

Challenging a common assumption in the field, Fitzsimmons-Doolan argues that public discourse in newspaper articles does not conflate nativism with restrictive language policy. However, as she herself notes, media discourse may or may not reflect what takes place in communities and classrooms or official language policies.

## Conclusion

In this chapter, we have identified four eras of theoretical and methodological development in the field of language policy and planning (LPP). The field was borne out of a utilitarian focus on helping developing nations and indigenous groups with language planning efforts, and the first wave (1960s–1970s) was characterized by a focus on frameworks, models, and heuristics. During the second wave of LPP research (1970s–1980s), the objectivist epistemologies embedded within the field were increasingly questioned, yet the general research focus on macro-level explanatory frameworks continued. The third wave of LPP research (1990s–2000s) was marked by a critical turn in LPP research and ushered in by Tollefson’s (1991) proposed historical-structural approach to language policy research. The fourth wave of LPP research (2000s–onward)—the “empirical turn”—has seen a growing number of



empirical studies of language policy processes, discourses, texts, and practices across multiple layers/levels/scales of LPP activity. This new wave of research has illuminated earlier theoretical frameworks, engendered theoretical debates focused on power and agency, and helped develop LPP-specific research methods. An emphasis on impact, engagements, ethics, and interrogation of research subjectivity in LPP processes will help define the field going forward and potentially proffer a new set of methodological tools that will take us into a new wave of LPP research.

## Resources for Further Reading

Corson, D. (1999). *Language policy in schools: A resource for teachers and administrators*. Mahwah, NJ: Lawrence Erlbaum Associates.

As advertised, this book is a valuable resource for teachers and administrators engaged in language planning and policy initiatives in schools. Corson portrays these initiatives as necessarily egalitarian, yet sustained by “emancipatory leaders,” if they are to be successful. Driving Corson’s proposals is a critical emphasis on social justice for students and teachers.

Hult, F. M., & Johnson, D. C. (Eds.). (2015). *Research methods in language policy and planning: A practical guide*. Malden, MA: Wiley-Blackwell.

This edited volume covers a wide variety of research methods used in language policy and planning research. Each chapter is written by a leading expert in the field with wide-ranging disciplinary backgrounds—economics, law, linguistics, education, political theory, and sociology. Each chapter explains the theoretical orientations and research methods involved in the different research traditions with accessible applications to LPP projects.

Johnson, D. C. (2013). *Language policy*. Basingstoke: Palgrave Macmillan.

Johnson’s book summarizes and synthesizes the research on language policy, drawing together the disparate findings and theoretical developments. He illuminates the competing definitions of “language policy,” the historical development of language policy theory, the effects of language policies throughout the world, and language policy research methodology. It is intended as a useful how-to guide for both neophyte researchers interested in taking up a study on language policy and a handy reference for language

policy scholars who want a book that combines the findings in the field. Specific research projects are proposed throughout the book.

McCarty, T. (2011). *Ethnography and language policy*. London: Routledge.

In this edited volume, McCarty has gathered authors who experienced ethnographers of language policy. As its name implies, it is particularly useful for those interested in ethnography but also focuses on discourse analysis and positionality in LPP research. The chapters deal with a wide variety of contexts and policies.

Ricento, T. (Ed.). (2006). *An introduction to language policy: Theory and method*. Malden, MA: Blackwell Publishing.

Ricento's edited volume deals with many of the major theories, and methodological traditions, in language policy research. As its name implies, it is an introductory overview of the field, with brief chapters written succinctly and accessibly. The theoretical contributions are especially robust and provide a valuable resource for both novice and experienced researchers.

## References

- Amir, A., & Musk, N. (2013). Language policing: Micro-level language policy-in-process in the foreign language classroom. *Classroom Discourse, 4*, 151–167.
- Barakos, E. (2016). Language policy and critical discourse studies: Towards a combined approach. In E. Barakos & J. W. Unger (Eds.), *Discursive approaches to language policy* (pp. 23–49). Basingstoke, UK: Palgrave Macmillan.
- Bonacina-Pugh, F. (2012). Researching 'practiced language policies': Insights from conversation analysis. *Language Policy, 11*, 213–234.
- Canagarajah, S., & Stanley, P. (2015). Ethical considerations in language policy research. In F. Hult & D. C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 33–44). Malden, MA: Wiley Blackwell.
- Chimbutane, F. (2011). *Rethinking bilingual education in postcolonial contexts*. Clevedon: Multilingual Matters.
- Cincotta-Segi, A. (2009). 'The big ones swallow the small ones.' Or do they? *The language policy and practice of ethnic minority education in the Lao PDR: A case study from Nalae*. Ph.D. Thesis. Canberra: The Australian National University.
- Cincotta-Segi, A. (2011). Talking in, talking around and talking about the L2: Three literacy teaching responses to L2 medium of instruction in the Lao PDR. *Compare, 41*(2), 195–209.

- Cobarrubias, J. (1983). Ethical issues in status planning. In J. Cobarrubias & J. A. Fishman (Eds.), *Progress in language planning: International perspectives* (pp. 41–85). Berlin: Walter de Gruyter.
- Davis, K. A. (1999). Dynamics of indigenous language maintenance. In T. Huebner & K. A. Davis (Eds.), *Sociopolitical perspectives on language policy and planning in the USA* (pp. 67–98). Amsterdam/Philadelphia: John Benjamins Publishing Company.
- De Costa, P. I. (2010). Language ideologies and standard English language policy in Singapore: Responses of a “designer immigrant” student. *Language Policy*, 9, 217–239.
- De Costa, P. I. (2014). Making ethical decisions in an ethnographic study. *TESOL Quarterly*, 48, 413–422.
- Fishman, J., Ferguson, C. A., & Das Gupta, J. (Eds.). (1968). *Language problems of developing nations*. New York: John Wiley & Sons.
- Fishman, J. A. (1994). Critiques of language planning: A minority languages perspective. *Journal of Multilingual & Multicultural Development*, 15, 91–99.
- Foucault, M. (1991). Governmentality. In G. Burchell, C. Gordon, & P. Miller (Eds.), *The foucault effect: Studies in governmentality* (pp. 87–104). Chicago: The University of Chicago Press.
- Fowler, R., Hodge, B., Kress, G., & Trew, T. (1979). *Language and control*. London: Routledge and Kegan Paul.
- Freeman, R. D. (1998). *Bilingual education and social change*. Clevedon: Multilingual Matters.
- Grin, F. (2003). Language planning and economics. *Current Issues in Language Planning*, 4, 1–66.
- Grin, F., & Vaillancourt, F. (2015). The economics of language policy: An introduction to evaluation work. In F. M. Hult & D. C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 21–32). Malden, MA: Wiley-Blackwell.
- Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Arnold.
- Haugen, E. (1959). Planning for a standard language in Norway. *Anthropological Linguistics*, 1, 8–21.
- Haugen, E. (1983). The implementation of corpus planning: Theory and practice. In J. Cobarrubias & J. A. Fishman (Eds.), *Progress in language planning: International perspectives* (pp. 269–290). Berlin: Walter de Gruyter.
- Heller, M. (2011). *Paths to post-nationalism: A critical ethnography of language and identity*. New York: Oxford University Press.
- Hopkins, M. (2012). Arizona’s teacher policies and their relationship with English learner instructional practice. *Language Policy*, 11, 81–99.
- Hornberger, N. H. (1998). Language policy, language education, language rights: Indigenous, immigrant, and international perspectives. *Language in Society*, 27, 439–458.

- Hornberger, N. H., & Johnson, D. C. (2007). Slicing the onion ethnographically: Layers and spaces in multilingual language education policy and practice. *TESOL Quarterly*, 41, 509–532.
- Hult, F. (2015). Making policy connections across scales using nexus analysis. In F. Hult & D. C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 217–232). Malden, MA: Wiley Blackwell.
- Hult, F. M. (2010). Analysis of language policy discourses across the scales of space and time. *International Journal of the Sociology of Language*, 2010, 7–24.
- Hult, F. M. (2014). Drive-thru linguistic landscaping: Constructing a linguistically dominant place in a bilingual space. *International Journal of Bilingualism*, 18, 507–523.
- Hult, F. M., & Johnson, D. C. (Eds.). (2015). *Research methods in language policy and planning: A practical guide*. Malden, MA: Wiley Blackwell.
- Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth: Penguin Books.
- Johnson, D. C. (2009). Ethnography of language policy. *Language Policy*, 8, 139–159.
- Johnson, D. C. (2011). Critical discourse analysis and the ethnography of language policy. *Critical Discourse Studies*, 8, 267–279.
- Johnson, D. C. (2013). Positioning the language policy arbiter: Governmentality and footing in the School District of Philadelphia. In J. W. Tollefson (Ed.), *Language policies in education: Critical issues* (pp. 116–136). London and New York: Routledge.
- Johnson, D. C. (2015). Intertextuality and language policy. In F. Hult & D. C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 166–181). Wiley Blackwell: Malden, MA.
- Johnson, D. C. (2017). Critical ethnography of language policy: A semi-confessional tale. In M. Martin-Jones & D. Martin (Eds.), *Researching multilingualism: Critical and ethnographic approaches* (pp. 105–120). London: Routledge.
- Johnson, D. C., & Johnson, E. J. (2015). Power and agency in language policy appropriation. *Language Policy*, 14, 221–243.
- Johnson, D. C., & Ricento, T. (2013). Conceptual and theoretical perspectives in language planning and policy: Situating the ethnography of language policy. *International Journal of the Sociology of Language*, 219, 7–21.
- Johnson, E. J. (2012). Arbitrating repression: Language policy and education in Arizona. *Language and Education*, 26, 53–76.
- Karjalainen, M. (2016). Language policy as a sociocultural tool: Insights from the University of Cape Town. *Higher Education Policy*, 29, 131–148.
- King, K. A. (2013). A tale of three sisters: Language ideologies, identities, and negotiations in a bilingual, transnational family. *International Multilingual Research Journal*, 7, 49–65.

- Kloss, H. (1969). *Research possibilities on group bilingualism: A report*. Quebec: International Center for Research on Bilingualism.
- Kochenov, D., & de Varennes, F. (2015). Language and law. In F. M. Hult & D. C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 56–66). Malden, MA: Wiley-Blackwell.
- Lee, T. S. (2009). Language, identity, and power: Navajo and Pueblo young adults' perspectives and experiences with competing language ideologies. *Journal of Language Identity Education*, 8, 307–320.
- Lin, A. M. (2015). Researcher positionality. In F. Hult & D. C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 4–21). Wiley Blackwell: Malden, MA.
- Madison, S. (2005). *Critical ethnography: Method, ethics, and performance*. Thousand Oaks, CA: Sage.
- Martin-Jones, M. (2015). Multilingual classroom discourse as a window on wider social, political and ideological processes. In N. Markee (Ed.), *The handbook of classroom discourse and interaction* (pp. 446–460). Malden, MA: Wiley-Blackwell.
- May, S. (2008). *Language and minority rights: Ethnicity, nationalism, and the politics of language*. New York: Routledge.
- McCarty, T. L. (Ed.). (2011). *Ethnography and language policy*. New York: Routledge.
- McCarty, T. L., Collins, J., & Hopson, R. K. (2011). Dell Hymes and the new language policy studies: Update from an underdeveloped country. *Anthropology and Education Quarterly*, 42, 335–363.
- Menken, K., & García, O. (Eds.). (2010). *Negotiating language policies in schools: Educators as policymakers* (pp. 249–261). London and New York: Routledge.
- Mortimer, K. S. (2013). Communicative event chains in an ethnography of Paraguayan language policy. *International Journal of the Sociology of Language*, 219, 67–99.
- Mortimer, K. S. (2016). Producing change and stability: A scalar analysis of Paraguayan bilingual education policy implementation. *Linguistics and Education*, 34, 58–69.
- Mortimer, K. S., & Wortham, S. (2015). Analyzing language policy and social identification across heterogeneous scales. *Annual Review of Applied Linguistics*, 35, 160–172.
- Pan, L. (2011). English language ideologies in the Chinese foreign language education policies: A world-system perspective. *Language Policy*, 10, 245–263.
- Perez-Milans, M. (2012). “Ah! Spain, that’s far away from China ” Reflexividad metodológica y movilidad en la etnografía sociolingüística crítica. *Spanish in Context*, 9, 219–243.
- Pérez-Milans, M. (2015). Language education policy in late modernity:(Socio) linguistic ethnographies in the European Union. *Language Policy*, 14, 99–107.

- Phyak, P. (2013). Language ideologies and local languages as the medium-of-instruction policy: A critical ethnography of a multilingual school in Nepal. *Current Issues in Language Planning*, 14, 127–143.
- Ramanathan, V. (2006). Of texts AND translations AND rhizomes: Postcolonial anxieties AND deracinations AND knowledge constructions. *Critical Inquiry in Language Studies*, 3, 223–244.
- Ramanathan, V. (2011). Researching-texting tensions in qualitative research: Ethics in and around textual fidelity, selectivity, and translations. In T. McCarty (Ed.), *Ethnography and language policy* (pp. 255–270). New York and London: Routledge.
- Ricento, T. (2000). Historical and theoretical perspectives in language policy and planning. *Journal of Sociolinguistics*, 4, 196–213.
- Ricento, T. K., & Hornberger, N. H. (1996). Unpeeling the onion: Language planning and policy and the ELT professional. *TESOL Quarterly*, 30, 401–427.
- Rubin, J., & Jernudd, B. H. (Eds.). (1971). *Can language be planned? Sociolinguistic theory and practice for developing nations*. Honolulu: The University Press of Hawaii.
- Ruiz, R. (1984). Orientations in language planning. *NABE Journal/National Association for Bilingual Education*, 8, 15–34.
- Shoba, J. A., & Chibbutane, F. (Eds.). (2013). *Bilingual education and language policy in the global south* (Vol. 4). New York and London: Routledge.
- Shohamy, E., & Gorter, D. (Eds.) (2009). *Linguistic landscape: Expanding the scenery*. New York: Routledge.
- Sonntag, S. (2009). Linguistic globalization and the call center industry: Imperialism, hegemony or cosmopolitanism? *Language Policy*, 8, 5–25.
- Stephens, C., & Johnson, D. C. (2015). “Good teaching for all students?”: Sheltered instruction programming in Washington state language policy. *Language and Education*, 29, 31–45.
- Tauli, V. (1974). The theory of language planning. In J. Fishman (Ed.), *Advances in language planning* (pp. 69–78). City: Publisher.
- Tollefson, J. W. (1991). *Planning language, planning inequality: Language policy in the community*. London: Longman.
- Tollefson, J. W. (2002). Introduction: Critical issues in educational language policy. In J. W. Tollefson (Ed.), *Language policies in education: Critical issues* (pp. 3–16). Mahwah, NJ: Lawrence Erlbaum Publishers.
- Tollefson, J. W. (2006). Critical theory in language policy. In T. Ricento (Ed.), *An introduction to language policy: Theory and method* (pp. 42–59). Malden, MA: Blackwell Publishing.
- Tollefson, J. W. (2013). Language policy in a time of crisis and transformation. In J. Tollefson (Ed.), *Language policies in education: Critical issues* (pp. 11–34). New York and London: Routledge.

- Tollefson, J. W. (2015). Historical-structural analysis. In F. M. Hult & D. C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 21–32). Malden, MA: Wiley-Blackwell.
- Wiley, T. G., & Wright, W. E. (2004). Against the undertow: Language-minority education policy and politics in the “age of accountability”. *Educational Policy*, 18, 142–168.
- Wortham, S., & Reyes, A. (2015). *Discourse analysis beyond the speech event*. New York: Routledge.



# 38

## Second Language Pragmatics

Soo Jung Youn

### Introduction

The last three decades have witnessed an increase in studies of second language (L2) pragmatics on a wide range of research issues along with various theoretical and methodological approaches to defining and operationalizing L2 pragmatics. Especially given the broad range of pragmatic targets studied, the critical understanding of theoretical and methodological approaches to L2 pragmatics is crucial in making valid research claims. To this end, this chapter provides an overview of research issues in L2 pragmatic learning, instruction, and assessment, followed by the discussion of data collection approaches in pragmatics research. Two sample studies (Al-Gahtani & Roever, 2012; Bella, 2012) are also discussed to illustrate the distinct data collection and analytic approaches in L2 pragmatics research.

### Defining L2 Pragmatics

Depending on theoretical perspectives and research methods, the definition and scope of L2 pragmatics can vary (Taguchi & Roever, 2017). Among many, a commonly cited definition of pragmatics is “the study of language from the point of view of users, especially of the choices they make, the constraints they

---

S. J. Youn (✉)

Department of English, Northern Arizona University, Flagstaff, AZ, USA

e-mail: [Soo-Jung.Youn@nau.edu](mailto:Soo-Jung.Youn@nau.edu)



encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication” (Crystal, 1997, p. 301). As alluded to in this definition, one’s pragmatic competence is determined by a constellation of elements, such as knowledge of linguistic forms, functions, contexts, norms of interaction, and social relationships among participants. Further, Leech’s (1983) distinction between sociopragmatics and pragmalinguistics offers a discussion of language users’ pragmatic knowledge and ability for use. Sociopragmatic competence refers to knowledge of contextual variables, such as participants’ relative power and social distance among participants, and the degree of imposition involved in speech acts (Brown & Levinson, 1987), which guide the appropriateness of social norms in specific contexts. On the other hand, pragmalinguistics focuses on the interface of pragmatics and conventions of language use. For example, when a student makes a recommendation letter request to a professor during office hours, pragmatically competent language users attend to mappings of these two dimensions of pragmatic competence, such as the relationship between the professor, the imposition, and conventional practices associated with a recommendation letter request (i.e., sociopragmatics) and the conventions of language use in making a request, such as modals and bi-clausal expressions (i.e., pragmalinguistics).

## Research Issues in L2 Pragmatics Research

Theoretical, empirical, and pedagogical interest in L2 pragmatics has resulted in widely established findings, such as the fact that grammatical competence cannot guarantee concomitant pragmatic competence (e.g., Bardovi-Harlig, 1999) and that pragmatic instruction is indeed beneficial (e.g., Taguchi, 2015). Building upon such established findings, this section discusses some research issues in pragmatic learning, instruction, and assessment.

### Learning

In early L2 pragmatics research, cognitive perspectives were dominant, resulting in the coined term interlanguage pragmatics (ILP) (Kasper & Blum-Kulka, 1993). Further, the majority of early pragmatics studies were conceptualized under speech act theory (Searle, 1976) and politeness theory (Brown & Levinson, 1987). Perhaps one of the most prominent early ILP pragmatics research was the Cross-Cultural Speech Act Realization Project (CCSARP) (Blum-Kulka, House, & Kasper, 1989). This project assigned

speech act data into certain categories (e.g., direct vs. indirect requests) and compared such categories between speaker groups. Sample Study 38.1 discussed below shows how pragmatic development can be investigated by drawing on the coding system developed in CCSARP.

### Sample Study 38.1

Bella, S. (2012). Pragmatic development in a foreign language: A study of Greek FL requests. *Journal of Pragmatics*, 44, 1917–1947.

#### Research Background

The study aims to examine developmental patterns of request in Greek foreign language learners with different linguistic backgrounds focusing on the linguistic resources realized in request across proficiency levels.

#### Research Problems/Gaps

Despite the extensive research on L2 learners' request development, the range of target languages investigated in request development remains limited.

#### Research Method

- *Type of research*: Cross-sectional research with qualitative and quantitative data.
- *Setting and participants*: 50 native speakers of Greek, 150 non-native speakers of Greek as a Foreign Language.
- *Instruments*: Discourse completion tasks (DCTs) on 12 different situations consisting of eight requests and four other situations, which functioned as distractors, and the interviews with the participants for verbal report data.
- *Data analysis*: Following the coding approach based on the Cross-Cultural Speech Act Realization Project (CCSARP) (Blum-Kulka et al., 1989), request head acts were coded using three degrees of directness: (a) direct, (b) conventionally indirect, (c) non-conventionally indirect.

#### Key Results

Low-intermediate learners displayed significantly lower frequencies of syntactic modifiers due to their lower degree of grammatical competence, yet displaying sociopragmatic awareness as evidenced in high frequencies of formulaic apologies in high imposition situations. Lower-level learners' requests are characterized as the basic/formulaic stage (e.g., formulaic use of mitigators), compared to those of more advanced learner groups and native speakers. Intermediate-level learner groups exhibited over-suppliance in external modifiers, which is less syntactically demanding yet compensate the lack of lexical and phrasal modification of request satisfying the need of being polite. The most noticeable pragmalinguistic resources that differentiated the learners' levels and native speaker groups involved the frequencies and types of lexical and phrasal modifiers.

#### Comments

It is important to note that the coding approach employed in CCSARP has been heavily criticized for the lack of clear category definitions. Despite such shortcomings, the coding scheme used in CCSARP can still satisfy the purpose of comparisons with previous studies and examine various resources utilized in pragmatic production data (Al-Gahtani & Roever, 2015).

Diverse theoretical approaches have been taken in pragmatic developmental studies (see Kasper, 2009, for review). For example, under the individual cognitive approach to L2 pragmatic development, pragmatics is conceptualized as the knowledge constructed in an individual's cognition. Thus, main research issues include how individual learners develop their ability to comprehend and produce pragmatic meanings (e.g., Taguchi, 2005). The interaction between interlanguage development and pragmatics has also been a longstanding research topic (Bardovi-Harlig, 1999, 2012; Li, 2014; Youn, 2014). Central to such topics, Bardovi-Harlig (2013) emphasizes the importance of employing appropriate research instruments depending on the focus of the study. For example, when investigating implicit and explicit pragmatic knowledge, written production tasks are more appropriate than conversational tasks (e.g., role-plays) because written production tasks (e.g., discourse completion tasks) afford more planning time than spontaneous and real-time role-play interaction.

In tandem with the individual cognitive approach to L2 pragmatics, Kasper (2006) called for a discursive approach to speech act pragmatics employing conversation analysis (CA), which respecifies the nature and scope of pragmatic competence by investigating how participants themselves co-construct pragmatic meanings and actions in talk-in-interaction. Such a call was motivated by the discontent with an *etic* perspective of the coding scheme prevalent in the speech act research tradition. CA examines how participants in social interaction collaboratively accomplish social actions and relations in sequential organizations by employing an *emic* perspective that focuses on how participants themselves display their understandings of their talk (see Kasper & Wagner, 2014, for an introductory discussion of CA). Sample Study 38.2 discussed below shows how requests can be investigated in sequential organizations that draw on the CA approach.

Apart from the individual cognitive approach and the discursive approach to pragmatics research, there are more distinct theoretical approaches (Kasper, 2009), such as sociocultural theories and language socialization, which furnish researchers with rich theoretical approaches to conceptualizing how L2 pragmatic competence develops.

## Instruction

The need for pragmatic instruction is not surprising, especially given that learners face unique challenges in learning opaque pragmatic rules that are context-specific and that mere input exposure itself is not sufficient for

pragmatic learning. Earlier instructional studies focused on the teachability of pragmatics, the benefits of instruction versus exposure, the effect of explicit and implicit instruction (e.g., Rose, 2005), and the role of pragmatic awareness (e.g., Bardovi-Harlig & Griffin, 2005). Thanks to the extensive body of instructional studies, the teachability of pragmatics and the benefits of pragmatic instruction have been well established (e.g., Jeon & Kaya, 2006; Taguchi, 2015).

Rose and Kasper (2001) argued that most aspects of pragmatics are amenable to instruction. The pragmatic targets focused in previous instructional studies vary, ranging from speech acts, such as academic e-mail requests (Nguyen, Do, & Nguyen, 2015), to job interview skills (Louw, Derwing, & Abbott, 2010) and sarcasm (Kim, 2014) (see Taguchi, 2015 for a list pragmatic targets in instructional studies). Depending on pragmatic targets, the effectiveness of teaching might differ. Linguistic simplicity and straightforward pragmatic rules attached to pragmatic targets, such as apology, can be more amenable to instruction, whereas more opaque pragmatic rules can be considerably challenging for learners. Taguchi's (2015) comprehensive review of 58 instructional intervention studies indicates that, overall, the positive effect of explicit instruction was reported and optimal instruction conditions for effective pragmatic instruction include "(1) explicit teaching with metapragmatic information and opportunities to produce target pragmatic forms; and (2) implicit teaching involving structured practice for processing pragmatic rules" (p. 29). In order to maximize such instructional conditions, further pedagogical attention can be directed toward strategy instruction (Taguchi, 2011), raising learners' awareness of pragmalinguistics and socio-pragmatics, and providing opportunities through which learners can reflect on cross-cultural differences (Ishihara & Cohen, 2010).

While the teachability and the benefits of pragmatic instruction are well established, Taguchi (2015) points out the need to ensure the validity of previous findings in instructional studies, that is the interaction between task characteristics and instructional effects. Depending on the choice of external measures employed in instructional studies, the effect of instruction can be different. For written DCTs, learners can spend more time to plan the response than role-plays that require spontaneous interaction with greater processing demands and monitoring skills. Thus, the instructional effect may appear more strongly on the task that does not have much processing demands, such as written DCTs. Since the effect of instruction can surface differently depending on the external measures used, researchers need to carefully select the targeted instruments that align with the nature of pragmatic targets.

## Assessment

Compared to the research on pragmatic learning and instruction, L2 pragmatic assessment research is relatively new, but has been growing along with the increased interest in developing external measures for instructional studies. Further, in large-scale language proficiency tests, the importance of assessing pragmatic competence as part of one's academic language proficiency has been increasingly recognized. As with other assessment practices, ensuring validity and reliability is also a pressing concern in assessing pragmatic competence. To this end, important issues include specifying the scope of pragmatic competence and operationalizing the selected aspects of pragmatic competence in relation to score uses and interpretation.

Given the broad range of pragmatic targets, previous pragmatic assessment research differs in terms of the coverage of pragmatic competence. In doing so, a constant tension between construct representation and practicality is observed (Roever, 2011). Early studies dealt with classic speech acts. For example, the pioneering work by Hudson, Detmer, and Brown (1995) focused on developing test instruments to assess three speech acts (request, refusal, apology) by manipulating three sociological variables from politeness theory (i.e., social distance, power distance among participants, ranking of imposition). Hudson et al.'s test instrument has been also translated into different languages, such as Japanese as a Foreign Language (Yamashita, 1996) and Korean as a Foreign Language (Youn & Brown, 2013). Despite its contribution, the construct coverage in Hudson et al. was limited to three speech acts. The underemphasis of pragmalinguistics was addressed in Roever (2005), who developed a web-based test of pragmalinguistics focusing on the comprehension of implicature, recognition of routine formulas, and knowledge of speech act strategies, which has broadened the coverage of pragmatic competence. The next generation of pragmatic assessment research has covered more highly interactive discourse in pragmatic assessment (e.g., Grabowski, 2009; Ross & Kasper, 2013; Walters, 2007; Youn, 2015). What is distinguishable in this line of work is the detailed qualitative attention to capturing how participants accomplish pragmatic actions in spoken interaction. Qualitative analyses, such as a CA approach, were employed to delineate the detailed features of pragmatic performance in interaction, which in turn function as key validity evidence.

In terms of operationalizing the pragmatic targets, various item formats have been employed. Hudson et al. (1995), for example, used DCTs, role-plays, self-assessment, and multiple-choice items. In spite of their practicality,

multiple-choice items were found to be difficult to obtain high reliabilities due to the challenge in developing valid distracters. Nonetheless, Liu's (2006) study shows that a careful test development approach to designing multiple-choice items can result in a high reliability of around 0.9. In order to measure more interactive aspects of pragmatic competence, both Grabowski (2009) and Youn (2015) employed role-plays. While role-plays can capture a range of conversational practices, the practicality of developing rating criteria and training raters to ensure reliability and validity is a main limitation. Further, in more complex pragmatic performance assessment, a considerable amount of effort to examine its validity needs to be made. Taken together, researchers should not choose research instruments based solely on convenience or convention. Rather, they should consider the extent to which the assessed aspects of pragmatic competence are aligned with the characteristics of research instruments. The next section further discusses various data collection approaches available in L2 pragmatics research.

## Data Collection Approaches in L2 Pragmatics Research

In the previous section, some research issues in the areas of L2 pragmatic learning, instruction, and assessment were discussed. For all areas, the crucial role of employing valid data collection approaches cannot be overemphasized. Various data collection methods in L2 pragmatics research have been well documented (e.g., Félix-Brasdefer, 2010; Kasper & Dahl, 1991; Kasper & Roever, 2005; Kasper & Rose, 2002). Table 38.1 summarizes the data collection methods in L2 pragmatics research which were discussed in Kasper and Rose (2002). These methods range from real-life authentic discourse to highly controlled production tasks. Three broad categories are discussed: (a) spoken interaction (authentic discourse, elicited conversation, role-plays), (b) questionnaires (discourse completion tasks, multiple choice, scaled response), and (c) self-report (interviews, think-aloud protocols, diaries). Challenges between highly controlled production tasks, which elicit comparable language features, and authentic discourse are not new in L2 pragmatics research. For the purposes of investigating a wide range of discourse features in spoken interactive productions, authentic discourse or role-plays are appropriate choices. However, given the challenge of eliciting constrained data from authentic discourse, talk elicited through conversation tasks and role-plays are useful alternatives to authentic interaction. For well-defined and predetermined aspects

**Table 38.1** Summary of data collection methods in L2 pragmatics research as discussed by Kasper and Rose (2002)

Categories		Description
Spoken interaction	Authentic discourse	Authentic discourse data include audio- and video-recorded interaction in any authentic settings or institutional encounters. While authentic discourse offers valid records of learners' pragmatic competence, comparing specific pragmatic practices of ordinary conversation across settings can be challenging, whereas the relatively recurrent and structured organization of institutional interaction allows comparisons of pragmatic actions and practices across settings and participants.
	Elicited conversation	Elicited conversation refers to data elicited from conversation tasks for the purpose of data collection without specifying different social roles to participants. While researchers can gain more control over elicited conversation, elicited conversation can be limited to a fairly narrow range of pragmatic practices since participants' social roles are not being manipulated.
	Role-plays	Role-plays are simulations of communicative encounters based on specified roles assumed for participants. As an alternative to authentic discourse and elicited conversation, researchers can focus on specific pragmatic actions and can manipulate participants' roles in role-plays under controlled conditions. Format of role-plays can vary, ranging from a single-turn response (closed role-plays) to interaction over numerous turns (open role-plays).

of pragmatic competence, questionnaires may be appropriate, whereas exploratory research purposes may require open-ended and participant-directed methods, such as self-reports. The decision of choosing data elicitation tools should match the research purpose. Given their wide uses, DCTs and role-plays, each representing the questionnaire and spoken interaction categories, are further discussed below along with an example study that illustrates each instrument type and how pragmatic productions are analyzed.

## Discourse Completion Tasks

One of the most widely used data collection methods in L2 pragmatics research is the DCT. The rich potential of DCTs is evident from its longstanding use in L2 pragmatics research in a wide range of speech acts. The typical

DCT format includes a situational description and an open slot to be completed by participants. In designing DCTs, three sociological variables, social distance (D), power distance among participants (R), and ranking of imposition (I), are often manipulated, as shown in the example DCTs in Sample Study 38.1 (Bella, 2012). Depending on the types of participants' responses, either written or oral DCTs can be designed. While DCTs can ensure relatively standardized and constrained responses from participants, researchers must be sure to use DCTs that fit their research purposes. As evidenced in the comparative research of written DCTs and authentic data (e.g., Golato, 2003), written DCTs are more appropriate for investigating speech act strategies and linguistic resources in realizing various speech acts. In addition, DCTs data elicit intuitional data rather than data on actual language use (Kasper & Rose, 2002).

Sample Study 38.1 conducted by Bella (2012) illustrates the use of DCTs in analyzing the production of request following the coding approach based on the Cross-Cultural Speech Act Realization Project (CCSARP) (Blum-Kulka et al., 1989).

## Role-Plays

As an alternative to authentic discourse and elicited conversation, role-plays have been widely employed when researchers want to focus on how participants produce and understand pragmatic actions in spoken interaction. In role-plays, researchers can focus on various conversational practices and manipulate participants' social roles under more controlled conditions, thereby allowing comparisons of specific aspects of pragmatic competence across participants and settings.

In designing role-plays, researchers can manipulate contextual variables, the roles designed for participants, the situational context in which the interaction occurs, and the communicative purpose of the interaction, as seen in the example role-plays in Sample Study 38.2.

Kasper and Dahl (1991) made a distinction between open- and closed role-plays depending on the degree of affordance in interaction. Closed role-plays elicit participants' reaction to an interlocutor's standardized and scripted initiation in a single turn. Thus, closed role-plays involve a minimal amount of interaction following a predetermined interactional outcome. On the other hand, open role-plays provide opportunities for participants to engage in interaction without predetermined interactional outcomes over multiple turns and different discourse management phases.



While open role-plays can elicit various conversational features, the extent to which they provide valid phenomenon of conversational practices in authentic contexts needs to be critically examined. Commonly perceived notions of elicited spoken interaction include few or no social consequences for the participants and the lack of generalizability as a research instrument (e.g., Bardovi-Harlig & Hartford, 2005; Gass & Houck, 1999). Toward such preconceived notions, the CA approach presents rigorous evidence of the nature of the elicited interaction. Huth (2010) offers a convincing account of how the elicited interaction can be socially consequential by showing the fundamental aspects of interaction that role-play participants draw on that are similar to those of naturally occurring conversation. An increasing line of CA-based studies on role-plays employed in institutionalized settings, such as language assessment, further confirms the potentials of role-plays as valid ways to examining conversational practices reflective of those in real-life contexts (e.g., Kasper & Youn, 2017; Okada, 2010). In particular, when the targeted interaction occurs in an institutional setting, the highly systematic and recurrent nature of institutional encounters generates opportunities to observe specific features across participants. Admittedly, even though the elicited role-play interaction will not be the same as those of authentic discourse, the elicited conversation via role-plays can be a useful alternative to authentic discourse that ensures some degree of authenticity and standardization.

As an example study of role-plays, Al-Gahtani and Roever (2012) is summarized below.

### Sample Study 38.2

Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics*, 33, 42–65.

#### Research Background

The study aims to examine how participants' L2 proficiency affects the interactional resources utilized in extended real-time request discourse.

#### Research Problems/Gaps

The previous research on L2 requests in developmental pragmatics commonly employed noninteractive data collection methods, such as discourse completion tasks.

#### Research Methods

- *Type of research*: Qualitative research using a cross-sectional research design.
- *Setting and participants*: 26 Saudi learners of Australian English at four distinct proficiency levels.

- *Instruments*: The role-plays of three request situations that differ in terms of the power distance among participants with low imposition and low social distance.
- *Data analysis*: A conversation analytic method was employed to examine how request sequences are co-produced by participants and how they systematically differ at different L2 proficiency levels. The analytic focus of the role-play data was to describe how participants achieve their request step by step in sequential organizations.

### Key Results

Distinct effects of participants' proficiency levels were apparent in the sequential organizations of learners' requests. The placement of pre-expansions (i.e., an optional sequence that precedes a first-pair part of request, such as how-are-you exchanges) differed noticeably between lower- and higher-level groups. Higher-level learners were more likely to supply supportive moves of pre-expansions before the actual request compared to lower-level learners. When leading the conversation, higher-level participants used a more proactive approach when interacting with an interlocutor, such as adjusting their production depending on the likelihood of acceptance, while the lower-level groups often relied on the interlocutor who elicited background information needed in making requests.

### Comments

Al-Gahtani and Roever (2012) show how researchers can analyze ways in which participants accomplish pragmatic actions in sequential organizations across multiple turns drawing on a CA approach. Taken together, the two sample studies above illustrate distinct data collection methods, which influence the language that is produced, and analytical approaches in analyzing the request data accordingly. While distinct in terms of their findings, the two studies offer a complementary account of the developmental patterns of L2 requests. The nature of elicited data from DCTs and role-plays needs to be clearly understood. Depending on the scope of pragmatic competence, researchers need to choose the most appropriate data collection methods and analytical frameworks. The decision of choosing data collection methods have to align with the theoretical stances adopted in a study so that researchers optimally meet the research purpose. Finally, misalignments among the theoretical approaches, the pragmatic targets, and the instruments used as external measures can seriously jeopardize the validity of research claims.

## Challenges and Future Directions

The boundaries of pragmatic competence are often not clear cut. The broad coverage of L2 pragmatics can present inherent challenges in narrowing down the scope of pragmatic competence and choosing appropriate research instruments. Further, the degree of appropriateness in L2 pragmatics is dependent on contexts, language users, and the target culture, which makes the mapping of language forms and functions more opaque. Future research on investigating context-specific pragmatic production data such as grammatical and interac-

tional features that employ either corpus linguistic approaches to academic language or a CA approach to institutional encounters can offer some references.

In terms of choosing data collection methods and sites, recent advancements in technology have expanded the scope of pragmatic competence and the nature of data collection approaches (e.g., Belz, 2007; Roever, Fraser, & Elder, 2014; Taguchi & Sykes, 2013). Learners' pragmatic production data can be collected in digitally mediated discourse, such as blogging, mobile games, and social networks (e.g., Eslami, Mirzaei, & Dini, 2015). The use of technology-enhanced authentic tasks is a noticeable trend in instructed pragmatic studies, including telecollaborations, wikis, video conferencing, and online discussions in virtual environments. For example, Sykes (2013) created a 3D space that simulates a Spanish-speaking interaction with computer-generated avatars to practice requests and apologies. Sydorenko (2015) also developed computer-delivered structured tasks in exploring the effect of oral practice pragmatic instruction. As seen in these examples, the integration of technology in L2 pragmatics research will be an important part of future research programs.

Context-specific research issues in pragmatic learning, instruction, and assessment need to be addressed. The learning environments that influence pragmatic developments go beyond classroom contexts, ranging from study abroad in English as lingua franca to various institutional settings. The unique benefits of accessing recurrent and systematic pragmatic production in institutional settings or ample pragmatic learning opportunities in study abroad contexts (e.g., Alcón-Soler, 2015; Hassall, 2015; Schauer, 2009) will continue to attract researchers' interests. Further, the emerging interest in applying a task-based language teaching approach to L2 pragmatics research (Taguchi & Kim, 2018) will generate various opportunities to investigate pragmatic instruction and learning issues in various real-life contexts. Taken together, diverse theoretical approaches and instructional frameworks will enable us to understand the complex interplay of the different elements underlying pragmatic competence, which continues to provide promises for future exploration.

## Resources for Further Reading

Félix-Brasdefer, J. C. (2015). *The language of service encounters: A pragmatic-discursive approach*. Cambridge: Cambridge University Press.

Drawing on a pragmatic-discursive approach, this volume offers a comprehensive account of service encounters using naturally occurring face-to-face

interaction data on different pragmatic practices in sequential organizations, prosody, and stylistic levels.

Ishihara, N., & Cohen, A. D. (2010). *Teaching and learning pragmatics: Where language and culture meet*. Harlow, UK: Pearson Longman.

This volume offers research-informed practical resources for teachers and teacher educators with various hands-on activities on diverse aspects of pragmatics in both written and spoken discourse.

Taguchi, N., & Sykes, J. (Eds.). (2013). *Technology in interlanguage pragmatics research and teaching*. Philadelphia: John Benjamins.

This edited volume discusses how recent advancements in technology have influenced pragmatics research and teaching practices.

Taguchi, N., & Roever, C. (2017). *Second language pragmatics*. Oxford and New York: Oxford University Press.

This volume offers a comprehensive review of L2 pragmatics research on learning, instruction, and assessment over three decades from diverse theoretical perspectives.

## References

- Alcón-Soler, E. (2015). Pragmatic learning and study abroad: Effects of instruction and length of stay. *System*, 48, 62–74.
- Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics*, 33, 42–65.
- Al-Gahtani, S., & Roever, C. (2015). The development of requests by L2 learners of modern standard Arabic: A longitudinal and cross-sectional study. *Foreign Language Annals*, 48, 570–583.
- Bardovi-Harlig, K. (1999). The interlanguage of interlanguage pragmatics: A research agenda for acquisitional pragmatics. *Language Learning*, 49, 677–713.
- Bardovi-Harlig, K. (2012). Formulas, routines, and conventional expressions in pragmatics research. *Annual Review of Applied Linguistics*, 32, 206–227.
- Bardovi-Harlig, K. (2013). Developing L2 pragmatics. *Language Learning*, 63, 68–86.
- Bardovi-Harlig, K., & Griffin, R. (2005). L2 pragmatic awareness: Evidence from the ESL classroom. *System*, 30, 401–415.

- Bardovi-Harlig, K., & Hartford, B. S. (Eds.). (2005). *Interlanguage pragmatics: Exploring institutional talk*. Mahwah, NJ: Erlbaum.
- Bella, S. (2012). Pragmatic development in a foreign language: A study of Greek FL requests. *Journal of Pragmatics*, *44*, 1917–1947.
- Belz, J. A. (2007). The role of computer mediation in the instruction and development of L2 pragmatic competence. *Annual Review of Applied Linguistics*, *27*, 45–75.
- Blum-Kulka, S., House, J., & Kasper, G. (Eds.). (1989). *Cross-cultural pragmatics: Requests and apologies*. Norwood, NJ: Ablex.
- Brown, P., & Levinson, S. D. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.
- Crystal, D. (1997). *A dictionary of linguistics and phonetics*. Oxford: Blackwell.
- Eslami, Z. R., Mirzaei, A., & Dini, S. (2015). The role of asynchronous computer mediated communication in the instruction and development of EFL learners' pragmatic competence. *System*, *48*, 99–111.
- Félix-Brasdefer, J. C. (2010). Data collection methods in speech act performance: DCTs, role plays, and verbal reports. In E. Usó Juárez & A. Martinex-Flor (Eds.), *Speech act performance: Theoretical, empirical, and methodological issues*. Amsterdam: John Benjamins.
- Gass, S. M., & Houck, N. (1999). *Interlanguage refusals: A cross-cultural study of Japanese-English*. Berlin: Mouton.
- Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics*, *24*, 90–121.
- Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking*. Unpublished Ph.D. dissertation, Columbia University.
- Hassall, T. (2015). Individual variation in L2 study-abroad outcomes: A case study from Indonesian pragmatics. *Multilingua*, *34*, 33–59.
- Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Technical Report #7). Honolulu, HI: University of Hawai'i at Manoa, Second Language Teaching and Curriculum Center.
- Huth, T. (2010). Can talk be inconsequential? Social and interactional aspects of elicited second-language interaction. *Modern Language Journal*, *94*, 537–553.
- Ishihara, N., & Cohen, A. D. (2010). *Teaching and learning pragmatics: Where language and culture meet*. Harlow, UK: Pearson Longman.
- Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 165–211). Amsterdam: Benjamins.
- Kasper, G. (2006). Speech acts in interaction: Towards discursive pragmatics. In K. Bardovi-Harlig, J. C. Félix-Brasdefer, & A. S. Omar (Eds.), *Pragmatics and language learning* (Vol. 11, pp. 281–314). University of Hawai'i at Manoa, HI: National Foreign Language Resource Center.

- Kasper, G. (2009). L2 pragmatic development. In W. C. Ritchie & T. K. Bhatia (Eds.), *New handbook of second language acquisition*. Leeds, UK: Emerald.
- Kasper, G., & Blum-Kulka, S. (1993). Interlanguage pragmatics: An introduction. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatics* (pp. 3–17). New York: Oxford University Press.
- Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition*, 13, 215–247.
- Kasper, G., & Roever, C. (2005). Pragmatics in second language learning. In E. Hinkel (Ed.), *Handbook of research in second language learning and teaching* (pp. 317–334). Mahwah, NJ: Earlbaum.
- Kasper, G., & Rose, K. R. (2002). *Pragmatic development in a second language*. Oxford: Blackwell.
- Kasper, G., & Wagner, J. (2014). Conversation analysis in applied linguistics. *Annual Review of Applied Linguistics*, 34, 171–212.
- Kasper, G., & Youn, S. J. (2017). Transforming instruction to activity: Roleplay in language assessment. *Applied Linguistics Review*. Advance online publication. <https://doi.org/10.1515/applirev-2017-0020>
- Kim, J. (2014). How Korean EFL learners understand sarcasm in L2 English. *Journal of Pragmatics*, 60, 193–206.
- Leech, G. (1983). *Principles of pragmatics*. London: Longman.
- Li, S. (2014). The effects of different levels of linguistic proficiency on the development of L2 Chinese request production during study abroad. *System*, 45, 103–116.
- Liu, J. (2006). *Measuring interlanguage pragmatic knowledge of EFL learners*. Frankfurt am Main: Peter Lang.
- Louw, K. J., Derwing, T. M., & Abbott, M. L. (2010). Teaching pragmatics to L2 learners for the workplace: The job interview. *The Canadian Modern Language Review*, 66, 739–758.
- Nguyen, T. T. M., Do, T. T. H., & Nguyen, A. T. (2015). Teaching email requests in the academic context: A focus on the role of corrective feedback. *Language Awareness*, 24, 169–195.
- Okada, Y. (2010). Role-plays in oral proficiency interviews: Interactive footing and interactional competencies. *Journal of Pragmatics*, 42, 1647–1668.
- Roever, C. (2005). *Testing ESL pragmatics*. Frankfurt, Germany: Peter Lang.
- Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing*, 28, 463–481.
- Roever, C., Fraser, C., & Elder, C. (2014). *Testing ESL sociopragmatics: Development and validation of a web-based test battery*. Frankfurt am Main: Peter Lang.
- Rose, K. R. (2005). On the effects of instruction in second language pragmatics. *System*, 33, 385–399.
- Rose, K. R., & Kasper, G. (Eds.). (2001). *Pragmatics and language teaching*. Cambridge: Cambridge University Press.
- Ross, S., & Kasper, G. (Eds.). (2013). *Assessing second language pragmatics*. Basingstoke, UK: Palgrave Macmillan.

- Schauer, G. (2009). *Interlanguage pragmatic development: The study abroad context*. London: Continuum.
- Searle, J. (1976). A classification of illocutionary acts. *Language in Society*, 5, 1–23.
- Sydorenko, T. (2015). The use of computer-delivered structured tasks in pragmatic instruction: An exploratory study. *Intercultural Pragmatics*, 12, 333–362.
- Sykes, J. (2013). Multiuser virtual environments: Learner apologies in Spanish. In N. Taguchi & J. Sykes (Eds.), *Technology in interlanguage pragmatics research and teaching* (pp. 71–100). Philadelphia: John Benjamins.
- Taguchi, N. (2005). Comprehension of implied meaning in English as a second language. *Modern Language Journal*, 89, 543–562.
- Taguchi, N. (2011). Teaching pragmatics: Trends and issues. *Annual Review of Applied Linguistics*, 31, 289–310.
- Taguchi, N. (2015). Instructed pragmatics at a glance: Where instructional studies were, are, and should be going. *Language Teaching*, 48, 1–50.
- Taguchi, N., & Kim, Y. (Eds.). (2018). *Task-based approaches to teaching and assessing pragmatics*. Amsterdam: John Benjamins.
- Taguchi, N., & Sykes, J. (Eds.). (2013). *Technology in interlanguage pragmatics research and teaching*. Philadelphia: John Benjamins.
- Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing*, 24, 155–183.
- Yamashita, S. O. (1996). *Six measures of JSL pragmatics* (Technical Report #14). Honolulu, HI: University of Hawai'i at Manoa, Second Language Teaching & Curriculum Center.
- Youn, S. J. (2014). Measuring syntactic complexity in L2 pragmatic production: Investigating relationships among pragmatics, grammar, and proficiency. *System*, 42, 270–287.
- Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, 32, 199–225.
- Youn, S. J., & Brown, J. D. (2013). Item difficulty and heritage language learner status in pragmatic tests for Korean as a foreign language. In S. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 98–123). Basingstoke, UK: Palgrave Macmillan.



# 39

## Language Testing and Assessment

April Ginther and Kyle McIntosh

### Introduction

Language testing is a relatively young, but rich, academic discipline within applied linguistics that is concerned with the valid and reliable representation, measurement, and evaluation of language proficiency in tests and assessments. Although it pertains to both first and second language acquisition, the field has developed primarily in relation to the latter, particularly with regard to the teaching and learning of English as a second language (ESL). The object of inquiry may be described alternately as linguistic knowledge, language ability, communicative competence, aptitude, performance, or achievement. Each of these terms is embedded in its own history, traditions, ideologies, and arguments about what constitutes the best starting point for test development and research. These terms also represent domains of inquiry that can be parsed into skills (reading, writing, listening, speaking), modalities (receptive/productive), linguistic subdomains (grammar, pronunciation, vocabulary), and/or subcompetencies (grammatical, sociolinguistic, discourse, pragmatic, strategic).

In this chapter, we use the term “language proficiency” to describe the area of inquiry. Rather than offering a definition, however, we will discuss the

---

A. Ginther (✉)

Department of English, Purdue University, West Lafayette, IN, USA

e-mail: [aginther@purdue.edu](mailto:aginther@purdue.edu)

K. McIntosh

Department of English and Writing, University of Tampa, Tampa, FL, USA

e-mail: [kmcintosh@ut.edu](mailto:kmcintosh@ut.edu)



ways in which researchers and test developers have approached the problem of definition. Since tests and assessments are representations and operationalizations of the underlying construct of language proficiency, what is selected for inclusion and—of equal importance—exclusion can provide a working definition of the construct.

Language is complex, and few, if any, of the skills, subdomains, or subcompetencies mentioned above have clear boundaries. Depending on the task, they can be expected to interact. Understanding these interactions is essential to language testing; we want to know how reading relates to writing, how structure relates to vocabulary, how grammatical competence relates to strategic competence, and so on. Nevertheless, our ability to represent and capture these relationships on any one test or assessment is limited. While the physical properties of objects can be measured directly, the interest of educational and psychological measurements lies in the representation and quantification of abilities that can only be captured indirectly. We are constrained not only by limited access to underlying mental representations and cognitive processes but also by the practical constraints (e.g., time, cost) associated with test administration.

Before declaring the task impossible, however, one should consider the rewards of engagement. As Spolsky (1995) noted:

Language testing is of particular interest because of the various competing factions that contribute to it. One of the reasons for my continuing fascination has been the way that it constantly forces practical and theoretical issues into fruitful tension. The needs of the tester regularly challenge the theorist, just as the findings of the theorist repeatedly tempt the tester. While it is fairly easy to come up with new assessment procedures, it remains difficult to explain exactly what is being measured, a situation that guarantees continuing productive stress. As if this first cause of strain were not enough, there is a second one provided by the fact that at least two disciplines have proprietary claims behind language testing: both language learning theorists and measurement experts have their own independent (and perhaps unresolvable) notions of what is involved. It is a field whose lush complexity promises stimulating exploration. (pp. 3–4)

For researchers, language testing actually suffers from an embarrassment of riches, but its saving grace is that it is grounded in—and by—practice. Furthermore, testing is ubiquitous and, for the foreseeable future, will continue to have consequences for learners, teachers, parents, administrators, policymakers, governments, employers, and employees alike.

## Historical Development

Spolsky (1981, 1995) divided the history of language testing into three periods: the prescientific, the psychometric-structuralist, and the psycholinguistic-sociolinguistic. While each period can be understood as a repudiation of the one that preceded it, together they reflect an expansion of the representation of language proficiency and the domain of inquiry. The prescientific is characterized by a reliance on continuous examination by both the learner and the teacher, with priority given to the judgment of a single examiner. The scientific-structuralist, finding fault with the narrow scope of a single teacher/examiner and the different standards that individual evaluators apply, placed emphasis on reliability (i.e., the consistency of scores across separate administrations or ratings), with priority given to objectivity over subjective judgment. The psycholinguistic-sociolinguistic, again finding fault with the narrow representation of language proficiency in the previous period, shifted away from reliability to broader concerns entailed in validity (i.e., the extent to which scores can be argued to represent an underlying construct), with priority given to the communicative functions of successful language use. Of these three periods, Morrow (1981) concluded:

We might characterize these in turn as the Garden of Eden, the Vale of Tears, and the Promised Land, and different tests (indeed different parts of the same test) can usually be seen to relate to one or other of these stages. (p. 10)

Spolsky's (1995) division of the history of language testing into discrete periods is often cited because it is useful for capturing the shifting priorities that have taken place. However, while values and domains of inquiry have changed over time, the primary methods associated with these periods (e.g., individual examination via interviews, multiple-choice formats in large-scale exams, the use of raters to evaluate written or spoken performance) can still be found within a single assessment or across tests used for particular purposes. Indeed, some aspects of language proficiency are best measured objectively, while others need to be evaluated by raters. In other words, language proficiency is something that can and *must* be measured *and* judged.

The issues raised by approaches associated with these three periods influenced the development of the Test of English as a Foreign Language (TOEFL) and continue to resonate throughout its subsequent revisions. The works of Robert Lado, a linguist, and John Carroll, a psychologist, represent an important juncture. Lado's (1961) *Language Testing: The Construction and Use of Foreign Language Tests* championed the psychometric-structuralist approach

by arguing that it is possible to identify and then develop a representative sample of structural and phonological items based on the similarities or differences between examinees' first and second languages. By carefully selecting (i.e., sampling) a set of items that target particular structural characteristics, identified on a continuum from easy to difficult, for the languages involved, Lado claimed that tests could be constructed to display the candidate's mastery of the second language.

That same year, Carroll's (1961) paper *Fundamental Considerations in Testing for English Language Proficiency of Foreign Students* was delivered at a conference sponsored by the Center for Applied Linguistics (CAL), the Institute of International Education, and the National Association of Foreign Student Advisors (NAFSA), which had been organized to address the perceived need for a test of English language proficiency for university admissions. Carroll contrasted the discrete-structuralist approach, represented by Lado (1961), with his own integrative approach that emphasized productive and receptive modes, "real world" settings, and communicative purposes. Rather than one approach replacing the other, however, this distinction captured what Carroll (1986) later deemed an important expansion of the concept of language proficiency.

While both the discrete-structuralist and integrative approaches to language testing influenced the development of the TOEFL, the former dominated in its earliest and longest-standing version: the paper-based test or TOEFL PBT. The first version consisted of 270 multiple-choice questions divided into four subsections: listening comprehension (requiring examinees to select the most appropriate paraphrase of a series of statements and answer questions about short interchanges between two speakers and a short lecture), reading comprehension (with items based on a series of short passages), vocabulary (selecting the best word for sentence completion or selecting synonyms), and writing (recognition of awkward sentences and inappropriate style). It is important to note that the multiple-choice format is not synonymous with discrete-point testing. Discrete-point items are focused on specific structural points of language (e.g., verb tense and aspect), while certain types of multiple-choice items can be integrative if they require examinees to venture beyond structure and consider broader discourse or pragmatic domains (e.g., main idea reading comprehension items). Despite this qualification, the first version of the TOEFL was decidedly discrete-structuralist in its orientation.

A fuller representation of Carroll's integrative approach would appear in 2005 with the introduction of the TOEFL internet-based test or TOEFL iBT, which incorporated the formerly stand-alone add-ons of writing and speaking

(TWE and TSE). The actual performance of these skills has become part of the standard test administration, in sharp contrast to the absence of speaking and the indirect measurement of writing with multiple-choice items in the earliest versions of TOEFL PBT. In addition, the writing and speaking subsections now include “integrated” tasks (i.e., examinees incorporate information presented in readings and lectures), along with independent items (i.e., examinees respond to a single source of information in the prompt). Productive writing and speaking tasks are rated by at least one human rater. The TOEFL iBT, while maintaining multiple-choice items in its listening and reading subsections, now leans more in the direction of Carroll’s nearly 50-year-old integrated and communicative approach. Perhaps the next revision will include interactive tasks (Davis, Laughlin, Gu, & Ockey, 2016), another step toward testing communication in “real-life” settings.

## Core Concepts

### Validity

In the first issue of the journal *Language Testing*, Davies (1984, p. 50) observed that language testing is a domain that draws heavily upon issues and methods central to psychological testing and educational measurement. Indeed, a working knowledge of the technical aspects of measurement is fundamental to the construction and analysis of tests in general, but the greater influence comes from the theoretical side.

The *Standards for Educational and Psychological Testing* are a product of the American Educational Research Association, the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). They were first published in 1966 and are revised as needed, with new editions appearing approximately every ten years. They include updates on current areas of concern (e.g., the 2014 *Standards* include a section on technology). More importantly, the revisions often present major theoretical shifts regarding measurement.

Prior to 1985, received wisdom identified four central types of validity: face validity, content validity, criterion-related validity, and construct validity. Of these types, *face validity*, or the appearance of a relationship between a test and the subject matter it is meant to represent, provides the least credible evidence for establishing validity. *Content validity*, like face validity, involves the relationship between the content of the test and the area tested but relies on empirical evidence and expert judgment rather than appearances. *Criterion-related validity*

involves the relationship of a particular test with other measures indicative of a similar ability (i.e., criteria). *Construct validity* is evaluated by investigating the underlying qualities that a test can be said to measure; it is an attempt to determine the degree to which selected concepts represent underlying constructs. For example, a test of grammatical competence might include items related to the tense, aspect, mood, and modality of the verbal system of a language, but could underrepresent the construct if limited only to verbal forms.

The 1985 *Standards* plainly state: “Validity [...] is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores” (p. 9). This theoretical shift reflected broader changes in thinking that had occurred in psychology and linguistics. Often described as the “cognitive revolution,” it was marked by the repudiation of behaviorism in psychology and structuralism/empiricism in linguistics (see Chomsky, 1957, 1965; Gardener, 1985; Miller, 2003). Messick (1975) stressed the importance of the relationship between testing and assessment on one hand and validation on the other: “A measure estimates how much of something an individual displays or possesses. The basic question is, What is the nature of that something?” (p. 957). Cronbach (1984) further emphasized validation studies when he said, “The end goal of validation is explanation and understanding. Therefore the profession is coming around to the view that all validation is construct validation” (p. 126). The good news for researchers is that the validation of a test is a process that can be applied to practically all types of data at any stage of development or use.

Messick (1989) presented guidelines for sorting validity evidence. While he argued that validity concerns are best addressed under the auspices of the unifying concept of construct validity, his matrix showed how evidence could be broken down into components associated with test score interpretation and use. Finally, validity is not something that can ever be established once and for all. Angoff (1988) is often quoted as saying, “Construct validation is a process, not a procedure; and it requires many lines of evidence [...]” (p. 27). Despite this emphasis on process, however, current conceptualizations of validity theory emphasize explication and, to a certain extent, procedure. Michael Kane has become the most widely recognized proponent of the argument-based approach to validation, which dominates current discussions of validity theory:

An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations

and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. (Kane, 2013, p. 1)

The argument-based approach draws from the work of British philosopher Stephen Toulmin (1958, 2001), who was famous for presenting the relationships between claims, data, and warrants as the basis for evaluating the quality of an argument. One of the best demonstrations of this approach can be found in Chapelle, Enright, and Jamieson's (2008) validation of the TOEFL iBT (see also Chapelle, Chung, Hegelheimer, Pendar, & Xu, 2010). Nevertheless, language testing remains concerned with the "nature of that something" and whatever our changing, developing, and expanding representations of the construct of language proficiency might entail.

## The Unitary Trait Hypothesis

Language testing research in the 1970s was so preoccupied with John Oller's Unitary Trait Hypothesis and cloze procedure that Stansfield (2008) remarked, "*one could call this the decade of John Oller*" (p. 312, italics original). Oller was influenced by Carroll's research on intelligence, as well as his use of factor analysis and the introduction of the discrete-point/integrative approach. The Unitary Trait Hypothesis championed the view that language proficiency was best represented by a single underlying factor associated with general intelligence instead of separate underlying components commonly represented by reading, writing, listening, and speaking.

Along with its theoretical importance, the Unitary Trait Hypothesis also had practical implications for score reporting. Factor analysis (i.e., a procedure used to represent the relationships among observed, correlated variables and then reduce them into groups called factors) is a standard statistical procedure used, in part, to provide evidence that supports both total and subscale score reporting, especially of large, standardized tests like TOEFL. If only one factor emerges in the analysis, then subscale score reporting becomes problematic. Oller's evidence for the Unitary Trait Hypothesis relied heavily on a particular type of factor analysis that was misapplied. Eventually, he conceded that the strong version of his hypothesis, which excluded the presence of any subcomponents, was incorrect.

Language testing researchers now agree that the best representations of language proficiency include both a strong general factor (i.e., the first factor) along with subcomponents. According to Alderson (1991), "[...] language

proficiency is both unitary and divisible at the same time, it seems” (p. 18). Additional evidence was provided by Sawaki, Stricker, and Oranje (2009), who compared model fit across four different models of the TOEFL using confirmatory factor analysis and found that the two best-fitting ones (i.e., a correlated specific-factor model and a second-order model with a general factor with correlated subcomponents) were comparable in terms of model fit (see Sample Study 39.1). Some preferred that the general factor disappear altogether and welcomed stronger evidence for a broader array of subcomponents (Alderson, 1991). Yet, the general factor has persisted. But what is it? An examination of models and frameworks provides some leads.

### Sample Study 39.1

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26, 5–30.

#### Research Background

The change from the computer-based version of the TOEFL (CBT) to the internet-based version (iBT) was accompanied by several design changes: most importantly, the addition of a mandatory speaking section and an integrated item that requires analysis of an audio or reading passage as a source for writing. Previous studies examined evidence in support of subscale and total score reporting through confirmatory factor analysis. The present study expands this tradition by examining the structure of the TOEFL iBT to determine its construct validity.

#### Research Problems/Gaps

- How appropriate is the TOEFL iBT score-reporting policy?
- What is the relationship between the constructs assessed in the four sections of the test (reading/listening/speaking/writing)?
- Which skills are reflected by performance on the integrated speaking or writing tasks?

#### Research Method

- Construct validation study of all four sections of iBT.
- Participants from 31 countries required to take both TOEFL iBT and CBT.
- Confirmatory factor analysis of items to identify difficulty levels.

#### Key Results

The findings support reporting five separate scores for the test: one for each section and a total score. The integrated speaking and writing tasks are shown to enhance construct representation, but do not comprise separate speaking and writing subcomponents and have a weak relationship with the reading and listening sections, due in part to their lower degree of integration in the test's design.

#### Comments

This study provides evidence for a general large factor, along with separable subcomponents or skills, which confirms Oller's contention that general language proficiency is vitally important in representation of the underlying construct.



## Models and Frameworks of Communicative Competence and Language Proficiency

Few articles have had greater influence on the development of both language testing and applied linguistics than Canale and Swain's (1980) *Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing*. Chomsky (1965) held that the proper domain of inquiry for linguistics was the investigation of linguistic competence as represented by the notion of "an ideal speaker-listener, in a completely homogeneous speech community who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest and errors (random or characteristic) in applying his knowledge of the language in actual performance [...]" (pp. 3–4). Hymes' (1972) notion of *communicative competence* served as a rebuttal from the applied linguistics community. Canale and Swain's work helped to formalize this alternative view when they presented the first model of communicative competence, which included *grammatical*, *sociolinguistic*, and *strategic* competences. Subsequent adaptations by Canale (1983) added *discourse competence*. These adaptations were later extended by Bachman (1990) and Bachman and Palmer (1996) to include the superordinate *pragmatic competence* (i.e., psychophysiological mechanisms) and affective variables. These studies also addressed the relationship between knowledge and skills and the interactions among model components, while stressing Canale and Swain's original emphasis on how language ability manifests in real-world domains.

The focus on language use in real-world communicative contexts expanded the restricted view of language based on structure and allowed testers to move away from the perceived overreliance on reliability. However, attempts to represent communicative competence by focusing on language ability for use and the inclusion of actual, real-time language performance quickly led to what Spolsky (1986) called the "triumph of functional, communicative tests" (p. 147). Nearly ten years later, McNamara (1995) likened the situation to the opening of Pandora's box.

### Proficiency Scales

The American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines are perhaps the best known instantiation of an attempt to represent communicative competence and language proficiency testing in a manner that has direct reference to classroom teaching. The preface to the guidelines reads:



The ACTFL Proficiency Guidelines are a description of what individuals can do with language in terms of speaking, writing, listening, and reading in real-world situations in a spontaneous and non-rehearsed context. For each skill, these guidelines identify five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The major levels Advanced, Intermediate, and Novice are subdivided into High, Mid, and Low sublevels. The levels of the ACTFL Guidelines describe the continuum of proficiency from that of the highly articulate, well-educated language user to a level of little or no functional ability. (ACTFL, 2012, p. 3)

Both teachers and testers readily acknowledge that the guidelines appeal to common sense, lend themselves to the development of classroom activities and tasks, and provide a common understanding of assumed progression. The full appeal of language proficiency guidelines can be seen in the rapid, widespread adoption of the Common European Framework of Reference (CEFR), which divides proficiency into six major categories: basic user (A1, A2), independent user (B1, B2), and proficient user (C1, C2). Part of its popularity may be attributed to the current situation in Europe, where many languages are in contact and users can cross relatively permeable national boundaries.

Both the ACTFL Guidelines and the CEFR share a common focus on users' abilities to accomplish particular tasks nested within the traditional skills of reading, writing, listening, and speaking, while also being grounded in real-life language use. Both ACTFL and CEFR tasks often appear as "Can Do" statements, providing concrete representations and explicit guidance about what to teach at each level. However, the CEFR goes beyond the traditional four skills by providing guidelines for linguistic competence (i.e., general linguistic range, vocabulary range), sociolinguistic competence (i.e., situational appropriateness, pragmatic competence), and strategic competence (i.e., identifying cues/infering), as well as thematic development, interaction, and mediation.

As the developers of the ACTFL Guidelines and the CEFR are careful to point out, these guidelines are not to be associated with any particular theory or method and do not constitute, in and of themselves, reliable or valid tests. In fact, the ACTFL website clearly states that their Guidelines "neither describe how an individual learns a language nor prescribe how an individual should learn a language, and they should not be used for such purposes." Unfortunately, they are often used for exactly that. Despite their appeal to common sense, guidelines are not valid representations of language proficiency, much less assessments of proficiency. Spolsky (1986) observed of the

ACTFL: “[A] single set of guidelines, a single scale, could only be justified if there were evidence of an empirically provable necessary learning order, and we have clearly had difficulty in showing this to be so even for structural items” (p. 154).

Nevertheless, the influence of the CEFR is so strong that virtually every commercial test has been linked to it (see, e.g., Papageorgiou, Tannenbaum, Bridgeman, & Cho, 2015), which provides one type of criterion-related evidence of its validity. However, as Fulcher (2004) cautioned, linking tests to the CEFR is not a simple matter. Just as the ACTFL Guidelines were criticized for being presented and promoted without an appropriate theoretical or research foundation (Bachman, 1988; Bachman & Savignon, 1986; Fulcher, 1996), so too has the CEFR met with pushback from the language testing community for lack of empirical evidence in support of its use (Davidson & Fulcher, 2007; Fulcher, 2004).

Validation research on the CEFR is presently underway, and it is likely that it will survive because of, or in spite of, researchers’ scrutiny. It is only fair to note that the Council of Europe (2003) published *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment*, which, according to Fulcher (2004):

[r]ightly stresses the need for theoretical and empirical evidence to support the validation of the claim, and usefully states that it is the responsibility of test providers to investigate the reliability and validity of their own tests before attempting to establish a link. (pp. 261–262)

More recently, Hawkins and Filipović (2012) employed a corpus-based approach to show that particular linguistic properties can be linked to L2 proficiency at different levels, thus raising the prospects of using large data sets to provide evidence in support of the structural/grammatical features of proficiency levels.

## Overview of Research Methods

The potential contributions of corpus-based studies lie not so much in the statistical methods that are employed, which are often descriptive, but rather in the size of the data sets that are examined. Access to large data sets was formerly a luxury reserved for those who administered large-scale tests, as these sets were typically not available to academic researchers.

## Powerful Methods

Relatively large data sets (of at least 100 subjects) invite and allow for the use of powerful statistical methods, including item response theory (IRT) and structural equation modeling (SEM). These methods have been widely applied and are of great value when analyzing language testing data. Applications and varieties of IRT, particularly the one-parameter Rasch model, were first presented in the second issue of *Language Testing* (Henning, 1984). To date, over 200 articles have been published in that journal alone that reference or apply the most popular variant, many-facet Rasch, to analyses of language testing data. Many-facet Rasch analysis has proven especially useful when looking at performance scales (e.g., Eckes, 2008) and rater performance (e.g., Davis, 2016). McNamara (1996) and Wind and Peterson (2017) have provided thorough discussions of its various applications.

The other popular method, SEM, refers to a variety of statistical methods that examine the fit between a theoretical representation of constructs and an observed set of data. Confirmatory factor analyses through SEM has become the method of choice for looking at the underlying structure of language tests (Sawaki et al., 2009; see Phakiti, Chap. 21), and has been applied to other important questions in language testing, such as the influence of examinee background characteristics on language test performance (Ginther & Stevens, 1998), the presence of theoretical constructs (Phakiti, 2008), and the influence of syntax and vocabulary knowledge on reading comprehension (Shiotsu & Weir, 2007).

## Classical Test Theory

The methods identified above are attractive because of the complex analyses they allow and the power of what they can reveal, but as Carroll (1986) reminded us, “The simpler, more traditional item analysis procedures can suffice in many instances” (p. 126). Classical test theory represents the set of methods associated with traditional item analyses. Briefly, this theory posits that observed scores are best understood as being represented by underlying, unobservable true scores that can be estimated by their measurable component parts: observed score variance and error variance. The object of traditional item analyses is to increase true score variance by decreasing observed error variance to the greatest extent possible. This includes the examination of item difficulty, item discrimination, and test and rater reliability. These analyses are readily accessible, even to novice test developers, and remain quite effective. The test analyses reported in Davies (1984), Fulcher (1997), and Chapelle et al. (2010) employ classical test theory to examine reliability and validity (see Sample Study 39.2).

### Sample Study 39.2

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing*, 14, 113–138.

#### Research Background

Given the temporal and organizational constraints involved in administering and scoring placement tests, little evidence has been gathered to establish their reliability and validity. In this article, Fulcher examines a placement test used at the University of Surrey to determine the level of English language support that students may need. To reduce bias on essay prompts, a new version of the test was designed that focuses on content accessible to well-informed lay audiences rather than disciplinary specialists.

#### Research Problems/Gaps

- What is the appropriate external criteria for conducting a “concurrent” validity study?

#### Research Method

- A placement test was administered to the entire population of international and domestic undergraduate and postgraduate students.
- The effect of primary language was established using chi-square test.
- Inter-rater reliability was calculated using correlations and descriptive statistics.
- Item difficulty, discrimination, and reliability of subtest sections were examined using classical test theory and a Rasch model.
- Construct validity of subtest sections was assessed using correlation and principal components analysis.
- Concurrent validity was investigated by comparing select students’ scores on placement test with their scores on TOEFL.
- Content validity was explored via faculty feedback on test items.
- Face validity was addressed via a questionnaire administered to students who required English language support.

#### Key Results

This study highlights the need for pilot testing and thorough analyses of item performance to determine the reliability of test results as a foundation for placement decisions and to reveal any changes that may occur in the language support needs of students over time. Comparing future forms of the test with the current version could help to maintain score interpretability, heighten test security, and gauge score gains.

#### Comments

Fulcher outlines a detailed set of methods that could be employed to examine the reliability and validity for in-house placement measures.

## Qualitative Analysis

While language testing is more commonly associated with quantitative methods, a host of qualitative analyses are also frequently applied. Banerjee and Luoma (1997) provided an excellent overview of the most common

qualitative research methods associated with test validation. A recent study by Torkildsen and Erickson (2016) looked at what test takers had to say about tests and the testing process. Test takers' opinions and insights were also highlighted in Yan, Thirakunkovit, Kauper, and Ginther (2016) for quality control of administrative processes. The use and interpretation of language proficiency test scores in university admissions processes were examined in Ginther and Elder (2014), using mostly qualitative methods.

## Challenges and Controversies

### Consequential Validity

Drawing from Messick's (1989) argument that validity must address not only the evidence explicating the construct, relevance, and utility of test scores but also the implications and consequences associated with test score use, the 1999 revision of the *Standards for Educational Testing* tapped into the idea of "intended benefits," which had been brewing for some time (p. 16). This helped to foster discussions of how best to evaluate the consequences of test use, especially in social and political domains. Although extending validity arguments into these domains has been problematic for test developers, particularly in instances where improper use may be attributed to them, it is important to recognize that validity arguments also represent the practices and beliefs of other stakeholders (e.g., students, teachers, school administrators).

Despite its inclusion in the revised *Standards*, consequential validity remains a controversial topic. While acknowledging the very real consequences involved in testing, critics like Mehrens (1997) and Popham (1997) have argued that this adds unnecessary complications to scientific discussions of validity, and should be left in the realm of public policy. Advocates have countered that the consideration of consequences is a necessary corollary to any actual or anticipated test score use (Chapelle et al., 2008; Linn, 1998; Shepard, 1993). Although the debate over the relative merits of consequential validity continues, there is general consensus that the use and interpretation of test scores have serious consequences for schools, school districts, and state governments where, too often, those responsible for making decisions based upon test scores lack the training necessary to ensure sound judgments.

### Assessment Literacy

The term assessment literacy, coined by Stiggins (1991), is difficult to define, largely because it entails a range of knowledge and skills that must be acquired

and utilized to varying degrees by assorted stakeholders engaged in different aspects of the complex process of constructing, implementing, scoring, interpreting, and/or evaluating particular forms of assessment in an array of educational domains. To complicate matters further, the aforementioned social and political consequences tend to arise whenever tests and other assessments are used for decision-making purposes in these domains.

According to Inbar-Lourie (2008), “[b]ecoming assessment literate requires the attainment of ‘a toolbox’ of competencies, some practical, some theoretical, on why, when and how to go about constructing a variety of assessment procedures” (p. 389). These include—but are not limited to—the skills involved in test development and validation; the knowledge required to make informed and principled score-based decisions; the ability to read and make sense of assessment-related research data; and an awareness of context, including historical and political dimensions. In addition, those involved must be well versed in current theories of language learning and able to apply these theories to the creation and utilization of valid, reliable measures of knowledge and proficiency in a given language. This includes addressing challenges to so-called “native speaker” norms posed by the global spread of English (see below).

Tensions can arise when different stakeholders possess different degrees of assessment literacy and hold different—and sometimes conflicting—perspectives based on their positions within an educational domain. Rea-Dickins (2001) distinguished between three “identities” of assessment: bureaucratic, which reflects the external mandates of administrators and policymakers; pedagogic, which contributes to the development of appropriate lesson plans and assignments; and learning, which focuses on what students gain. There is often a divide, both real and imagined, between those who are directly involved with pedagogy and learning in the classroom (i.e., students, teachers, tutors) and those who are involved with the more bureaucratic aspects (i.e., test designers, administrators, policymakers). Undoubtedly, the growing emphasis placed on standardized testing as part of national and local educational reform efforts has led to negative forms of washback, such as decreased motivation and teaching to the test. However, Inbar-Lourie (2008) countered that this is a false dichotomy, one she likens to the split between quantitative and qualitative research. Instead, multiple perspectives are needed to inform assessment practices and research.

While the extent and intensity of involvement in developing assessment literacy varies among stakeholders, there is room for improvement across the board. In fact, a number of researchers (Fulcher, 2012; Weigle, 2007) have stressed the need for teachers to develop more acute awareness of assessment principles and practices in order to critically evaluate and, if necessary, resist externally mandated, high-stakes standardized tests, especially since they are

often the ones held accountable for student performance. Ideally, classroom instructors should factor assessment into every stage of the course, from planning the syllabus to giving final exams. This requires setting measurable objectives, deciding how to determine if those objectives have been met, and scoring students' performance (Weigle, 2007).

Likewise, administrators and policymakers should be able to provide satisfactory answers to queries from teachers, students, and parents about how and why a particular test is being used. As studies have shown (O'Loughlin, 2013; Ginther & Elder, 2014), those who use high-stakes tests for admissions and placement decisions at the university level often just check to see that a minimum score has been met without having a clear sense of how that score was determined or what it means with regard to a student's ability to perform in a classroom or real-world setting. This is particularly unsettling at a time when many universities are actively recruiting international students in an effort to offset the loss of other revenue streams. In addition, large-scale standardized language tests like TOEFL and IELTS are being used outside of college admissions and placement in noneducational domains such as immigration (Fulcher, 2012).

## World Englishes and English as a Lingua Franca

Work on World Englishes (WE) by Kachru (1985) and others (e.g., Berns, 2008; Nelson, 2011) has challenged the notion of an ideal native-speaker, long promoted in theoretical and applied linguistics, and helped to legitimize varieties other than standard British or American English. Meanwhile, English as a lingua franca (ELF) scholars like Seidlhofer (2001) and Jenkins (2006) have advocated for a more flexible contact language that could serve as a communicative resource for so-called "non-native" and "native" speakers alike. Both traditions have criticized language tests, especially large-scale ones like TOEFL, that continue to use native English speaker (NES) norms as the basis for items and assessment, despite the fact that non-native English speakers (NNES) are now the majority (Davidson, 2006).

To illustrate the complications that arise from language testing's attempts to adhere to a single norm, Spolsky (1993) shared an apocryphal story about the shock that some British testing experts felt when Lado, following the publication of his 1961 book, visited Cambridge and showed them copies of his tests that, to their disbelief, were written in standard American English. Spolsky then relayed his own experience with international students at Indiana University who had studied British English before arriving in the US and were



therefore worried about being disadvantaged on oral exams. From there, he presented an overview of the history of the exam, from its invention in ancient China to its function across the British Empire, before settling on the rise of TOEFL to show how different approaches to testing (e.g., performance-based, multiple-choice) and the normative varieties they favor (e.g., British English, American English) have helped to support colonialist and neocolonialist aims.

In the same issue of the journal *World Englishes*, Lowenberg (1993) examined the morpho-syntactic and stylistic differences between NES and NNES varieties to show that, while certain universal rules concerning parts of speech do exist, there are important variations, like the countability of nouns and the use of prepositions in phrasal verbs, that occur as often between British and American English as they do other varieties. Lowenberg challenged the assumption of adhering to these norms, especially when users speak localized varieties like Indian or Singaporean English, and urged test developers to be more sensitive to such differences in the future.

Nevertheless, WE and ELF continue to present special challenges for research and development in language testing. As Dimova (2017) pointed out, the variability of pronunciation across varieties and the lack of codification have made it difficult to measure proficiency without describing a standard. However, exposure to a variety is likely to have a positive effect on intelligibility (Berns, 2008; Major, Fitzmaurice, Bunta, & Balasubramanian, 2005), although it would be almost impossible to find a single rater who was familiar with every variety of English currently spoken in the world.

To allay concerns that speakers of nonstandard varieties will necessarily be disadvantaged when taking tests based on traditional NES norms, there is evidence to suggest that rater bias may not be an insurmountable problem. For example, Zhang and Elder's (2011) study of NES and NNES raters on the spoken section of the College English Test (CET-SET) in China found both groups were equally severe in their judgments of linguistic resources, although NES raters focused on content and fluency more than their NNES counterparts. To better assess the degree to which NES bias influences rating, Hsu (2016) constructed a "rater attitude instrument" that measured the expectations and tendencies of experienced NES and NNES raters toward Indian English on samples taken from the IELTS speaking section. Her results showed that, while over half of the raters preferred British or American English, the variety spoken by the test taker did not impact their scoring decisions in the majority of cases. Rather, communicative effectiveness and task completion weighed more heavily, which is consistent with recommendations from WE and ELF scholars (see Jenkins, 2006).



## Conclusion

As we have indicated in this chapter, each successive expansion of the construct of language proficiency has added to its richness in terms of description. This capacity to grow and adapt points toward a bright future for the field of language testing. While challenges and controversies remain, they also provide researchers with ample room to develop questions, test theories, apply frameworks and methods, and further their agendas. Most importantly, language testing requires that researchers, teachers, and other stakeholders acquire and maintain appropriate degrees of assessment literacy so that they may actively engage with theory and practice as they develop, administer, and score valid and reliable tests, interpret complex and occasionally contradictory data, and make principled decisions about when and how test scores are to be used.

## Resources for Further Reading

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

This book offers a broad, practical overview of foundational issues in language testing. Fulcher embeds the development of language tests within their historical, educational, and socioeconomic contexts. Within these contexts, he shows how the purpose of a test should serve as the starting point for its development and evaluation. He then presents the statistical underpinnings of standardized tests in a straightforward manner that allows readers to understand and appreciate classical test theory, while recognizing its limits. Fulcher uses examples and analogies to make his points, resulting in a text that is both informative and readable, even when covering the more arcane aspects of his subject.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.

Kane begins the article with a discussion of the shift from criterion-related to construct validity. His distinction between weak and strong paradigms alerts readers to the problems that arise from a lack of clear criteria for the evaluation of evidence. He then presents strategies that may be used to construct a fully-fledged validity argument. Kane contends that it is not the test

itself that is valid or invalid, but rather the inferences drawn on the basis of the test scores (e.g., a particular examinee is unprepared to enter or exit a course of study). His discussion of how successful arguments are developed, strengthened, or undermined has implications that extend beyond educational measurement and language testing.

Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests*. London: Longman.

Shohamy provides a concise overview of the often undeniably negative consequences of test use in this influential book, which introduced the idea of “critical language testing” as a subdomain of critical pedagogy. She draws on the power of personal narratives to illuminate the dark side of testing; indeed, who among us does not have a personal horror story that involves taking a test? This volume furthers the ongoing discussion of consequential validity initiated by Messick (1989). The book’s greatest contribution is found in its policy-oriented second half, where Shohamy outlines her vision of a democratic approach to language testing and addresses the rights and responsibilities of test takers.

## References

- ACTFL. (2012). *ACTFL proficiency guidelines* (Revised). Alexandria, VA: American Council on the Teaching of Foreign Languages.
- Alderson, C. (1991). Language testing in the 1990s: How far have we come? How much further have we to go? In A. Sarinee (Ed.), *Current developments in language testing: Anthology Series 25* (pp. 1–27). Singapore: Regional Language Centre.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 9–13). Hillsdale, NJ: Lawrence Erlbaum.
- Bachman, L. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition*, 10, 149–164.

- Bachman, L., & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal*, 70, 380–391.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Banerjee, J., & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, Volume 7: Language testing and assessment* (pp. 275–287). Dordrecht: Kluwer Academic.
- Berns, M. (2008). World Englishes, English as a lingua franca, and intelligibility. *World Englishes*, 27, 327–334.
- Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). New York: Longman.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47.
- Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (2nd ed., pp. 313–321). New York: McGraw Hill.
- Carroll, J. B. (1986). LT + 25, and beyond. *Language Testing*, 3, 123–129.
- Chapelle, C., Chung, Y., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, 27, 443–469.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. New York: Routledge.
- Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: M.I.T. Press.
- Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for languages: Learning, teaching and assessment*. Cambridge: Cambridge University Press.
- Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper and Row.
- Davidson, F. (2006). World Englishes and test construction. In B. B. Kachru, Y. Kachru, & C. Nelson (Eds.), *The handbook of world Englishes* (pp. 709–717). Hoboken, NJ: Wiley-Blackwell.
- Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching*, 40, 231–241.
- Davies, A. (1984). Validating three tests of language proficiency. *Language Testing*, 1, 50–69.
- Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33, 117–135.

- Davis, L., Laughlin, V., Gu, L., & Ockey, G. (2016, March). *Face-to-face speaking assessment in the digital age: Interactive speaking tasks on-line*. Paper presented at the Georgetown University Roundtable, Washington, DC.
- Dimova, S. (2017). Pronunciation assessment in the context of world Englishes. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 49–66). New York: Routledge.
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25, 155–185.
- Fulcher, G. (1996). Invalidating validity claims for the ACTFL oral rating scale. *System*, 24, 163–172.
- Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing*, 14, 113–138.
- Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, 1, 253–266.
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9, 113–132.
- Gardener, H. (1985). *The mind's new science*. New York: Basic Books.
- Ginther, A., & Elder, C. (2014). *A comparative investigation into understandings and uses of the TOEFL iBT test, the International English Language Testing Service (academic) test, and the Pearson Test of English for Graduate Admissions in the United States and Australia: A case study of two university contexts*. ETS research report No. TOEFLiBT-24. Retrieved from [https://www.ets.org/research/policy\\_research\\_reports/publications/report/2014/jtms](https://www.ets.org/research/policy_research_reports/publications/report/2014/jtms)
- Ginther, A., & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish language examination. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum.
- Hawkins, J., & Filipović, L. (2012). *Criterion features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.
- Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing*, 1, 123–133.
- Hsu, T. H.-L. (2016). Removing bias towards World Englishes: The development of a rater attitude instrument using Indian English as a stimulus. *Language Testing*, 33, 367–389.
- Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics. Selected readings* (pp. 269–293). Harmondsworth: Penguin.
- Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25, 385–402.
- Jenkins, J. (2006). Current perspectives on teaching world Englishes and English as a lingua Franca. *TESOL Quarterly*, 40, 157–181.
- Kachru, B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. Widdowson (Eds.), *English in the*

- world, teaching and learning the language and literatures* (pp. 11–30). Cambridge: Cambridge University Press.
- Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.
- Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17, 28–30.
- Lowenberg, P. H. (1993). Issues in validity in tests of English as a world language: Whose standards? *World Englishes*, 12, 95–106.
- Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2005). Testing the effects of regional, ethnic and international dialects of English on listening comprehension. *Language Learning*, 55, 37–69.
- McNamara, T. F. (1995). Modelling performance: Opening Pandora's box. *Applied Linguistics*, 16, 159–179.
- McNamara, T. F. (1996). *Measuring second language performance: A new era in language testing*. New York: Longman.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16, 16–18.
- Messick, S. (1975). The standard program: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.
- Miller, G. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, 7, 141–144.
- Morrow, K. (1981). Communicative language testing: Revolution or evolution? In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing*, 38 (pp. 9–26). London: The British Council.
- Nelson, C. (2011). *Intelligibility in world Englishes*. Hoboken, NJ: Blackwell.
- O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, 30, 363–380.
- Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels*. Research Memorandum-15-06. Princeton, NJ: ETS.
- Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, 25, 237–272.
- Popham, W. J. (1997). Consequential validity: Right concern – Wrong concept. *Educational Measurement: Issues and Practice*, 16, 9–13.
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18, 429–462.
- Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26, 5–30.

- Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, 11, 133–158.
- Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education*, 19 (pp. 405–450). Washington, DC: AERA.
- Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24, 99–128.
- Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 5–30). Frankfurt am Main: Peter Lang.
- Spolsky, B. (1986). A multiple choice for language testers. *Language Testing*, 3, 147–158.
- Spolsky, B. (1993). Testing across cultures: An historical perspective. *World Englishes*, 12, 87–93.
- Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.
- Stansfield, C. (2008). Where we have been and where we should go? *Language Testing*, 25, 311–326.
- Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, 72, 534–539.
- Torkildsen, L. G., & Erickson, G. (2016). “If they’d written more...” – On students’ perceptions of assessment and assessment practices. *Education Inquiry*, 7, 137–157.
- Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.
- Toulmin, S. (2001). *Return to reason*. Cambridge, MA: Harvard University Press.
- Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, 16, 194–209.
- Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 35, 161–192.
- Yan, X., Thirakunkovit, S., Kauper, N., & Ginther, A. (2016). What do test takers say: Test-taker feedback as input for quality control. In J. Read (Ed.), *Post-admission language assessments of university students* (pp. 157–183). Switzerland: Springer.
- Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, 28, 31–50.



# 40

## Linguistic Landscape

David Malinowski

### Introduction

The methodologies for the study of linguistic landscape (LL) are as diverse as its own disciplinary origins and allegiances. With roots in such fields as language policy and urban sociolinguistics, frequent borrowings from Hallidayan social semiotics and multimodal discourse studies, and a current home in applied linguistics, linguistic landscape is identified as much as a collection of *approaches* for data collection and analysis as it is a *field* unto itself. Studies in linguistic landscape have, to take a few examples, developed quantitative analytic methods to gauge the visibility of multiple languages across city neighborhoods as they “symbolically construct” the public space (Ben-Rafael, Shohamy, Hasan Amara, & Trumper-Hecht, 2006), conducted critical literacy and action research to study the development of language awareness among elementary school children in multilingual societies (Dagenais, Moore, Sabatier, Lamarre, & Armand, 2009), and employed phenomenological approaches to ethnographic observation in the “architecting” of tattooed “corporeal landscapes” (Peck & Stroud, 2015). As Hult (2014) notes succinctly, “LL analysts variously employ quantitative and qualitative techniques and are guided by a broad range of theoretical underpinnings” (p. 510).

As with any younger field, a degree of methodological uncertainty is to be expected as linguistic landscape endeavors to define its own object of study, a

---

D. Malinowski (✉)

Department of Linguistics and Language Development,

San José State University, San José, CA, USA

e-mail: [david.malinowski@sjsu.edu](mailto:david.malinowski@sjsu.edu)

© The Author(s) 2018

A. Phakiti et al. (eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology*,  
[https://doi.org/10.1057/978-1-137-59900-1\\_40](https://doi.org/10.1057/978-1-137-59900-1_40)



dilemma articulated by Spolsky in his “prolegomena to a sociolinguistic theory of public signage” (Spolsky, 2009, p. 25): “Whatever we call it, is linguistic landscape a phenomenon calling for a theory, or simply a collection of somewhat disparate methodologies for studying the nature of public written signs?” Some researchers continue to rely upon the objectivist definition of linguistic landscape offered by Landry and Bourhis (1997, p. 25), “the language of public road signs, advertising billboards, street names, place names, commercial shop signs and public signs on government buildings” or, more concisely, “linguistic objects that mark the public space” (Ben-Rafael et al., 2006, p. 7). However, many others have taken issue with the delimitations imposed by either of the *Ls* in “LL”: Gorter, for instance, noted problems with the term “landscape” and the frequency of studies set in densely populated urban settings, and suggested “linguistic cityscape” instead (Gorter, 2006, p. 2), while Jaworski and Thurlow (2010a) presented a rigorous argument for consideration of the “semiotic” beyond just the “linguistic.” Although the focus of the preponderance of earlier studies was (and continues to be) upon *writing*, inscriptions, and other visible designs, a growing concern with other forms of languaging and symbolic practices has made less focused glosses for LL more attractive, such as Blackwood, Lanza, and Woldemariam’s “the study of language in the public space” (2016, p. xvi). Finally, the apparently common denominator in all linguistic landscape studies, that is, the *public* that is held to be both a necessary condition for and a product of people’s diverse semiotic activities (e.g., Shohamy, 2015), has also been probed and complexified, as in Hanauer’s (2009) analysis of wall space in a microbiology laboratory, Zabrodskaja’s (2014) observation about the influence of public policy on “private space and mind” in Estonia (p. 115), and the ambiguities of graffiti as a public text (e.g., Pennycook, 2009). Linguistic landscape research, then, employs a diverse and expanding palette of methodological approaches relating observable textual traces to situated semiotic practices in order to better understand interrelationships of language, power, and society. This reality is exemplified in the two sample studies at the end of this chapter.

## Core Issues: Finding a Balance Between Quantitative and Qualitative Approaches

At a broad level, there is recognition that linguistic landscape research began with a predominantly quantitatively oriented methodological focus, with strong ties to variationist sociolinguistics and other traditions in the positivist social



and linguistic sciences. Backhaus' (2007) monograph *Linguistic Landscapes*, comparing evidence of multilingualism in a corpus of approximately 11,800 signs observed across 28 sites in Tokyo, was a case in point, and is widely regarded as foundational in the field. Other prominent collections of papers appearing at this time, including a 2006 special issue of the *International Journal of Multilingualism* and the chapters in Shohamy and Gorter (2008), also illustrated a strong concern for analysis of the distribution and prominence of linguistic features/codes across geographic space, often through comparisons of two focal sites. In recent years, statistical surveys and frequency comparisons have assumed even greater nuance and complexity, as in Peukert's analysis of correlations between visible diversity in the LL and "usage structures," or, as Peukert explains, "the actual utilization of a concrete spatial unit involving social actions and practices independent of the functions of language" (Peukert, 2015, p. 30).

The late 2000s also saw a pushback against descriptivist and distributional approaches, with calls for greater responsiveness to complexity, change, and subjective meanings in linguistic landscape research or, as Coupland and Garrett (2010, p. 12) argued, "more [sensitivity] to historical processes and contexts, as well as to textual nuances." Their qualitative methodological approach, involving the identification and analysis of culturally meaningful *frames*, was one response adopted by others as well (e.g., Kallen, 2010; Moriarty, 2014). Meanwhile, others have employed the ethnographically based discourse analysis approaches of *geosemiotics* and *nexus analysis* to analyze social action in place (e.g., Hult, 2014; Pietikäinen, Lane, Salo, & Laihiala-Kankainen, 2011). And, despite ongoing doubts as to what precisely is meant by "ethnography" in linguistic landscape (see O'Connor & Zentz, 2016, p. 29), a significant thread of research identifies itself with ethnographic methods and traditions (Blommaert, 2013; Collins & Slembrouck, 2007; Lou, 2016; Malinowski, 2009).

More recently still, there have been explicit arguments made in favor of a balanced approach between methodological approaches, partly due to a perception that the "qualitative shift" has unfairly sidelined quantitatively oriented studies. Conducting new surveys at the sites of his own previous research on the visibility of France's regional languages, Blackwood (2015) reaffirms the value of multi-sited quantitative investigations in allowing for spatial (cross-site) and temporal (diachronic) inquiry, and concludes, "a symbiotic approach, where the quantitative and qualitative approaches feed into one another, is the ideal *modus operandi*" (p. 40). In point of fact, much recent research in linguistic landscape is making more self-conscious use of multiple methods to bolster findings (Gorter, 2013, p. 199). However, across

the field of linguistic landscape there has been and continues to be vigorous discussion about such “fundamental” issues as the definition of the “research site,” appropriate methods for site selection and delineation and, as will be discussed below, the very meaning of the term “sign.”

## **A Methodological Debate Writ Small: “What Is a Sign?”**

Although this question may seem simple in light of the plethora of street signs, billboards, and other public texts that find themselves at the center of linguistic landscape research, defining the unit of analysis (or “phenomenon of interest”) has been far from straightforward. The definition of a sign in a linguistic landscape study, of course, has implications for if and how signs as material objects can be counted and classified—or whether they should.

Backhaus’ (2007) definition of a sign as employed in his quantitative surveys of Tokyo neighborhoods is a frequent reference point: he defines it as “any piece of written text within a spatially definable frame” (p. 66). The fact that this definition treats as equal “anything from the small handwritten sticker attached to a lamp-post to huge commercial billboards” was one reason for Cenoz and Gorter (2006), for instance, to offer a definition that sees not spatially definable frames but “an establishment” (e.g., a shop, a restaurant), with all of its publicly displayed texts counted as a single sign. Others seeking to count and classify the entirety of a given signed environment have taken issue with other definitional issues, such as whether to admit mobile texts into analysis: whereas Backhaus, Gorter, and many others typically exclude “all non-stationary objects” (Backhaus, 2007, p. 67), Sebba (2010, p. 61), for instance, argues for the inclusion of “everyday mobile texts” like banknotes, stamps, and bus tickets since these, too, comprise the “textually mediated world” and help to index linguistic vitality. Kallen’s (2010) notion of “the detritus zone” (p. 53), of trash as part of the linguistic landscape, is a further extension of this logic.

If recognizing and counting signs as discrete, countable objects in the material world raise these kinds of challenges, then taking stock of the intertextuality of signs, and signs as embedded in, or constituted by, perception, communication, and social action, raises others. Scarvaglieri, Redder, Pappenhagen, and Brehmer (2013), for instance, propose a “functional-pragmatic” approach that focuses upon “the interactive process between author and reader” (p. 55) whereby “each sign documents a unit of textual linguistic action.” Such an action-focused approach does not refute the salience of written texts within spatially definable frames but foregrounds

social semiotic processes themselves as focal points of data collection and analysis. Historical studies making use of archival photographs, film, literary references, and other mediums of representation (e.g., Pavlenko & Mullen, 2015); material-discursive approaches that seek to retrace chains of semiotization across stratified social spaces (Stroud & Mpendukana, 2009); ethnographic research that foregrounds the subjective perceptions, readings, and verbal accounts of participants (e.g., Garvin, 2010); and research in such domains as virtual/online linguistic landscapes (Ivković & Lotherington, 2009) and corporeal landscapes (Kitis & Milani, 2015) dramatically unsettle, expand, and refigure the possible meanings of “signs” in linguistic landscape.

## **An Ongoing Challenge: Defining the LL Subject-in-Place**

As indicated above, the definition of “linguistic landscape” as a field is far from settled—a fact that may have as much to do with the field’s reticence to treat landscapes as more than the static, public backdrop of the city (Pennycook, 2009, p. 304) as it does the relative newness of LL studies. While the question of how research in linguistic landscape speaks to more geographically centered fields (as in Nash, 2016, writing in *Landscape Studies*) may not be relevant to the immediate concerns of applied linguists, it nonetheless raises a more general issue of (self-)definition: What, and who, are the subjects of a linguistic landscape research agenda that takes both of its identifiers (“language” and “landscape”) seriously?

## **Growth of LL Corpora, Big(ger) Data Approaches, and GIS Applications**

Methodological innovations for researching linguistic landscape “at scale”—that is, for amplifying methods of collecting, analyzing, and representing sets of data by factors of magnitude larger than in previous years—suggest paths forward. Recent calls for a unified canon of variables in quantitative linguistic landscape research (e.g., Amos & Soukup, 2016) bespeak the potential for the systematization of data gathering and categorization techniques across geographic and research contexts, and thus address problems of sampling and generalizability of findings (cf. Gorter, 2013, p. 201). The use of such techniques as hypothesis-driven stratified judgment sampling (relying extensively upon published geo-referenced demographic data; see Soukup, 2016) and

participant response heatmaps (Amos, 2016), for instance, represents measured responses to the seeming impossibility of undertaking “an authoritative, comprehensive study of all the public space in a given city” (Blackwood, 2015, p. 41). Meanwhile, experiments with surveying the linguistic landscape virtually, using publicly available and all-encompassing “street view” services such as Google Street View, would seem to suggest that such an “authoritative, comprehensive” study may, after all, be possible (see Park & Yang, 2016). Work utilizing Geographic Information Systems (GIS), digital mapping, data visualization, and other sociodemographic research on linguistic variation offers new approaches for understanding multilingualism, language contact, and historical change at levels of granularity and scope before considered impossible.

### **Pushing Modes and Manners of Knowing the Linguistic Landscape, in the First Person**

Challenging the notion of the linguistic landscape as “objective fact,” a growing number of studies over the last decade have recognized the subjective and sometimes political nature of residents’, merchants’, customers’, tourists’, and even researchers’ attribution of meanings into and from public texts. In some cases, researchers have elicited responses from participants in interviews and surveys by showing them images (e.g., Collins & Slembrouck, 2007); in others, researchers have accompanied or video-recorded participants as they react to the linguistic landscape in real time, through walking (or driving) tours (e.g., Garvin, 2010). In rarer cases, the researcher self-reflexively addresses the question of her or his own subjectivities that may impact significantly upon such questions as framing of research questions, site selection, and collection and analysis procedures (e.g., Chmielewska, 2010; Milani, 2014).

Of course, to open research to greater reliance upon human senses and perceptions is also to admit more diverse perspectives on what is meant by, and what surpasses, “language” in the public sphere (cf. Jaworski & Thurlow, 2010b). In particular, the so-called linguistic *soundscape* or *audioscape*—the aural “scene” that is perceptible in a place allowing for the documentation of “how social spaces are created and developed over time through linguistic action” (Scarvaglieri et al., 2013, p. 63)—has received attention as it either complements or contradicts the languages on visible public texts (Mitchell, 2010). More recently, Pennycook and Otsuji (2015) have posited the inter-

relation of the sensory landscape with the linguistic and the semiotic; they argue that smellscape, in particular, “make central the capacity of the senses to connect across time and space” (p. 200) and illustrate the propensity of lived spaces to interpellate people in ways both expected and beyond control. In this calculation, the linguistic landscape extends far beyond typical definitions of “languages on public and commercial signs in a given territory or region” (Landry & Bourhis, 1997, p. 23) and implicates human subjects’ own ways of knowing and being in that landscape.

In light of these considerations, we may surmise that the task of reconciling the “linguistic” with the “landscape” and thus of defining the *subject of the field* is not to privilege either the objectivity of language in the material world or the subjectivities of its readers, writers, and walkers at the expense of the other. It is (also) to relate the two together in ways that reveal how meaning-making in place may be an “emergent property deriving from the interactions between people, artefacts, and space” (Pennycook, 2016, p. 10). Seen from this perspective, one methodological challenge for the field of linguistic landscape will be to introduce more elements of non-representationality (cf. Thrift, 2007) and reflexivity into its heavily representational habits of seeing, framing, and studying signs.

## Limitations and Future Directions

In light of the marked interdisciplinarity of the field of linguistic landscape, and in consideration of some of the unique challenges in designing research methods to reconcile knowledge of language and society with people and place, in this final section we suggest an agenda of three possible research directions that, at the same time, recognize some of the limitations of knowledge generation in the field. These areas may be seen as mutually related, building upon the thematic concerns outlined in the previous sections. They do not in and of themselves seek to resolve the perhaps larger questions of finding an appropriate balance (and rich points of contact) between quantitative and qualitative methods within and across linguistic landscape research projects, or of more rigorously and transparently representing methodological choices and procedures in linguistic landscape research projects—both in and of themselves significant concerns (see, e.g., Blackwood, 2015; Gorter, 2013).

## Operationalizing Questions of Materiality and Embodiment

Drawing upon Kress and van Leeuwen's (1996) *Grammar of Visual Design*, Scollon and Scollon (2003) asserted that meanings made by the materiality of signs and the semiotic environment at large are inseparable from, for instance, linguistic meaning. Yet, as Chmielewska (2010) argues, the linguistic/material world of signs in place is not simply a more complex object to be read from an abstracted vantage point. Rather, the material world of necessity includes the embodied subjectivities of those within it. In this sense, the politics, techniques, and consequences of emplacement—of both language and people in place—may further become focal objects of study. Approaches such as Stroud and Mpendukana's (2009, 2010) material ethnography of multilingualism and Pennycook and Otsuji's (2015) pursuit of sensory landscapes and spatial repertoires outline productive paths forward; making linguistic landscape accountable to systemic practices of gender-ing, race-ing, class-ing, and ageing (to name a few) in and of landscapes—and of bodies as landscapes—will also be a priority (Milani, 2014; Peck & Stroud, 2015).

## Researching Mobility and Mobilizing Research

The transit of discourses uncovered in Stroud and Mpendukana's (2009) material ethnographic approach, Coulmas' (2009) account of linguistic *landscaping* as process, and Chmielewska's (2010) perspective on the evolving nature of a sign reader's interpretation of the LL<sup>1</sup> are but a few of the many acknowledgments of the importance of *mobility* in current linguistic landscape studies (see Moriarty, 2014). Where Sebba's (2010) argument in favor of analyzing "the totality of public texts, fixed or circulating, within a specified location" (p. 61) pushed up against objectivist understandings of the linguistic landscape as a body of texts fixed in place, Hult's (2014) "drive-thru linguistic landscaping" on the highway system of San Antonio, Laitinen's (2014) 630-kilometer documentation of the Finnish linguistic landscape by bicycle, and Garvin's (2010) walking tour interviews point to various possibilities for mobilizing research, at least partially in the persons of the researchers themselves. In particular, the growing interest in transience, invisibility, and exclusion in the linguistic landscape and the transformation of spaces through demonstrations and protest (Martín Rojo, 2016; Rubdy & Ben Said, 2016) attests to the urgency of developing methods of mobile documentation and analysis.

## Virtual Linguistic Landscapes

The task of accounting for materialities, embodiments, and mobilities in linguistic landscape research assumes further complexities when considering the prevalence of virtual and digitally augmented technologies and environments. Chun's (2014) study of semiotic mobility in the Occupy Movement and Lee's (2015) employment of literacy studies, geosemiotics, and "contextualized" linguistic landscape interpretive frameworks to investigate the enregisterment of "textspeak" or internet-specific language in physical public places are two instantiations of this effort. Still, in light of the foundational concern in linguistic landscape studies with issues of language (choice, variation, function, symbolism, etc.) in *public* texts, key questions remain around the *publicness* of texts and textual-material practices in virtual, online, and hybrid contexts (see also Berezkina, 2015).

Jones (2010) offers one path forward by using mediated discourse analysis (Norris & Jones, 2005) to understand how people's (often) simultaneous engagement in online and offline activity is subject to "a complex nexus of overlapping and competing attention structures in the discourses in place, the interaction order, and the historical bodies of participants" (p. 164). *Attention* and mediated *action*, then, become the analytic foci, still with a keen eye to the discourses in place that have preoccupied linguistic landscape researchers for close to two decades.

At a more general level, too, we note that the potential utility of computer-mediated communication and digitally mediated discourse studies for explicating some of the methodological and theoretical dilemmas of linguistic landscape studies has not gone unnoticed: Androutsopoulos' recent methodological comparison of computer-mediated communication (CMC) and linguistic landscape (Androutsopoulos, 2014) finds several parallel developments between the fields, including a common concern for the analysis of (primarily) written languaging practices in complex and dynamic semiotic environments, and a continuum of approaches "between a 'purely textual' and a more 'ethnographic' approach" (p. 75).

This dilemma—that is, between reading LL-as-text and LL-as-place—harks back to the opening remarks about the double nature of landscape that, in the view taken up in this chapter, may confound attempts to find a unique or singular "linguistic landscape methodology." However, seen from another angle, the very act of debating the double-sidedness of landscape and resulting ambivalences of the field through rigorous research on questions of materiality, movement, and virtualization (to name a few) may itself *be* the point, giving shape and direction to a field that, like any other, cannot see itself but from the outside.



### Sample Study 40.1

Scarvaglieri, C., Redder, A., Pappenhagen, R., & Brehmer, B. (2013). Capturing diversity: Linguistic land- and soundscaping. In J. Duarte & I. Gogolin (Eds.), *Linguistic superdiversity in urban areas: Research approaches* (pp. 45–74). Amsterdam: John Benjamins Publishing.

#### Research Background

Recognizing the difficulty in traditional LL approaches to understand the lived nature of linguistic diversity, Scarvaglieri et al. conduct a multi-tiered linguistic landscape and soundscape analysis of two demographically contrasting neighborhoods in Hamburg, Germany. The authors begin with the assertion that traditional techniques employed in linguistic landscape studies to count and categorize of signs (e.g., top-down/bottom-up) overlook diverse author-reader interactive processes; they employ a functional-pragmatic approach to reveal “language as a complex form-function nexus anchored in reality as social practice” (Redder, 2008, p. 133, cited on p. 50).

#### Research Problems

The goals of the chapter are themselves largely methodological. Through their analysis, the authors set out to investigate:

- how the performance of linguistic diversity on the micro-level of linguistic action can be described and analyzed.
- what methods can be used to connect these micro-level analyses to the macro-level of social diversity.

#### Research method

- *Type of research:* Functional-pragmatic discursive analysis
- *Setting and participants:* Two adjacent neighborhoods in the St. Georg district of Hamburg, Germany—one middle class and gentrified and the other lower class
- *Instruments/techniques:* Language documentation in monolingual and multilingual signs with team-based in-person observation protocols and photography; two weeks of “soundwalk” observations, hand documentation, and recordings of verbal language use
- *Data analysis:* Statistical distribution analysis of language use in 1034-sign corpus, followed by in-depth functional-pragmatic analysis of samples for writer-reader interaction system, speech actions in public and semi-public spaces

#### Key Results

The study portrays a heterogeneous picture of visual and verbal language practices comprising the public sphere in both neighborhoods, even as German was found to be the most commonly used language. The analysis confirms a functional and social division within St. Georg, with migration-induced multilingualism characterizing one neighborhood and high “symbolic value” language use indicative of a “gentrified multilingualism” (p. 69) in the other. Additionally, despite challenges of audio recording in public space, the authors remark, “the recording and subsequent analysis of authentic spoken discourse in public places remains a methodological desideratum within the approach of linguistic soundscaping” (p. 70).



### Comments

In addition to demonstrating the rich possibilities of a mixed-method study carried out by a multi-member research team, this study contributes to the trend in the field toward problematizing subjective and hard-to-see elements of writer-reader relations, disrupting tendencies to assign fixed meaning to the mere presence of a language or languages on a sign. Furthermore, the integration of soundscape and situated audio analysis bears significant implications for understanding LL with respect to issues such as materiality, mobility, and virtualization.

### Sample Study 40.2

Peck, A., & Stroud, C. (2015). Skinscapes. *Linguistic Landscape*, 1, 133–151.

#### Research Background

In the context of post-apartheid South Africa where, as the authors explain, “the question of which bodies fit into what spaces remains highly contentious” (p. 134), this study presents a qualitative investigation into the affective and performative dimensions through which bodies and place write themselves into each other. The research builds upon persistent questions in linguistic landscape literature, including ethnic coexistence in place, erasure and memorialization, and history and mobility in place, while foregrounding the dual role of the body both in and as landscape.

#### Research Problems

As was the case with the Scarvaglieri et al. chapter discussed in Sample Study 40.1, one of Peck and Stroud’s goals is to advance methodologies for LL studies by synthesizing and applying techniques in a new domain. While the reading and writing practices, mobility, and even gaze of human subjects in the linguistic landscape have been recognized in the literature, for the large part, their bodies have not. Peck and Stroud ask how “bodies that matter” in historically and racially divided places are created through the negotiation and narrativizing of semiotics of and on the skin.

#### Research Method

- *Type of research*: A hybrid of more traditional ethnographic methods with phenomenologically motivated *sensory* observation and recording.
- *Setting and participants*: Primary data gathering took place over approximately six months with the owner, staff, and clients of a small tattoo shop in a multicultural, multiracial Cape Town neighborhood.
- *Instruments/techniques*: Ethnographic methods included participant observation, interviews, and field notes, while phenomenologically oriented observations were captured with tools including head-mounted cameras for video recording.
- *Data analysis*: Discourse analysis utilizing transcripts from five case studies coded for, among other features, participants’ touch, affect, posturing, and gaze during and around processes of bodily inscription (tattooing).

### Key Results

The study illustrates how practices of tattooing, as one form of bodily inscription, cast together participants' lived experience of place and time with social discourses of race, gender, class, occupation, and other grounds of identity formation and contestation. The authors argue that practices of "touch" are invaluable in elucidating possible future pathways of visibility and movement for bodies in question and suggest the value of a corporeal sociolinguistics approach for the field of linguistic landscape.

### Comments

Through an intensely proximal focus achieved through sustained close observation, interaction, and video recording, this research project opens up some of the more expansive and transdisciplinary concerns visible in linguistic landscape research. It directly relates the inking of bodies with those bodies' own trajectories through historically raced, gendered, and classed spaces, as *inscriptions* in place. As such, it offers one approach to bridging scales of analysis in LL research, while calling attention to the human mechanisms through which social structures of difference are contested and re-enacted.

## Resources for Further Reading

Following themes highlighted in this chapter, these resources are intended to suggest expansive thinking about methodological possibilities for linguistic landscape research, in the areas of *language* (King & Carson, 2016), *place* (Presner et al., 2014), and *the body* (Kozel, 2007).

King, L., & Carson, L. (Eds.). (2016). *The multilingual city: Vitality, conflict and change*. Bristol and Buffalo: Multilingual Matters.

Long a central concern in linguistic landscape studies, the "vitality" of languages and multilingualism in urban contexts forms the centerpiece of this edited volume. Through a synthesis of micro and macro approaches, the book's collection of chapters advances the concerns of plurilingual citizenship, education, policy, and work taken up in the LUCIDE (Languages in Urban Communities: Integration and Diversity for Europe) network's multi-year, 18-city research project.

Kozel, S. (2007). *Closer: Performance, technologies, phenomenology*. Cambridge, MA: MIT Press.

In this book, Kozel pursues "thoughtful and philosophically well-grounded first-person, or subjective, approaches to research" (p. 15) by conducting a

series of phenomenological analyses of dance and performances. These movement events are tracked and visualized with wearable and motion-sensing technologies, demonstrating how flesh and digital technologies are interwoven, ever closer than they have been before.

Presner, T., Shepard, D., & Kawano, Y. (2014). *HyperCities: Thick mapping in the digital humanities*. Cambridge, MA: Harvard University Press.

A collection of chapters exploring the theory and methods behind *HyperCities*, an online multilayered platform “for exploring, learning about, and interacting with the layered histories of city and global places,” this multidisciplinary volume thoughtfully elaborates a critical geospatial and narrativizing methodology known as *deep mapping*.

## Note

1. As she writes of the prototypical “receiver” of environmental information: “She actively searches for, organizes, classifies and selects information *while* moving among, around objects, in relation to their surfaces, to changing points of *surface attachment* or *implacement*” (Chmielewska, 2010, p. 277; italics in original).

## References

- Amos, W. (2016). Chinatown by Numbers: Defining an ethnic space by empirical linguistic landscape. *Linguistic Landscape*, 2, 127–156.
- Amos, W., & Soukup, B. (2016). *Towards a canon of variables in quantitative LL research*. Presented at the 37th International LAUD Symposium, Landau/Pfalz, Germany.
- Androutsopoulos, J. (2014). Computer-mediated communication and linguistic landscapes. In J. Holmes & K. Hazen (Eds.), *Research methods in sociolinguistics: A practical guide* (pp. 74–90). Hoboken: Wiley-Blackwell.
- Backhaus, P. (2007). *Linguistic landscapes: A comparative study of urban multilingualism in Tokyo*. Clevedon: Multilingual Matters.
- Ben-Rafael, E., Shohamy, E., Hasan Amara, M., & Trumper-Hecht, N. (2006). Linguistic landscape as symbolic construction of the public space: The case of Israel. *International Journal of Multilingualism*, 3, 7–30.
- Berezkina, M. (2015). Russian in Estonia’s public sector: “Playing on the borderline” between official policy and real-life needs. *International Journal of Bilingual Education and Bilingualism*, 20, 417–427.

- Blackwood, R. (2015). LL explorations and methodological challenges: Analysing France's regional languages. *Linguistic Landscape*, 1, 38–53.
- Blackwood, R., Lanza, E., & Woldemariam, H. (Eds.). (2016). *Negotiating and contesting identities in linguistic landscapes*. New York: Bloomsbury.
- Blommaert, J. (2013). *Ethnography, superdiversity and linguistic landscapes: Chronicles of complexity*. Clevedon: Multilingual Matters.
- Cenoz, J., & Gorter, D. (2006). Linguistic landscape and minority languages. In D. Gorter (Ed.), *Linguistic landscape: A new approach to multilingualism* (pp. 67–80). Clevedon: Multilingual Matters.
- Chmielewska, E. (2010). Semiosis takes place or radical uses of quaint theories. In A. Jaworski & C. Thurlow (Eds.), *Semiotic landscapes: Language, image, space* (pp. 274–291). London and New York: Continuum.
- Chun, C. W. (2014). Mobilities of a linguistic landscape at Los Angeles City Hall Park. *Journal of Language and Politics*, 13, 653–674.
- Collins, J., & Slembrouck, S. (2007). Reading shop windows in globalized neighborhoods: Multilingual literacy practices and indexicality. *Journal of Literacy Research*, 39, 335–356.
- Coulmas, F. (2009). Linguistic landscaping and the seed of the public sphere. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 13–24). New York: Routledge.
- Coupland, N., & Garrett, P. (2010). Linguistic landscapes, discursive frames and metacultural performance: The case of Welsh Patagonia. *International Journal of the Sociology of Language*, 205, 7–36.
- Dagenais, D., Moore, D., Sabatier, C., Lamarre, P., & Armand, F. (2009). Linguistic landscape and language awareness. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 253–269). New York: Routledge.
- Garvin, R. T. (2010). Responses to the linguistic landscape in Memphis, Tennessee: An urban space in transition. In E. Shohamy, E. Ben-Rafael, & M. Barni (Eds.), *Linguistic landscape in the city* (pp. 252–271). Bristol, UK: Multilingual Matters.
- Gorter, D. (2006). Introduction: The study of linguistic landscape as a new approach to multilingualism. In D. Gorter (Ed.), *Linguistic landscape: A new approach to multilingualism* (pp. 1–6). Clevedon: Multilingual Matters.
- Gorter, D. (2013). Linguistic landscapes in a multilingual world. *Annual Review of Applied Linguistics*, 33, 190–212.
- Hanauer, D. I. (2009). Science and the linguistic landscape: A genre analysis of representational wall space in a microbiology laboratory. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 287–301). New York: Routledge.
- Hult, F. M. (2014). Drive-thru linguistic landscaping: Constructing a linguistically dominant place in a bilingual space. *International Journal of Bilingualism*, 18, 507–523.
- Ivković, D., & Lotherington, H. (2009). Multilingualism in cyberspace: Conceptualising the virtual linguistic landscape. *International Journal of Multilingualism*, 6, 17–36.

- Jaworski, A., & Thurlow, C. (2010a). Introducing semiotic landscapes. In A. Jaworski & C. Thurlow (Eds.), *Semiotic landscapes: Language, image, space* (pp. 1–40). London and New York: Continuum.
- Jaworski, A., & Thurlow, C. (Eds.). (2010b). *Semiotic landscapes: Language, image, space*. London and New York: Continuum.
- Jones, R. H. (2010). Cyberspace and physical space: Attention structures in computer mediated communication. In A. Jaworski & C. Thurlow (Eds.), *Semiotic landscapes: Language, image, space* (pp. 151–167). London: Continuum.
- Kallen, J. L. (2010). Changing landscapes: Language, space and policy in the Dublin linguistic landscape. In A. Jaworski & C. Thurlow (Eds.), *Semiotic landscapes: Language, image, space* (pp. 41–58). London and New York: Continuum.
- Kitis, D. E., & Milani, T. M. (2015). The performativity of the body. *Linguistic Landscape, 1*, 268–290.
- Kress, G., & van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. London: Routledge.
- Laitinen, M. (2014). 630 kilometres by bicycle: Observations of English in urban and rural Finland. *International Journal of the Sociology of Language, 2014*, 55–77.
- Landry, R., & Bourhis, R. Y. (1997). Linguistic landscape and ethnolinguistic vitality. *Journal of Language and Social Psychology, 16*, 23–49.
- Lee, C. (2015). Digital discourse@public space: Flows of language online and offline. In R. H. Jones, A. Chik, & C. A. Hafner (Eds.), *Discourse and digital practices: Doing discourse analysis in the digital age* (pp. 175–192). New York: Routledge.
- Lou, J. J. (2016). *The linguistic landscape of Chinatown: A Sociolinguistic ethnography*. Bristol: Multilingual Matters.
- Malinowski, D. (2009). Authorship in the linguistic landscape: A multimodal-performative view. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 107–125). New York: Routledge.
- Martín Rojo, L. (Ed.). (2016). *Occupy: The spatial dynamics of discourse in global protest movements*. Amsterdam: John Benjamins.
- Milani, T. M. (2014). Sexed signs: Queering the scenery. *International Journal of the Sociology of Language, 2014*, 201–225.
- Mitchell, T. D. (2010). “A Latino community takes hold”: Reproducing semiotic landscapes in media discourse. In A. Jaworski & C. Thurlow (Eds.), *Semiotic landscapes: Language, image, space* (pp. 168–186). London & New York: Continuum.
- Moriarty, M. (2014). Languages in motion: Multilingualism and mobility in the linguistic landscape. *International Journal of Bilingualism, 18*, 457–463.
- Nash, J. (2016). Is linguistic landscape necessary? *Landscape Research, 41*, 380–384.
- Norris, S., & Jones, R. H. (Eds.). (2005). *Discourse in action: Introducing mediated discourse analysis*. London: Routledge.
- O’Connor, B. H., & Zentz, L. R. (2016). Theorizing mobility in semiotic landscapes: Evidence from South Texas and Central Java. *Linguistic Landscape, 2*, 26–51.

- Park, E., & Yang, J. (2016). A sociolinguistic analysis of a commercial district in Seoul: A linguistic landscape approach. *The Sociolinguistic Journal of Korea*, 23, 37–63.
- Pavlenko, A., & Mullen, A. (2015). Why diachronicity matters in the study of linguistic landscapes. *Linguistic Landscape*, 1, 114–132.
- Peck, A., & Stroud, C. (2015). Skinscapes. *Linguistic Landscape*, 1, 133–151.
- Pennycook, A. (2009). Linguistic landscapes and the transgressive semiotics of graffiti. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 302–312). New York: Routledge.
- Pennycook, A. (2016). Posthumanist applied linguistics. *Applied Linguistics*, Advance Access, 1–18. <https://doi.org/10.1093/applin/amw016>
- Pennycook, A., & Otsuji, E. (2015). Making scents of the landscape. *Linguistic Landscape*, 1, 191–212.
- Peukert, H. (2015). Urban linguistic landscaping: Scanning metropolitan spaces. In M. Laitinen & A. Zabrodska (Eds.), *Dimensions of sociolinguistic landscapes in Europe: Materials and methodological solutions* (Vol. 7, pp. 29–51). Frankfurt: Peter Lang.
- Pietikäinen, S., Lane, P., Salo, H., & Laihiala-Kankainen, S. (2011). Frozen actions in the arctic linguistic landscape: A nexus analysis of language processes in visual space. *International Journal of Multilingualism*, 8, 277–298.
- Rubdy, R., & Ben Said, S. (Eds.). (2016). *Conflict, exclusion and dissent in the linguistic landscape*. New York: Springer.
- Scarvaglieri, C., Redder, A., Pappenhagen, R., & Brehmer, B. (2013). Capturing diversity: Linguistic land- and soundscaping. In J. Duarte & I. Gogolin (Eds.), *Linguistic superdiversity in urban areas: Research approaches* (pp. 45–74). Amsterdam: John Benjamins.
- Scollon, R., & Scollon, S. B. K. (2003). *Discourses in place: Language in the material world*. London: Routledge.
- Sebba, M. (2010). Discourses in transit. In A. Jaworski & C. Thurlow (Eds.), *Semiotic landscapes: Language, image, space* (pp. 59–76). London and New York: Continuum.
- Shohamy, E. (2015). LL research as expanding language and language policy. *Linguistic Landscape*, 1, 152–171.
- Shohamy, E., & Gorter, D. (Eds.). (2008). *Linguistic landscape: Expanding the scenery*. New York: Routledge.
- Soukup, B. (2016). *English in the linguistic landscape of Vienna, Austria (ELLViA): Outline, rationale, and methodology of a large-scale empirical project on language choice on public signs from the perspective of sign-readers*. Vienna English Working Papers, 25. Retrieved from <http://anglistik.univie.ac.at/views/>
- Spolsky, B. (2009). Prolegomena to a sociolinguistic theory of public signage. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 25–39). New York: Routledge.

- Stroud, C., & Mpendukana, S. (2009). Towards a material ethnography of linguistic landscape: Multilingualism, mobility and space in a South African township1. *Journal of Sociolinguistics*, 13, 363–386.
- Stroud, C., & Mpendukana, S. (2010). Multilingual signage: A multimodal approach to discourses of consumption in a South African township. *Social Semiotics*, 20, 469–493.
- Thrift, N. (2007). *Non-Representational theory: Space, politics, affect*. Abingdon, Oxon and New York: Routledge.
- Zabrodskaia, A. (2014). Tallinn: Monolingual from above and multilingual from below. *International Journal of the Sociology of Language*, 2014, 105–130.



# 41

## Researching Academic Literacies

David Bloome, Gilcinei T. Carvalho, and Sanghee Ryu

### Introduction

A pivotal issue in the conduct of research on learning to use written language in academic domains is whether literacy is conceptualized as essentially an autonomous set of cognitive processes or whether it is conceptualized as diverse social practices involving written language situated in social and ideological contexts. The singular word “literacy” is used when referring to the first conception of written language use (hereafter academic literacy). Researchers holding this conception focus on the interaction of an individual reader/writer and one or more texts. Theoretical debates concern what cognitive processes are involved, what reader factors, what textual factors, and how the pathway from novice to expert is conceptualized. The plural word “literacies” is used when referring to the second conception of written language use (here-

---

D. Bloome (✉)

Department of Teaching and Learning, College of Education and Human Ecology,  
The Ohio State University, Columbus, OH, USA

e-mail: [bloome.1@osu.edu](mailto:bloome.1@osu.edu)

G. T. Carvalho

Universidade Federal de Minas Gerais, Belo Horizonte – MG, Brazil

e-mail: [gilcineicarvalho@gmail.com](mailto:gilcineicarvalho@gmail.com)

S. Ryu

Research Center of Korean Language and Literature Education, Korea University,  
Seoul, South Korea

e-mail: [gratefuleverything@gmail.com](mailto:gratefuleverything@gmail.com)



after academic literacies). Researchers holding this conception focus on the particular social practices involving the use of written language by people in specific academic communities (Street, 2015). They ask in what social and academic situations are which social practices of written language use employed? By whom? With whom? How? For what purposes? With what meanings and social consequences?

Building on the heuristic offered by Lea and Street (2006), the purpose of this chapter is to discuss three perspectives of researching learning to use written language in academic domains. The first perspective is associated with academic literacy. Researching literacy from this perspective focuses on identification and acquisition of underlying, autonomous, cognitive processes and strategies associated with expert use of written language in and across academic domains (Scardamalia & Bereiter, 1991). The second and third perspectives are associated with academic literacies. The second perspective asks how social practices involving the use of written language (hereafter literacy practices) vary across social contexts and how literacy practices within an academic community are acquired (Wingate, 2015). A major framework associated with this second perspective is that of academic socialization. The third perspective views academic literacies as inherently dialectical social practices continuously refracted. Questions here are asked about how literacy practices within a specific academic domain are being refracted and how the tensions inherent to refracting influence meaning-making and the academic domain itself. Since refracting is inherently an historical process (people refract previous uses of written language), researchers focus attention on how people's literacy practices reflect and refract the shared social and cultural ideological contexts of engagement in academic literacy practices.

### **Researching Academic Literacy as Interaction Between an Individual Reader/Writer and a Text(s)**

When academic literacy is defined as essentially an interaction between the reader/writer and an academic texts, attention is focused on the cognitive processes and strategies involved in that interaction (Goldman et al., 2016). Student progress in learning academic literacy is conceptualized as acquisition of the requisite cognitive processes and strategies approaching those employed by an expert in the academic field.

What cognitive processes and strategies experts in an academic field use in their reading and writing continue to be debated. One approach to identifying those cognitive processes has been the use of verbal reports and protocol analysis, asking individuals to think aloud as they are engaged in reading and writing

academic texts (Afflerbach, 2009). Researchers have examined the knowledge experts bring and use in their uses of written language in a designated academic field and how it is distinct from the knowledge of novices (e.g., Alexander, 2003; Shanahan, Shanahan, & Misischia, 2011). At issue is not the enactment of a set of discrete cognitive processes and strategies, but the more complex cognitive processes of the integration of diverse knowledge structures and the construction of an overall knowledge structure (Goldman & Bloome, 1997). Studies of academic texts may reveal hidden demands academic texts may make on readers; such insights are useful if incorporated into studies of how experts actually read, use, and produce those texts (Prior & Bilbro, 2012). Knowledge of the demands that specific academic texts and tasks might make on readers and writers may be used in experimental studies that vary the conditions of interaction between an individual and texts.

Attention is also paid to the context of individual-text interaction, with context conceptualized as factors such as the instructional task, individual differences (e.g., motivation), learning conditions (e.g., how many students are in the class), and the nature of instruction (e.g., the focus, organization, structure, duration of instruction). The social and cultural context, defined as students' ethnic, linguistic, socioeconomic status, and experiential background knowledge, is conceptualized as factors that facilitate or constrain each individual student's use and acquisition of the designated cognitive processes and strategies. As discussed later, this is a different definition of context than taken up in the other two perspectives.

Questions about the teaching and learning of academic literacy from an individual-text(s) perspective have been addressed within a process-product model. Process factors, defined as the broad range of activities and social and psychological processes that occur in the classroom, are statistically linked with products such as a score on a test of reading within a specific academic domain. Factors outside of the classroom that might affect the relationship of a classroom process and a product (such as contextual factors) are labeled "presage" factors. These include the education of the teacher and student socioeconomic and cultural backgrounds. Employing a process-product model for researching academic literacy requires operationalizing and quantifying cognitive processes and strategies, presage factors, and products so that they can be statistically related.

For example, consider the application of an academic literacy perspective in a study of the teaching and learning of argumentative writing in response to literature in a tenth-grade English language arts classroom. Researchers might begin by identifying what cognitive processes and strategies experts in the field employ using case studies, experiments involving varying academic texts, and think-aloud protocols. Similar studies might be conducted

with novices, high school students of various levels of academic achievement. Using a process-product research design, researchers might then explore in a large number of classrooms what process and presage factors are related to students' movement toward acquisition of those cognitive processes and strategies identified as those experts in the field use. A process-product design might be combined with an instructional intervention experiment designed to focus on the acquisition of targeted cognitive processes and strategies, and the experimental condition compared with a control condition.

Recently, researchers have taken as substantive the differences in the nature of individual-text interaction across academic and disciplinary domains (Shanahan et al., 2011). This has led to identification of differing cognitive processes and strategies in the interaction of individuals and texts across academic and disciplinary domains as well as differences in how students acquire acumen with those academic domain-specific cognitive processes and strategies needed to make appropriate meanings of the academic texts within an academic domain (Goldman et al., 2016).

## Researching Academic Literacies as Academic Socialization

The concept of academic literacies begins with the premise that the uses of written language across academic fields are sufficiently and substantively different to constitute heterogeneous practices. The emphasis in researching academic literacies is thus focused on describing the literacy practices within disciplinary fields and how those literacy practices might be acquired (Geisler, 2013; Lea & Street, 1998).

The concept of "literacy practice" is key to this perspective (Street, 1995, 2000). Briefly stated, a literacy practice is a way of using written language in a specific type of social situation that is shared with others and learned (see Baynham & Prinsloo, 2009). A key aspect of literacy practices is that they are social; the basic unit of analysis is not the individual and the text, but people acting and reacting to each other in culturally driven ways (social practices) involving the use of written language (Bloome, Carter, Christian, Otto, & Shuart-Faris, 2005). A literacy practice is thus an abstraction, a shared cultural model for how to use written language, materially realized in social events. Implied in such a view of academic literacies is an assumption of in-group members who share a set of literacy practices. Acquiring the literacy practices of a group, such as an academic community, is also an issue

of social identity (Ivanic, 1998). That is, a person gains the social identity of a member of an academic community by publically engaging in the literacy practices of that academic community in the appropriate situations. Becoming a member of an academic community may involve a kind of apprenticing (Lave & Wenger, 1991) in which people acquire the literacy practices of the community as they move from the periphery toward the center of the community. So conceptualized, researching the acquisition and development of academic literacies focuses on describing how people are socialized into the community's social practices including its literacy practices.

Researching socialization to an academic community of practice has primarily been grounded in theories of learning derived from sociocultural psychology (e.g., Rogoff, 1990; Vygotsky, 1978) and theories of language socialization (e.g., Duff, 2010; Schieffelin & Ochs, 1986). With regard to the first, attention is paid to how zones of proximal development are constructed, for example, through the interaction of teachers and students and how students move toward the mastery of the targeted academic literacy practices. Researchers employing a language socialization framework focus on how people (teachers and students) use both spoken and written language to promote socialization to the culture of the academic community (to its ways of thinking, valuing, acting, believing, and feeling) and how engagement in the social practices, activities, and events of the community socializes people to the literacy practices of the community.

Research on socialization in general, and on acquiring academic literacies in specific, has shown that the socialization process involves active engagement both by those who are members of the community (e.g., teachers) and those who are becoming members. However, research has also shown that some students resist taking on the social identity of an academic community as they view doing so as requiring them to defenestrate their social identity as a member of a cultural or social group associated with their family, home, ethnic, racial, or gender community.

Social context is a key construct within an academic literacies conception. Literacy practices are situated in social contexts. A social context includes multiple levels ranging from face-to-face situations to academic communities to social institutions as well as broadly considered cultural ideological contexts. Social context here includes recognition of multiple and potentially conflicting social and cultural ideological contexts (e.g., the context of the academic community, the context of the student's ethnic

community, as well as national contexts; see Russel, Lea, Parker, Street, & Donahue, 2009). What is key to this definition of social context is that social context is not viewed as separate from and external to the interaction of an individual and a text, but rather that the essential unit of analysis is situated literacy practices. Social context, thus, is not conceptualized as an environment in which literacy events and literacy practices occur, but rather an acknowledgment that any literacy practice is inseparably a part of a history of social events and social practices whose meaningfulness is intimately derived from the cultural ideologies that people have constructed as part of those histories and as part of their lives in and across communities and social institutions.

Ethnographic approaches are often used to identify and describe the literacy practices within an academic community as well as to describe how students are socialized into an academic community and its literacy practices. An ethnographic approach to academic literacies emphasizes the cultural and social aspects of language use, how the people in the academic community define and interpret their literacy practices (an emic perspective), the complex and multi-leveled social contexts of those literacy practices, and how different social contexts relate to each other.

For example, consider research from an academic literacies as socialization perspective of a tenth-grade English language arts classroom. An ethnographic perspective would focus attention on the culture of the classroom and on how the academic literacy practices in that classroom both helped constitute and reflect the classroom culture as representative of the academic community of literature scholarship. The teacher, as a representative of the academic community of literary scholars, brings its literacy practices into the classroom. Because the classroom is a different social context than the academic community of literary scholars, questions are asked about how those literacy practices are recontextualized in the classroom.

From an ethnographic perspective, there are at least two communities to which the students are being socialized: the community of literary scholars and the classroom community. Questions are therefore asked about how the literacy practices of the classroom are reflective of what actually occurs in the community of literary scholars. Within the context of the classroom, the community of literary scholars may be more so an imagined community (an imagined community that may have little basis in the realities of the professional, workplace, or civic communities that any tenth-grade student is likely to join later in life). Questions would also be asked about how the cultural lives of the students outside of the classroom influenced the classroom culture and its literacy practices (see Sample Study 41.1).

### Sample Study 41.1

Castanheira, M. L., Street, B. V., & Carvalho, G. T. (2015). Navigating across academic contexts: Campo and Angolan students in a Brazilian university. *Pedagogies: An International Journal*, 10, 70–85.

#### Research Background

This study examines the university experiences of two groups of nontraditional students whose previous generation had very limited access to university.

#### Research Method

The study uses ethnographic and qualitative methods, including observation and interviewing.

- A central research question was what counted as appropriate writing from the perspectives of the students and from the teachers.
- Data analysis involved close analysis of interview transcripts, student writing, and observed classroom events.

#### Key Result

The university is not a homogenous environment but rather involves a complex range of literacy practices; the study points out tensions involved in language uses, especially the differences between registers and the possibilities of positioning whether and how an author should present his or her point of view. A series of hidden features of academic literacies was revealed; students need access to these hidden features. The study also pointed to differences in views about appropriate uses of written language between students and teachers. Teachers need to understand student views if they are to be successful with teaching appropriate uses of academic literacy practices.

#### Comments

Fundamental aspects of the study's logic of inquiry were, first, to hold as inherently connected the relationship between global and institutional, social and cultural contexts and how people used written language in local situations to communicate and position themselves, and, second, to make visible aspects of written language use in academic contexts that are often hidden and invisible and subsequently prevent some students from acquiring academic literacy practices.

More generally, questions are also raised about what matters, to whom, when, where, and how in the concerted enactment of the classroom literacy practices. That is, does it matter to participants (teacher, students, and others involved in that classroom education) whether the literacy practices of the language arts classroom are those of the literary community? Or, are other things at stake such as socialization to those literacy practices that will provide access to higher education? Another way to conceptualize what is at stake is to conceptualize the acquisition of classroom academic literacy practices as the acquisition of cultural capital that can be exchanged for economic capital, social capital, and symbolic capital (Bourdieu, 1977).

## Researching Academic Literacies as Chronotopic, Dialectical Social Practices

As people use written language within a social situation, how they use it reflects past literacy practices while also addressing the newly evolving social situation. From this perspective, the question to ask about the use of written language is not what literacy practice it is, but rather how does that use of written language reflect and refract the uses of which literacy practices that have gone before, that are anticipated in the future, and of what meaningfulness and social consequences are the dialectics created through the reflection and refraction?

Shifting the focus from the literacy practice itself to the dialectics of reflection and refraction eschews the nominalization of academic literacies and foregrounds academic literacies as a way of acting on and in the specific social events and social contexts in which people find themselves. As such, any engagement with academic literacy practices is an enactment of power relations. In the first and second perspectives, power is defined by the accumulation of either cognitive processes and strategies or social practices, respectively. The larger the accumulation, the more power an individual (or group) has. From the perspective of academic literacies as chronotopic, dialectical social practices, power is better understood as a situated process involving interactional negotiations of what is happening, who has the rights to do what, with whom, how, for what purposes, and with what interpretations and social consequences (see Sample Study 41.2).

### Sample Study 41.2

Gutiérrez, K. D. (2008). Developing a sociocritical literacy in the third space. *Reading Research Quarterly*, 43, 148–164.

#### Research Background

A theoretical construct for addressing differences in the cultural backgrounds of teachers and students is that of “Third Space,” the intersection of what teachers, as representatives of the dominant academic and institutional culture, and students, as historically located in their own cultural, linguistic, and social communities, bring to learning and development such that instructional conversations are transformative spaces. As public universities in the United States consider how to serve the higher education needs of immigrant students, they need to address how to create “Third Spaces.”

#### Research Method

- Ethnographic and social-design study of a four-week summer residence program for migrant high school students.
- Gutierrez analyzed representative classroom interactions and focal students’ writing.
- Data analysis focused on multiple dimensions of the teaching and learning context, including the social organization of the academic field, the complexities of



instructional conversations, the tensions and synergies between the individual student and collective learning community, and student's everyday literacy practices.

### Key Result

Gutiérrez shows how the "Third Space" created at the institute by using students' everyday literacy practices facilitated students' development of critical social thought, a collective identity as historical actors, and extended their repertoire of academic writing. Students were able to move beyond the strictures of a given academic genre and incorporate aspects of writing practices from their own communities providing agency and opportunity for critical reflective practice, which Gutierrez labels "sociocritical literacy."

### Comments

A fundamental aspect of the study's logic of inquiry is that the social, cultural, and political contexts of the lives of students and teachers are not distinct from their engagement in learning to use written language in academic learning. A second fundamental aspect is that instructional contexts need to value students' agency and the dialogic interaction between teacher and students in refracting extant academic literacy practices.

Part of these social consequences include the negotiation of the *in situ* cultural ideology (i.e., the broader cultural interpretive framework to be employed in giving meaning to what is happening and in defining the people involved). These negotiations of power relations occur in a historical context; that is, what has come before and what is anticipated to come later plays a role in the negotiations.

Researchers exploring academic literacies from this perspective have used critical ethnographic approaches and critical discourse analysis to make visible the power relations inherent in the enactment of literacy practices. Like ethnographic research and discourse analysis in general, researchers describe and interpret the uses of written language in specific social situations and their social contexts grounded in an emic perspective. Of special focus are (a) the dialectics that arise through how literacy practices are adapted (refracted) to the specific social situation including its past and anticipated history and (b) the construction of chronotopes and their implications for personhood. Bakhtin (1935/1981) used the term "chronotope" to refer to how an author crafted how a protagonist moved through time and space. As used in the study of academic literacies, chronotope refers to a cultural ideology that provides an interpretive framework for movement through time and space with implications for the construction of personhood. By personhood we mean the explicit and implicit way of defining what it means to be human, what are the kinds of people in the world, what characteristics do they have, what rights and privileges are inherent to being human. As Williams notes, "a definition



of [written] language is implicitly or explicitly, a definition of human beings in the world” (1977, p. 21). Some scholars have argued that promulgating a definition of personhood is a major dimension of the teaching and learning of academic literacies, although implied and unacknowledged, and that debates about how to teach reading and writing and how to conceive of academic literacies are more so debates about personhood (Egan-Robertson, 1998).

For example, consider again the tenth-grade language arts classroom. The teacher had structured her instructional units to build on each other, moving from learning how to write a literary argument to using the writing of a literary argument to learn. She had also created instructional activities across the academic year whereby students had to incorporate the experiences of their diverse cultural communities into the academic literacy practices they were learning while respecting and appreciating how others had adapted those academic literacy practices. Implied is a chronotope of movement through time and space. The student was not distinct from her/his family and community and thus conceptualized themselves as part of that history. Thus, in the interpretation and discussion of a short story, an African-American student invoked the history of slavery, a student of Cambodian heritage invoked her family’s history of escaping the Killing Fields and immigrating to the United States, a Somali student invoked her religion of Islam and the attacks upon her and other Somalis in her local community, students whose families had moved from Appalachia incorporated those family experiences to the interpretation and discussion of that short story, and so forth for other students with other cultural and historical backgrounds. Yet, at the same time, the students brought to bear their experiences in previous grades with reading, interpreting, and discussing literature, while recognizing that at times these multiple framings could be brought together while at other times they remained distinct. Such a conception of personhood eschews a deficit model as no one framing for interpreting and discussing literature is viewed as *the* framing by which others are evaluated and it allows students to define themselves as mutually occupying valued positions and identities in the academic community, the classroom community, and their home ethnic/cultural and historical community. The personhood of students, here, is defined as involving both the capacity for and movement toward actualization of empathy toward others’ interpretive frames. They are conceptualized as human beings who have moved from an egocentric and ethnocentric view of the world to a pluralistic view constituted by the recognition that engagement in literacy practices involves a dialectic among diverse refractions of given academic literacy practices (example from Newell et al., 2015). By contrast, in a different tenth-grade literature classroom, students were being taught the features of literary genres, with each instructional unit addressing a different genre. The movement through time and space in that classroom can be characterized as cyclical (one

instructional genre unit followed by the next that could be sequenced in any order), with students acquiring increasing knowledge of genre features. Implied here is a definition of personhood in which students are repositories of bits of knowledge rather than actors with others on the worlds in which they live. Inherent in this repository definition of personhood is a hierarchy (some students hold more bits in their repository than others) and a deficit model (some students have deficiencies in depositing enough bits into their repository). Understanding academic literacy practices involves not only the learning process but how teaching choices influence students' meaning-making.

## Final Comments

We began by distinguishing between academic literacy and academic literacies, with the former focusing attention on the interaction of the individual and a text (or texts) and the latter defining the use of written language as situated social practices. We then presented three perspectives, with the first, which we labeled an individual-text perspective, focused on the cognitive processes and strategies employed by individuals in their interaction with a text(s) potentially mediated by a broad range of contextual factors. From this perspective, the nature of the texts and tasks in an academic domain places particular cognitive demands upon readers and writers. What students are taught and learn are those cognitive processes and strategies that experts in the particular academic domain employ in their uses of written language. The second perspective, which we labeled academic literacies as academic socialization, conceptualizes uses of written language as social practices that are sufficiently and substantively distinct across communities such that knowing the literacy practices in one academic community does not provide access to those of another. As such, there is not a literacy but rather literacies. Students are socialized in their classrooms to the academic literacy practices of an academic community. This socialization process is complex in part because teachers and students need to orchestrate the relationship of three community contexts: the academic community (e.g., the academic community of literary scholars), the classroom community (e.g., a tenth-grade English language arts classroom), and the students' home communities (e.g., family literacy practices). The third perspective, which we labeled academic literacies as chronotopic, dialectical social practices, focuses attention of the dialectics that are inherent to the social and sociolinguistic processes of reflection and refraction of extant academic literacy practices. As students engage in the academic literacy practices that their teachers make available, they have to adapt those practices to the particular social, cultural, and historical situations in which they find

themselves, and in the process of adaptation tensions arise which provide potential for the social construction of new meanings (Gutierrez, 2008). Students' refractions of academic literacy practices may also bump up against the ways in which other students have refracted academic literacy practices which, depending on whether the instructional context has promoted appreciation and mutuality, has the potential to create new and multiple understandings both of an academic text(s) and academic community as well as cultural worlds in which the students live. The teaching and learning of academic literacy from this third perspective is also framed by recognizing that the social interaction of teachers and students around academic literacy practices constructs an implied chronotope for defining personhood.

Given the differences in underlying epistemological and ontological assumptions across the three perspectives (Lillis & Scott, 2007), a synthesis of them would be superficial. The questions asked, the units of analysis, what counts as data, the interpretive frameworks, and the conception of what is being studied all differ. There are also differences in how building knowledge in the field is conceptualized and how that knowledge might be employed by educators. Researching from the perspective of academic literacy as cognitive processes and strategies contributes to the knowledge base in the fields of educational studies and literacy studies deductively and inductively by incrementally adding knowledge warranted by empirical study. From this perspective, a research study either adds knowledge to a theoretical model, contradicts one or more theoretical constructs of a theoretical model, or provides findings such that a new theoretical model is needed. From the perspective of the study of academic literacies (both the second and third perspectives discussed earlier), the knowledge base is built abductively (analogic reasoning) through a series of situated cases each of which provides insights about social and cultural patterns that can inform educators and others about how uses of written language might play out within analogous social and educational situations (Lillis, Harrington, Lea, & Mitchell, 2015).

In our view, rather than force a superficial synthesis across perspectives, researchers of academic literacy(ies) might conceptualize research on academic literacy(ies) metaphorically as a kind of heteroglossia within a chronotopically diffuse space (i.e., with multiple definitions of movement through time and space and of personhood). Any particular research study, whatever its perspective, adds a voice to the evolving conversation. The common starting point to all perspectives is considering the complexities of literacy practices even in academic contexts where it seems to be apparently more homogeneous or more regulated.

## Resources for Further Reading

Lea, M., & Street, B. (2006). The “academic literacies” model: Theory and application. *Theory into Practice, 45*(4), 368–377.

The authors extend the concept of academic literacies from its original location in higher education to K-12 education. The authors note three models for conceptualizing academic literacies: a study skills model which foregrounds individual cognitive skills, an academic socialization model which foregrounds students’ acculturation into disciplinary and subject area-based literacy practices, and what they call an academic literacies model that foregrounds meaning-making, identity, power, and authority.

Lillis, T., & Scott, M. (2007). Defining academic literacies research: Issues of epistemology, ideology and strategy. *Journal of Applied Linguistics, 4*, 5–32.

This article addresses the confusion around the multiple definitions of academic literacies. It does so by first considering the geopolitical and historical contexts of the term, attending to the increase of students attending higher education and the use of deficit models to guide curriculum, instruction, and policy. The authors then consider an epistemology for academic literacies located in literacy as social practices with an emphasis on ideology and transformation.

Newell, G., Bloome, D., Hirvela, A., with VanDerHeide, J., Wynhoff Olsen, A., & Lin, T.-J. (with Buescher, E., Goff, B., Kim, M., Ryu, S., & Weyand, L.) (2015). *Teaching and learning argumentative writing in high school English language arts classrooms*. New York: Routledge.

This book examines the teaching and learning of argumentative writing differentiating between structuralist models, in which the focus is on the elements of an argument (claim, warrant, evidence, etc.), and social practice models, in which the focus is on the substance of the argument and the social relations of writers and readers. The authors discuss the underlying rationality for using argumentation for academic learning.

Wingate, U. (2015). *Academic literacy and student diversity: The case for inclusive practice*. Bristol: Multilingual Matters.

This book argues for a change in policy and practice with regard to how academic literacies are taught. It notes that the context for conceptualizing an academic literacies pedagogy is recognition of the ethnic and linguistic diversity of students. It addresses common misunderstandings regarding students' academic literacy needs, existing pedagogical models, and proposes a new model labeled an inclusive model.

## References

- Afflerbach, P. (2009). Verbal reports and protocol analysis. In M. Kamil, P. Mosenthal, P. Pearson, & R. Barr (Eds.), *Handbook of reading research, volume III* (pp. 163–179). New York: Routledge.
- Alexander, P. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher*, 32, 10–14.
- Bakhtin, M. (1935/1981). In M. Holquist (Ed.), *The dialogic imagination: Four essays by M. M. Bakhtin*. Austin: University of Texas Press.
- Baynham, M., & Prinsloo, M. (2009). Introduction. In M. Baynham & M. Prinsloo (Eds.), *The future of literacy studies* (pp. 1–20). New York: Palgrave Macmillan.
- Bloome, D., Carter, S., Christian, B., Otto, S., & Shuart-Faris, N. (2005). *Discourse analysis and the study of classroom language and literacy events A Microethnographic approach*. Mahwah, NJ: Erlbaum.
- Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge: Cambridge University Press.
- Castanheira, M. L., Street, B., & Carvalho, G. T. (2015). Navigating across academic contexts: Campo and Angolan students in a Brazilian university. *Pedagogies. An International Journal*, 10, 70–85.
- Duff, P. (2010). Language socialization into academic discourse communities. *Annual Review of Applied Linguistics*, 30, 169–192.
- Egan-Robertson, A. (1998). Learning about culture, language, and power: Understanding relationships among personhood, literacy practices, and intertextuality. *Journal of Literacy Research*, 30, 449–487.
- Geisler, C. (2013). *Academic literacy and the nature of expertise: Reading, writing, and knowing in academic philosophy*. Routledge.
- Goldman, S. R., & Bloome, D. (1997). Learning to construct and integrate. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer* (pp. 169–182). Washington, DC: American Psychological Association.
- Goldman, S. R., Britt, A., Bown, W., Cribb, G., George, M., Greenleaf, C., et al. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist*, 51, 219–246.

- Gutierrez, K. (2008). Developing a sociocritical literacy in third space. *Reading Research Quarterly*, 43, 148–164.
- Ivanic, R. (1998). *Writing and identity. The discursual construction of identity in academic writing*. Amsterdam and Philadelphia: John Benjamins.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, England: Cambridge University Press.
- Lea, M. A., & Street, B. (2006). The ‘academic literacies’ model: Theory and applications. *Theory into Practice*, 45, 368–377.
- Lea, M. R., & Street, B. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education*, 23, 157–172.
- Lillis, T., Harrington, K., Lea, M., & Mitchell, S. (Eds.). (2015). *Working with academic literacies. Case studies towards transformative practice*. Fort Collins, CO: The WAC Clearinghouse.
- Lillis, T., & Scott, M. (2007). Defining academic literacies research: Issues of epistemologies, ideology and strategy. *Journal of Applied Linguistics*, 41, 5–32.
- Newell, G., Bloome, D., Hirvela, A., Van Der Heide, J., Wynhoff Olsen, A., & Lin, T.-J. (with Buescher, E., Goff, B., Kim, M., Ryu, S., & Weyand, L.). (2015). *Teaching and learning argumentative writing in high school English language arts classrooms*. New York: Routledge.
- Prior, P., & Bilbro, R. (2012). Academic enculturation: Developing literate practices and disciplinary identities. In M. Castelló & C. Donahue (Eds.), *University writing: Selves and texts in academic societies* (pp. 19–31). Brill. Retrieved July 13, 2016, from <http://booksandjournals.brillonline.com/content/books/b9781780523873s003>
- Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. New York: Oxford University Press.
- Russel, D. R., Lea, M., Parker, J., Street, B., & Donahue, T. (2009). Exploring notions of genre in “academic literacies” and “writing across the curriculum”: Approaches across countries and contexts. In C. Bazerman, A. Bonini, & D. Figueiredo (Eds.), *Genre in a changing world* (pp. 395–423). Fort Collins, CO: The WAC Clearinghouse/Parlor Press.
- Scardamalia, M., & Bereiter, C. (1991). Literate expertise. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 172–194). Cambridge, UK: Cambridge University Press.
- Schieffelin, B., & Ochs, E. (1986). Language socialization. *Annual Review of Anthropology*, 15, 163–191.
- Shanahan, C., Shanahan, T., & Misischia, C. (2011). Analysis of expert readers in three disciplines. *Journal of Literacy Research*, 43(4), 393–429.
- Street, B. (1995). *Social literacies: Critical approaches to literacy in development, education and ethnography*. London: Longman.
- Street, B. (2000). Literacy events and literacy practices: Theory and practice in the ‘New Literacy Studies’. In M. Martin-Jones & K. Jones (Eds.), *Multilingual literacies: Reading and writing different worlds* (pp. 17–30). Philadelphia: John Benjamins Publishing Company.

- Street, B. (2015). Academic writing: Theory and practice. *Journal of Educational Issues*, 1, 110–116.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Williams, R. (1977). *Marxism and literature*. Oxford: Oxford University Press.
- Wingate, U. (2015). *Academic literacy and student diversity: The case for inclusive practice*. Bristol: Multilingual Matters.

# Index<sup>1</sup>

## A

- AAAL, *see* American Association for Applied Linguistics
- Absolute fit indices, 472–474
- Abstracts, 201, 202
- Academic Collocation List* (ACL), 361
- The *Academic Formulas List*, 361
- Academic Keyword List (AKL), 361, 372n3
- Academic literacies, 887–898
- academic socialization, 890
  - chronotopic practices, 894
  - definition of, 888
  - dialectical social practices, 894
  - interaction between individual reader/writer and text(s), 888
- Academic socialization, 888, 890–893, 897
- Acceptability judgments, 289, 294, 300
- See also* Grammaticality judgments
- Acoustic analyses, 726–727
- Acquisition, 361, 570, 596, 731, 813, 888
- metaphor, 41
  - planning, 813
- See also* Instructed second language acquisition (ISLA)
- Action research, 84, 89, 125, 270, 712, 869
- Active training paradigms, 303
- “Adjusted *R* Square,” 409
- Adult language acquisition, 690–692
- Advanced statistics, 459–496
- Agency, loss of, 168
- Agreement rate, 549
- Akaike information criterion, 473
- Alpha, 557, 559
- Alternate sign languages, 795
- American Association for Applied Linguistics (AAAL), 7, 8, 20
- American Community Survey (ACS), 744, 751
- American Council on the Teaching of Foreign Languages (ACTFL), 853
- Proficiency Guidelines, 854
- American Educational Research Association, 849
- American Psychological Association (APA), 560, 849

---

<sup>1</sup>Note: Page numbers followed by ‘n’ refer to notes.



- American Sign Language, 795
- Analysing stage, of empirical research, 14
- Analysis of variance (ANOVA), 137, 277, 395, 409, 410, 417, 501–518, 524, 535, 559, 756, 804
- assumptions of, 505–506
  - effect size and, 511–514
  - output and interpretation of results, 509–511
  - procedure, 409
  - regression, *t*-tests and GLM and, 514–515
  - significance testing with, 508–509
  - use and misuse of, 515–518
- Anderson-Darling (AD) test, 536
- Anderson-Rubin*, 451
- Anglophone countries, 83
- Annotated corpus, 40
- Annual Review of Applied Linguistics* (journal), 778
- ANOVA, *see* Analysis of variance
- Applied linguistics (AL), 5–25, 80, 82–84, 89, 92, 93, 103, 270, 287, 365, 377, 423, 523, 541, 542, 572, 845, 869
- correlation and simple linear regression in, 395–418
  - area view and notion of shared variance ( $r^2$ ), 402
  - assumptions of, 415–416
  - coefficient *a*, 414
  - coefficient *b*, 414
  - coefficient  $\beta$ , 414
  - computation and significance testing of  $R^2$ , 409–411
  - graphical representation of Pearson's *r*, 400–402
  - Model Summary, 406
  - overview of, 395–397
  - Pearson's Product-Moment Correlation Coefficient (*r*), 397
  - regression coefficients, 411–413
  - R* Square, 408
  - simple linear regression, 403
- evolution of research, 6
- methodological reform in, 7–8
  - primary and secondary research, 10
  - published books in, 9
  - research approaches and methodologies, 10 (*see also specific entries*)
- Applied Linguistics* (journal), 6, 7
- Applied research, 11
- Approximate replications, 148, 149
- non-supportive, 152, 153
  - supportive, 152, 153
- Area under the curve (AUC), 291
- Area view, 402
- Argument-based approach, 850, 851
- Artificial grammar research, 304
- Artificial language, 304–305
- Assessment literacy, 858–860, 862
- Asymmetric participation format, 226
- Attractiveness, 290, 856, 870
- Attractive questionnaire, 280–281
- Audience, literature review, 139
- Audio-recorded verbal interaction, 630
- Autobiographical research writing, 600
- Autobiography, 240, 599
- interviews, 228, 229
  - studies, 600–601
- Auto-cyberethnography, 786, 787
- Autoethnography, 600–601, 786
- Average reduced frequency (ARF), 360
- B**
- Bad data, 523
- Balanced bilingual, 683
- Bartlett*, 451
- Bartlett's test, 433, 441
- of sphericity, 434
- Basic research, 11
- Bayesian inference, 8
- Before the Dissertation*, 195
- Beta ( $\beta$ ) coefficients, 415

- Beverly Costa, 276
- Bias-corrected and accelerated (BCa) method, 533
- Biber Tagger, 570–571
- Big ‘D’ Discourse, 42–43  
analysis, 48
- Bilingualism and Emotion  
Questionnaire (BEQ),  
273–275
- Bilingualism and multilingualism,  
681–697  
approaches to researching, 685–694  
core issues in, 682–684  
future challenges, 697  
methodological considerations and  
challenges, 695–697
- Bilingualism research, 234, 269, 695
- Bilingual Language Interaction  
Network for Comprehension  
of Speech (BLINCS), 687
- Bilingual languages, 274, 290, 298,  
303, 745, 748, 796, 800, 814,  
818
- Bilinguals, typology of, 696
- Biographical case studies, 602, 603
- Biographies, 88, 596
- Black South African English (BSAE),  
725
- BLINCS, *see* Bilingual Language  
Interaction Network for  
Comprehension of Speech
- Bootstrapping, 8, 25, 392, 502,  
531–534, 536
- British Association for Applied  
Linguistics, 8
- C**
- CA approach, 87, 832, 834, 838–840
- California Healthcare Interpreting  
Association, 765
- Cambridge International Corpus*, 360
- the *Cambridge Learner Corpus* (CLC),  
365
- Canadian Department of  
Multiculturalism and  
Citizenship, 745
- Canary Row*, 801
- Case studies, 85, 104, 116, 270, 276,  
364, 376, 602, 606, 725–726,  
730, 743, 746, 879
- Categorical data, 524, 525, 536, 538  
measuring frequency differences,  
526–530  
simple frequencies, 526  
statistics for, 530
- Categorical variables, 530
- Causal indicators model, 425
- CDA, *see* Critical discourse analysis
- Center for Applied Linguistics (CAL),  
848
- CFA, *see* Confirmatory factor analysis
- CHAT application, 693
- Chatbots, 58–62, 64–68, 70, 72
- CHILDES system, 688
- Childhood bilingualism, 688–690
- Chinese Communist Party, 724
- Chi square test, 473, 486–487, 492,  
527–529
- Chomskyan competence approach, 692
- Chronotopes, 895
- Chronotopic practices, 894–897
- Circumstantial bilinguals, 768, 769
- CLAN application, 693
- Classical item analysis (CIA), 560
- Classical test theory (CTT), 417, 558,  
561, 856–857
- Classroom, 85, 89, 91, 92, 250, 261,  
281, 287, 359, 362, 369, 370,  
525, 557, 559, 572, 622, 625,  
629, 641, 664, 669, 672, 674,  
725, 732, 749, 793, 795, 799,  
801, 820, 853, 854, 859, 860,  
889, 890, 892–894, 896, 897  
instruction, 108  
language learning, 72  
research studies, 799
- Closed niche group, 276–277

- Closed role-plays, 837  
 Cluster analysis, 645  
 Coarse score, 451  
 Code-mixing, 692–693  
 Coders, 543, 544, 554–555  
 Code switching, 240, 276, 290, 529, 620, 683, 684, 687, 691, 692, 721  
 Coefficient, 408, 411  
    $a$ , 414  
    $b$ , 414  
    $\beta$ , 414–415  
 Cognitive approach, 832  
 Cognitive processes, 288, 297, 301  
 Cognitive revolution, 850  
 Cognitive strategy, 454  
 Cohen's  $d$  effect size, 63  
 Coh-Metrix, 571–572  
   indices, 574  
 Collecting stage, of empirical research, 14  
 Collocation, 359, 361, 362, 364, 368, 369, 569, 820  
*Common European Framework of Reference for Languages* (CEFR), 366, 854  
 Communalities, 437  
 Communality variance, 426  
 Communication, 127, 229, 234, 240, 252, 274, 275, 359, 375, 376, 378, 379, 382–384, 387, 588, 615, 616, 629, 634, 721, 728–731, 761–766, 769, 800, 830, 849, 872, 877  
 Communicative blunders, 242  
 Communicative competence, models and frameworks of, 853  
 Community language, 741–757  
 Comparative fit index, 473  
 Complementary approaches, 561  
 Completely standardised solution, 461  
 Composite scores, 451  
 Computer-aided error analysis (CEA), 364  
 Computer-mediated communication, 381  
 Computer-mediated discourse analysis (CMDA), 382  
 Computer programmes, 426  
 Conceptual knowledge, 125  
 Conceptual replication, 149, 150, 153  
   non-supportive, 152, 153  
   supportive, 152, 153  
 Concurrent designs, 107–109  
 Concurrent validity, 857  
 Confidentiality issues, 276  
 Confirmability, 82  
 Confirmatory factor analysis (CFA), 70, 395, 424, 425, 454, 455, 459–464, 852  
   critical requirements and statistical assumptions for, 468–469  
   EQS and, 477–495  
   model of reading performance, 463, 464  
   software for, 469  
   steps in, 469–477  
 Conjoint interviews, 234  
 Connective ethnography, 785  
 Consequential validity, 858  
 Consistent Akaike information criterion, 473  
 Construct validation, 850  
*Contemporary Corpus of American English* (COCA), 361  
 Content analysis, 89  
 Content validity, 849, 857  
 Context, 378  
 Context-specific research issues, 840  
 Contingency table, 526, 527  
 Contrastive Interlanguage Analysis (CIA), 366  
 Convergent validity, 424  
 Conversational analytic research, 796  
 Conversation analysis (CA), 38–40, 85–87, 384, 615, 617, 647, 671, 730, 781, 832

- Corpora, 723
- Corpus, 813  
 analysis, 382  
 data, 525  
 linguistics, 39, 40, 110, 705,  
 723–725
- Corpus-based analyses, 363
- Corpus-based approaches, 731
- Corpus-based inquiry, 728–730
- Corpus-derived frequencies, 360
- Corpus of Online Registers of English  
 (CORE), 526
- Corpus research methods, 360  
 for language teaching and learning,  
 359–370  
 challenges and controversial  
 issues, 369  
 lexis-grammar in context, 363  
 role of frequency, 360  
 role of learner corpus data, 364
- Correction equation, 546
- Correlation, 136, 278, 368, 395, 396,  
 400, 401, 416, 424, 426–428,  
 434, 443, 444, 446, 452, 454,  
 531, 545–549, 553, 557, 559,  
 574, 583, 584, 586, 857,  
 871  
 coefficient, 545, 546
- Correlation matrix, 428, 431–434,  
 440, 444, 470, 583  
 and EFA statistical requirements,  
 431
- Correlations, between factor scores and  
 composite scores, 453
- Co-speech gesture, 794–796
- Council of Europe (2003), 855
- Create A Research Space (CARS), 202,  
 204
- Creating Contexts: Writing Introductions  
 Across Genres*, 202
- Credibility, 82
- Criterion-referenced (CR), 557
- Criterion-related validity, 849
- Critical discourse analysis (CDA), 88,  
 89, 640, 782, 816, 895
- Critical ethnography, 818
- Critical incidents, 605
- Critical language policy (CLP), 814,  
 815
- Critical languages, 748, 751
- Critical period hypothesis, 44, 690,  
 692
- Critical theory, 90
- Cronbach's alpha coefficient, 430  
 of each factor, 446–449
- Cross-Cultural Speech Act Realization  
 Project (CCSARP), 830, 831,  
 837
- Cross-factor loadings, 426
- Cross-language semantic priming, 298
- Cross-linguistic patterns, 290
- Cross-sectional research, 11
- Cross-validation, 289
- Cultural context, definition of, 889
- CUSUM technique, disadvantage of,  
 711, 715n1
- D**
- Data analysis, 10, 12, 14, 17, 18, 20,  
 108, 110, 112, 136, 137, 258,  
 378, 423, 560, 568, 604–607,  
 640, 878, 879, 893, 894  
 methods, 92, 598
- Data coding, reliability analysis of  
 instruments and, 541–562  
 classical test theory reliability, 561  
 external reliability, 543–555  
 internal-consistency reliability, 555  
 types of reliability, 543
- Data collection, 648, 800–801, 840  
 approach, 835–838  
 methods, 90, 91, 639  
 sites, 747, 748
- Data gathering, 89
- Data mining, 229

- Data reuse and participants' right, 177–178
- Data, types of, 427
- Data visualisation, 22
- 'D' discourse analysis, 43, 48
- Decision task, behavioral measures, 288–289
- Degrees of freedom, 471
- Democratic language policy, 815
- Dependability, 82
- Descriptive forensic linguistics, 704, 706
- Descriptive statistics, 427, 432, 469, 506, 508
- Design, 82, 84, 103, 104, 106–109, 111–113, 115–117, 125, 126, 133, 270, 276, 277, 288, 289, 293, 296, 298, 301, 305, 359, 360, 362, 366, 369, 370, 387, 409, 411, 418, 532, 542, 544, 548, 554, 560, 561, 727, 852, 870, 890
- Designing stage, of empirical research, 13–14
- Dialectical social practices, 894
- Diary study, 688–689
- Difficulty factor, 427
- Digital discourses research and  
     methods, 375–387  
     approaches, 382–383  
     challenges and controversial issues, 381–382  
     context, 378  
     current/core issues, 377  
     data collection, processing and analysis, 383–384  
     research ethics, 384–387  
     text and interaction, 379–381
- Digital ethnography, 382, 385, 387
- Digital literacies, 261, 380
- Digital mapping, 22
- Digital media, 227, 376, 377, 384, 387
- Digital tools, 376
- Direct Oblimin, 443, 444
- Discourse analysis, 89, 380, 382, 384, 386, 568, 597, 598, 730, 816, 817, 871, 879, 895
- Discourse completion tasks (DCTs), 833, 834, 836, 837, 853
- Discriminant validity, 424
- Disseminating stage, of empirical research, 14
- Dissertation, *see* Doctoral dissertation
- Dissertations, 746, 747
- Distribution-free data, 538
- Distribution, normality of, 427
- Distribution tests, 535
- Diverse theoretical approaches, 832
- Doctoral dissertation, 183, 184, 186, 187, 192, 195, 196n1
- Double correction, 546
- Dual system hypothesis, 689–690
- Dummy coding, 530
- E**
- EAL, 203, 212, 215
- Early participant observation, 253
- Ecological validity, 227, 271, 708
- Effect indicators model, 424
- Eigenvalue less than 1, 436
- Elicited data techniques, 692
- Elicited imitation (EI) technique, 293, 328–331  
     considerations, 330–331  
     history and purpose of, 328–329  
     validation, 329–330
- ELL, *see* English language learners
- ELT Journal*, 125
- Embedded designs, 108
- Emblems, 795
- Embodied cognition, 39, 40
- Embodiment, 876
- EmoLex, 578
- Emotional self, 263
- Empirical knowledge, 125, 129

- Empirical research, stages of, 13
- Empirical study, 58–60
- English as a lingua franca (ELF), 369, 720–722, 728, 733, 860  
methodologies of, 727–732
- English as an international language (EIL), 203, 720–722, 733  
methodologies of, 727–732
- English as a second language (ESL), 108, 624, 845
- English for Academic Purposes (EAP), 206, 212, 213
- English language classroom, 624
- English language learners (ELL), 170, 171
- English Profile Programme, 366
- Entextualization, 34, 35
- Epistemic relativism, 230
- Epistemic reorientation, 619
- Epistemologies, 40–45, 81, 82, 199, 812
- EQS program, 465, 469  
analysis in, 485  
CFA in, 477–495  
model specifications, 483, 484  
opening data file in, 478  
outputs, 483–488  
spreadsheet, 479
- Equamax, 443
- ERP, *see* Event-related potential
- Error editor, 364
- Error taxonomies, 365
- ESRC Centre for Research on Bilingualism in Theory & Practice, 693
- Ethical applied ethical research, historical development, 166
- Ethical applied linguistic research, 163–180  
challenges and controversies, 169–179  
current core issues, 167–169
- Ethical research, 241  
practices and social change, 22
- Ethics, 8, 14, 17, 276, 280, 382, 765, 770, 811, 818, 821  
guidelines, 164, 171  
training, 165, 167
- Ethnographic approaches, 892
- Ethnographic-based inquiry, 730–731
- Ethnographic inquiry, 725–726
- Ethnographic interviews, 228, 229
- Ethnographic methods, 750, 879
- Ethnography, 84, 86, 250, 251, 259, 615, 785, 787, 816, 871, 876  
of language policy, 815
- Ethnolinguistic communities, 743
- Ethnomethodology, 615
- Event-related potential (ERP)  
technique, 44, 288, 294–297, 686
- Evidence-based practice, 124
- Exact replication, 148, 151  
non-supportive, 153  
supportive, 153
- Excavation, 229
- Excel files, 272
- Expanding circle, 720
- Explanatory designs, 109
- Explicit knowledge, 690
- Exploratory designs, 111
- Exploratory factor analysis (EFA), 417, 423–455  
common *vs.* error factors in, 425  
composite scores, 451  
computer programmes, 426  
and confirmatory factor analysis (CFA), 424  
correlation matrix and EFA  
statistical requirements, 431–434
- Cronbach's alpha coefficient, 446
- entering, checking and cleaning data, 428
- extraction method, 434
- factor scores, 451

- Exploratory factor analysis (EFA) (*cont.*)  
 item-level descriptive statistics,  
 430–431  
 models, 455  
 naming a factor, 449–450  
 number of factors, 436–442  
 Pearson correlation, 453  
 and principal component analysis  
 (PCA), 424–425  
 reliability coefficient, 430–431  
 requirements and statistical  
 assumptions for, 426–428  
 rotated factor matrix, 446–448  
 rotation method, 442  
 uses of, 423–424
- External reliability, 543
- Extraction method, 434–437
- Eye movement data, 290, 291,  
 299–301
- Eye tracking, 686  
 data, 290, 674
- F
- Face-threatening act (FTA), 259
- Face-to-face or in-person interview, 227
- Face validity, 849, 857
- Factor analysis, 433, 851
- Factor correlation matrix, 444
- Factor extraction method, 427, 435,  
 436
- Factor loading, 425, 426, 444–447,  
 454
- Factor rotation, 435, 444
- Factor scores, 451–453
- Fieldnotes, 86, 249–252, 641, 879  
 analysing, 258–260  
 current/core Issues, 250–252  
 essential processes/steps, 252–256  
 including in, 262–263  
 reliability/validity, 257, 258  
 research background, 254, 260  
 research method, 254, 260  
 research question, 260  
 technology and, 263  
 writing, 256–257, 261–262
- Final regression analysis, 586
- Fine-grained clausal indices, 577
- Fine-grained phrasal indices, 577
- First language (L1), 173, 202, 239,  
 278, 279, 283, 294, 297, 298,  
 300, 301, 304–306, 329, 341,  
 350, 351, 362, 367, 453, 575,  
 720, 727, 732, 754, 800, 848
- Fisher's exact test, 529, 530
- Fishman's Legacy, 742–745
- Five-factor model, 447
- Fixations, 290
- Fleck, S., 31–50
- fMRI, *see* Functional magnetic  
 resonance imaging
- Focus group Interviews, 225–242, 723  
 research method, 237
- Focus, literature review, 139
- Follow-up questionnaires, 114
- Foot pedal, 624
- Foreign language, 273, 274, 276, 290
- Forensic linguistics, 250, 703–714  
 action research, disciplinary  
 engagement and knowledge  
 mobilisation and, 712–714  
 data identification and collection,  
 706–708  
 ethical issues in data collection and,  
 708–710  
 quantitative analysis and statistics  
 and, 710–711  
 research, 703–705
- Fractional view, of bilingualism, 682,  
 683
- Frequency, 201, 205–207, 210, 259,  
 278, 279, 288, 299, 359, 366,  
 368, 369, 372n1, 382, 525,  
 569–571, 576, 577, 582, 724,  
 831, 870, 871  
 role of, 360–363

- F*-test statistic, 510  
 Full-blown ethnographies, 264  
 Functional magnetic resonance imaging (fMRI), 686  
 “Fundamental Considerations in Testing for English Language Proficiency of Foreign Students,” 848
- G**
- Gender imbalance, 273  
 Generalizability theory (G-theory), 561  
*General Service List* (GSL), 360  
 Genre analysis, 85, 200, 384  
*Genre Analysis: English in Academic and Research Settings*, 199  
 Gentrified multilingualism, 878  
 Geographic Information Systems (GIS), 22  
 Gesture, 383, 621, 645  
   coding, 802, 805  
   emblems, 795  
   pantomimes, 795  
   phases, 794, 797  
   research background, 797, 803  
   research method, 798, 803  
   research problems/gaps, 797, 803  
   signs, 795  
   speech-linked, 795  
   types of, 794–795  
   types of analysis, 804  
 GIS, *see* Geographic Information Systems  
 Goals, literature review, 139  
 Goodness-of-fit criteria, 472–474  
*Google Forms*, 270, 280  
 Google Street View, 22, 874  
 Google Translate, 21  
*Grammar and Beyond 4* (2013), 365  
 Grammar learning paradigms, 304  
*Grammar of Visual Design*, 876  
 Grammatical competence, 853  
 Grammatical elicitation techniques, advantage of, 693  
 Grammaticality judgments, 314–325, 333n4, 750  
 Grammatical knowledge, 293, 330  
 Grant-funded study, 625  
 Granularity, 38  
 Graphics, 61  
 Greek foreign language, 831  
 Grounded narrative inquiry, 598  
 Grounded theory, 81, 82, 87, 88, 92, 104, 259, 555, 598  
 Group differences, analyzing, 501–518  
 Group interviews, 234  
 G-test, 528
- H**
- $h^2$ , 426  
 Habits of mind, 31–50  
*Heritage Language Journal (HLJ)*, 742, 746, 747, 749–751  
 Heritage languages (HL), 293, 741–757  
   in dissertations, 750–752  
   learning, 83  
 Heteroglossia, 898  
 HL-CL Articles, 746, 747  
 HL-CL research, 752–753  
   research background, 753  
   research method, 753, 754  
   research problems/gaps, 753  
 HL-CL schools, 754  
 Homogeneity, 556  
   of variance, 505–506  
 Human interaction, 617, 619  
 Hypothetical population distributions, 531
- I**
- Identity, 84, 211, 230, 377, 595, 616, 720, 777–787, 852, 880, 891



- Identity (*cont.*)  
 case study and, 779–780  
 challenges and controversial  
 decisions, 782–784  
 conversational analysis (CA) and,  
 781  
 critical discourse analysis and, 782  
 future directions, 785–787  
 narrative inquiry and, 780 (*see also  
 specific entries*)
- Idiolect, 707
- Ill-constructed questionnaire, 282
- Implicit knowledge, 133, 690
- Implicit learning, 690
- IMRD research article structure, 204
- IMRD structure, 207, 208
- Incommensurability, 45–48
- Independent observations, 505
- Independent-samples *t*-test, 532
- In-depth interviews, 86, 105
- Indexicality, 778, 781, 783
- Indian English (IE), 727
- Indigenous languages, 741–757
- Individual bilingualism, 683
- Individualism, cognitive accounts, 596
- Inductive inquiry, 87
- Inferential statistics, 61, 62
- Inflation, 137
- Informed consent, 174–176  
 significance of, 708–709
- Inner circle, 720, 723–725
- Innovations, 7, 8, 20, 21, 146, 204,  
 349, 661, 692, 714, 817, 873
- Institutional frame, interview, 226
- Institutional Review Board (IRB),  
 163–180, 180n1, 385, 560
- Instructed second language acquisition  
 (ISLA), 663–675  
 challenges and controversial issues,  
 672–673  
 core issues in, 664–671  
 limitations and future directions,  
 673–675
- Instrument, 83, 270, 276, 396, 411,  
 416, 423, 424, 426, 449, 453,  
 523, 529, 538, 547, 556,  
 563n1, 833, 839, 878, 879  
 design, 227  
 forms, 544–547  
 reliability, 547–550
- Instruments for Research into Second  
 Languages (IRIS) database,  
 333n2, 671
- Intended benefits, 858
- Interaction, 209, 226, 250, 271, 299,  
 375, 547, 596, 639, 721, 794,  
 846, 875, 887, 888  
 hypothesis, 616 (*see specific entries*)
- Interactional competence, 617
- Interaction analysis, 615–634  
 challenges, 633–634  
 core issues, 618–620  
 examples, 624–633  
 kinds of data, 620–622
- Intercultural communication, 731
- Interculturality, 239, 240
- Interdisciplinarity, 204, 264, 762, 769,  
 817, 875
- Interlanguage pragmatics (ILP), 830
- Interlanguage variation, 36
- Intermediate-level learner groups, 831
- Internal-consistency reliability, 543,  
 555–561
- Internal reliability, 543
- International Association of Forensic  
 Linguists, 714
- International Corpus of Learner English*  
 (ICLE), 365
- International Journal of Bilingual  
 Education and Bilingualism*,  
 280
- International Journal of Multilingualism*,  
 871
- International Journal of Speech,  
 Language and Law, The*  
 (journal), 714

- Internet, 270
- Internet-based research, 271
- Internet-based samples, 274
- Interpersonality, research article, 210
- Interpersonal metafunction, 651
- Inter-rater reliability, 857
- Intersectionality, 778
- Intersubjectivity, 632
- Interview, 84, 89, 91, 225–242, 252, 282, 287, 382, 534, 554, 578, 651, 723, 847, 874, 893
- challenges and controversial issues, 238–240
  - conjoint, 234
  - core issues, 229
  - focus group, 234–237
  - format, 230, 231
  - group, 234
  - historical development, 226, 227
  - language and interculturality, 239, 240
  - as naturalistic *vs.* contrived, 238, 239
  - open or open-ended interviews, 232, 233
  - rapport, 239
  - researchers, 241
  - as research instrument, 229
  - research problems/gaps, 237
  - semi-structured interviews, 233, 234 (*see specific entries*)
  - as social practice, 230
  - structured or standardized interview, 231, 232
- Intra-rater reliability, 553, 554
- Introduction-Method-Results-Discussion (IMRD) structure, 200, 201
- Introspective data, 692
- Introspective measures, 105
- Introspective verbal reports, 339–354
- Investigative forensic linguistics, 704, 706, 710
- IRB, *see* Institutional Review Board
- IRIS digital repository, 17
- Irreducibility thesis, 42, 44
- IRT, *see* Item response theory
- ISLA, *see* Instructed second language acquisition
- Item-level descriptive statistics, 430
- Item parceling, 451, 467
- Item response theory (IRT), 21–22
- Items, numbers of, 427
- J**
- Jaccard's coefficient, 711
- Japanese Learner English (JLE) Corpus, 578
- Journal of Second Language Writing* (journal), 204
- Journal of Sociolinguistics* (journal), 249
- Judgment tasks
- grammaticality judgments, 314–325
  - magnitude estimation, 325–326
  - preference judgments, 326–327
  - truth-value judgments, 327
- Just-identified model, 471
- K**
- Kaiser criterion, 436
- Kaiser-Meyer-Olkin (KMO), 433, 434, 441
- Kappa  $\kappa$ , 555
- KIAP project, 210, 211
- Knowledge, 45, 81–83, 117, 125, 126, 130, 139, 140, 210, 230, 277, 287, 289, 293, 304, 306, 381, 385, 387, 396, 397, 554, 557, 573, 588, 608, 616, 633, 634, 730, 732, 832, 845, 849, 853, 856, 858, 859, 875, 888, 889, 897, 898
- collection, 229
  - components of, 40

- Knowledge (*cont.*)  
 individual, 44  
 kinds of, 42
- Knowledge-construction process, 211
- Kolmogorov-Smirnov (KS) test, 505, 507, 536
- Kruskal-Wallis test, 535, 536
- Kurtosis, 397, 430
- L
- Labov, William, 36
- Language acquisition, bilingual, 687–692
- Language and Communicative Practices*, 42
- Language and law, *see* Forensic linguistics
- Language and Law/Linguagem e Direito* (journal), 714
- Language, as evidence, 704, 705, 707, 708
- Language assessment, 542, 838  
 methodology, 695
- Language community, defining, 170–172
- Language game, 46
- Language ideologies, 84
- Language Interaction Data Exchange System (LIDES), 693
- Language learning, 84, 264, 273, 288, 304, 306, 359, 380, 418, 545, 577, 596, 601, 616, 617, 633, 634, 731  
 career, 88, 605  
 history narratives, 605  
 memoirs, 598, 600  
 paradigms, 287, 288, 290, 296, 297, 301–303, 305, 306
- Language Learning* (journal), 7, 155
- Language Loyalty in the United States*, 742–744, 746
- Language modes and choices, 696
- Language naming, 696–697
- Language pairing, 696
- Language planning, 811, 812  
 processes of, 812
- Language policies, 418, 644, 720, 721, 816  
 arbiters, 815  
 research background, 819  
 research method, 820
- Language policy and planning (LPP)  
 research, 811–821  
 challenges and controversial Issues, 814–817  
 core issues, 812–813  
 intermediary stage, 813  
 limitations, 817–818  
 research background, 818  
 research methods, 819
- Language Problems of Developing Nations*, 813
- Language processing, 287–291, 294, 297, 301, 307
- Language production, 292–293
- Language proficiency, 227, 293, 547, 557, 845, 847, 848, 851–854, 858, 862  
 models and frameworks of, 853
- Language research methods, 287
- Language teaching, 207, 597, 599  
*Language Teaching* (journal), 8, 146, 155  
*Language Teaching Research* (journal), 125–126, 534, 601
- Language Testing* (journal), 849, 856
- Language testing and assessment, 845–862  
 challenges and controversies, 858–861  
 core concepts, 849–855  
 historical development, 847–849  
 overview of research methods, 855–858
- Language Testing Research Colloquium (LTRC), 7

- Language Testing: The Construction and Use of Foreign Language Tests*, 847
- Large-scale assessments, 556
- Large-scale language proficiency, 834
- Latency, psycholinguistic measure of, 290–292
- Late/sequential bilingual, 684
- Lauper, Cyndi, 281
- Layering, 20
- Learner beliefs, 754
- Learner corpus data, role of, 364–368
- Learning, 364, 369, 370
- Levene's test, 508
- Lexical decision task (LDT), 289, 294, 297
- Lexical entries, 362
- Lexical-level learning, 303
- Lexis-grammar, in context, 363–364
- LGBT individuals, 92
- LIDES, *see* Language Interaction Data Exchange System
- Life records, 596
- Likelihood ratio test, 528, 529, 708
- Likert scale, 272, 274, 275, 281, 320, 325, 727
- Likert-type scale data, 430
- Lime*, survey online authoring tool, 271
- Linear correlation ( $r$ ), 397
- Linearity, 427
- Linear regression model, 404, 411, 417, 579, 586, 587
- Linguistic approaches, 692–694
- Linguistic audioscape, 874
- Linguistic ethnography, 258, 615
- Linguistic Inquiry and Word Count (LIWC), 573–574, 578
- Linguistic landscape (LL), 869–880
  - challenges, 873–875
  - core issues, 870–873
  - embodiment, 876
  - materiality, 876
  - mobility, 876
  - studies, 91
  - virtual linguistic landscapes, 877
- Linguistic soundscape, 874
- Linguist List*, 280
- LISREL adjusted goodness-of-fit index, 473
- LISREL goodness-of-fit index, 473
- Literacies, 212, 260, 261, 274, 380, 541, 561, 744, 753, 812, 877, 887
- Literacy practice, concept of, 890, 891
- Literature reading, 127
- Literature review, 123–126, 188, 189
  - components of, 129
  - critical nature of, 130–131
  - data organizing, 128, 129
  - defining problem, 126
  - searching for literature, 126
  - selecting studies, 127
  - stages, 126
  - structuring, 129, 130
- LMTEST, 482
- Longitudinal analysis, 64–67
- Longitudinal research, 11, 617
- Longman Grammar of Spoken and Written English* (LGSWE), 364
- Louvain English for Academic Purposes* dictionary (LEAD), 362
- Lower-level learners, 831
- Low-intermediate learners, 831
- LTRC, *see* Language Testing Research Colloquium
- M**
- Macmillan English Dictionary for Advanced Learners* (MEDAL), 365
- Magnitude estimation, 325–326
- Magnitude-of-effect (ME) estimates, 476

- Many-facet Rasch measurement  
 (MFRM), 561  
 utility of, 554
- Mardia's coefficient, 469
- Masked priming, 297
- Massive open online classes (MOOCs),  
 575
- Matched-guise technique, 727
- Materiality, 634, 642, 876, 877, 879
- Maximum deviation (MD), 291
- ME estimates, *see* Magnitude-of-effect  
 estimates
- Means comparison analysis, 60–64
- Mean substitution, 430
- Measurement data, 525
- Mechanical Turk*, 271
- MEDAL2, 367
- Mediated Discourse Analysis, 640,  
 643–644, 649, 653, 877
- Member checks, use of, 173
- Member validation, 82
- Memoirs, 600–601
- Meta-analysis, 133, 134, 136, 137,  
 141, 449, 542, 671, 673, 675
- Metacognitive strategy, 454
- Meta-data, 383
- Metadiscourse, 209, 211, 212
- Metafunctions, semiotic resources, 645
- Metaphors for learning, 41–42
- Methodologies, 55–73, 123, 125, 127,  
 129, 137, 139, 201, 215, 241,  
 292, 368, 423, 541, 602, 606,  
 633, 721, 722, 733, 741, 746,  
 869, 870, 879  
 of EIL and ELF, 727–732  
 evolution of, 742–744  
 and methods, distinction between,  
 192–193
- Methods, research articles, 205
- Microethnography, 86, 615, 618
- Mini-language, 305–306
- Mixed effect modelling, 8
- Mixed-method approach, 282
- Mixed methodology, 103–117  
 designs, 106  
 research background, 108
- Mixed methods, 106, 226, 362, 879  
 analysis, 115  
 approach, 193  
 designs, 106, 107, 113, 114  
 limitations of, 117  
 Pros and Cons to, 116, 117  
 implementation, 115  
 planning, 114
- Mixed methods research (MMR), 12,  
 14, 103, 104, 106–107  
 challenges in, 116, 117
- Mobility, 595, 877, 879  
 and mobilizing research, 876
- Mode, definition of, 652–653
- Model-based analysis, 68–72
- Model fit indices, 486–488
- Model Summary*, 406, 408
- Moderator, 236  
 analysis, 137
- The Modern Language Journal*, 105, 250
- Monolingualism, 276
- Monte Carlo simulation method, 439
- Moral obligation of researchers, after  
 publication, 178–179
- Motivation, 56, 58, 59, 64, 200, 225,  
 396, 543, 602, 859, 889
- Mouse-tracking methodology, 291, 292
- Move, defined, 200
- Multifaceted measurement models, 561
- Multilingual interaction, 87
- Multilingualism, 92, 93, 273, 276, 284,  
 752, 815, 871, 874, 876, 878  
 research, 269
- Multimodal (Inter)action Analysis,  
 640–643, 652, 653  
 historical and psychological aspects,  
 649–650  
 objects and artefacts, 650–652  
 research background, 642  
 research method, 642

- Multimodal analysis, 639–653  
 approaches, 640  
 challenges and controversial issues, 649  
 method, 651  
 research background, 651
- Multimodal Analysis Image Software, 646
- Multimodal Conversation Analysis, 647–648, 650, 652, 653
- Multimodal data analysis, 639
- Multimodality, 87
- Multiple imputation, 430
- Multiple recording devices, 236
- Multivariate analysis of variance (MANOVA), 454
- Multivariate normality, 469, 484
- Multivariate statistical technique, 423
- N**
- Narrative, 84, 132, 228, 229, 260, 386, 424
- Narrative analysis, 597–599, 603–604  
 challenge of, 608  
 data, 604–607
- Narrative inquiry, 88, 780  
 in applied linguistics research, 595–597  
 approach to research, 597–599  
 autobiographical and biographical studies, 599, 600
- Narrative Inquiry* (journal), 596
- Narrative interview approach, 228
- Narrative knowledge, 599
- Narrative research, 597
- Narrative writing, 603  
 for publication, 607–608
- National Association of Foreign Student Advisors (NAFSA), 848
- National Council on Measurement in Education (NCME), 849
- National Foreign Language Center (NFLC), 745
- Native ethnography, 251
- Native speaker, 860
- Naturalistic and corpora data, limitations of, 693
- Naturalistic and experimental data, 692, 693
- Natural language, 304
- Natural language processing (NLP)  
 tools, role for, 567–589  
 approaches, 589  
 Biber Tagger, 570  
 Coh-Metrix, 571  
 Linguistic Inquiry and Word Count (LIWC), 573  
 L2 Syntactic Complexity Analyzer (SCA), 573  
 sample analysis, 578  
 SEANCE, 577  
 spoken and written discourse analysis, 568  
 Suite of Automatic Linguistic Analysis (SALAT), 574  
 TAACO, 575  
 TAALES, 576  
 TAASSC, 577  
 tools, 588
- Near-perfect  $R^2$  (shared variance), 428
- Near-zero  $R^2$  (shared variance), 428
- Network analysis, 382
- Neuro- and psycholinguistic approaches, to bilingualism, 685–687
- Neuroimaging techniques, 294
- New Academic Vocabulary List*, 361
- New General Service List*, 360
- NHST, *see* Null hypothesis significance testing
- Nigerian English (NE), 724
- Nominal data, 427
- Non-forensic data, 706
- Nonintrospective data, 692

Non-native speaker, 860  
 Non-norm fit index, 473  
 Nonparametric bootstrap analysis, 532, 533  
 Nonparametric hypothesis tests, 535–536  
 Nonparametric statistics, 531  
 Nonparametric tests  
   bootstrapping, 531–534  
   distribution tests, 535–536  
   nonparametric hypothesis tests, 535–536  
   permutation test, 531–534  
 Non-perfect correlations, 398  
 Non-random sampling, 272  
 Nontraditional data, 523  
 Normality, of distribution, 427  
 Normally distributed data, 505, 507  
 Norm fit index, 473  
 Novel word learning, 303  
 Novices, 200, 204–209, 214, 231, 633, 854, 856, 887, 888, 890  
 Null hypothesis, 535  
 Null hypothesis significance testing (NHST), 134, 508–510, 672

**O**

Oblique rotation, 443  
 Observation, 86, 91, 228, 249–264, 279, 281, 376, 380, 382, 387, 525, 527, 541, 554, 558, 620, 641, 648, 725, 728, 730, 731, 749, 869, 870, 878–880, 893  
 Observational self, 263  
 Observer's paradox, 688  
 Occupy Movement, 877  
 One-way ANOVA, 137  
 Online questionnaire, 269–284  
   ball rolling, 279–280  
   limitation and future directions, 282–283  
 Ontology, 81, 82

Open-ended interviews, 232, 233  
 Open interview, 232, 233  
 Opinionated words, 256  
 Options dialog box, 445  
 Oral practice pragmatic instruction, 840  
 Oral proficiency interview (OPI), 578  
 Organizational Control, 551  
 Organization, literature review, 139  
 Original test statistic, 531  
 Orthogonal rotation, 443  
 Our interview society, 226  
 Outer circle, 720, 723–725  
 Outliers based on perfect, 428  
 Outliers based on zero, 428  
 Overidentified model, 471–472  
*Oxford Learner's Dictionary of Academic English*, 372n3

**P**

Pantomimes, 795  
 Paper-based questionnaires, 270, 271, 275, 282  
*Parallel analysis*, 436, 439–440  
 Parallel forms, 544  
 Parametric statistics, 277  
 Parsimony fit indices, 473, 474  
 Participant observation, 84, 91, 251, 382, 644, 648, 725, 879  
 Participant's right to confidentiality, 172–174  
 Participation metaphor, 41, 42, 44  
 Part-of-speech (POS), 569, 574  
   tagger, 571, 575, 578  
*Pattern Grammar*, 569  
 Pearson correlation, 454, 582  
 Pearson product-moment correlation coefficient, 397–400, 545, 546, 548, 553  
 Pearson's *r*, 396–400  
   graphical representation of, 400–402  
 Pedagogical definitions, 753

- Pedagogical techniques, 290  
 Pedagogy-oriented argumentation, 201  
 Permutation test, 529, 531–534, 536  
 Personal interviews, 227  
 Personhood, 895–898  
 Perspective, literature review, 139  
 PET, *see* Positron emission tomography  
 Phi coefficient, 530  
*The Phonology of English as an International Language*, 727  
*Phrasal Verb Pedagogical List*, 362  
 Phraseology, 362, 370  
 Physical environment, 378  
 Pilot test, 281, 282  
 Placement test, 857  
*Planning Language, Planning Inequality*, 814  
 Political bilingualism, 683, 695  
 Polyglot, 683  
 Population effect, 137  
 Positionality, 778, 783  
 Positioning theory, 615  
 Positron emission tomography (PET), 686  
 Post hoc tests, 503, 510, 511  
 Powerful methods, 856  
 Power relations, 230, 894, 895  
 Practical knowledge, 125, 126  
 Pragmalinguistics, 834  
 Pragmatic competence, 853  
 Pragmatic developmental studies, 832  
 Pragmatic learning, 833  
 Preference judgments, 326–327  
 Presage factors, 889  
 Pre-service educators, 92  
 Primary research, 10, 11  
 Priming paradigms, 297–299  
 Principal axes factor (PAF) method, 427, 434, 435, 440, 441, 443  
     initial and extraction values based on, 442  
 Principal component analysis (PCA), 424, 434–439, 441, 530  
 Probability sampling, 272  
 Procedural ethics, 169  
 Process factors, 889  
 Process-product research design, 890  
 Productive' bilingual, 683  
 Product-moment, 397  
 Professional, 89, 226, 280, 376, 379, 386, 387, 563n2, 725, 763–765, 767–771, 892  
 Professionalization, 762, 766, 768  
 Professional questionnaire, 280–281  
 Proficiency scales, 853–855  
 Promax, 443  
 ProQuest database, 742, 748–751  
 Psycholinguistic methods, 287–306  
 Psycholinguistic paradigms, 297–303  
 Psycholinguistic perspective, 796  
 Psycholinguistics, 361, 576, 588  
*Publication Manual of the American Psychological Association*, 475  
 Public language, 870, 875  
 Public space, 869  
 Python script, 579
- Q**  
 QQ plots, 536  
 Qualitative, 103, 108, 205, 226, 249, 270, 381, 721, 869, 893  
     *See also specific entries*  
 Qualitative analysis, 857–858  
 Qualitative applied linguistics research, 598  
 Qualitative approaches, varieties, 85  
 Qualitative data collection, 104, 108, 109, 111, 114, 117  
 Qualitative methodological approach, 12, 871  
 Qualitative methodology, 79–93, 104  
 Qualitative methods, 79, 103  
 Qualitative-quantitative dichotomy, 105–106  
 Qualitative replication, 152–154



- Qualitative research (QR), 11–12, 79, 86, 87, 90–93, 449, 542, 555, 598, 673, 858, 859  
 epistemology, 81  
 historical development, 80  
 ontology, 81, 82  
 philosophical and methodological premises, 81  
 types of, 84  
 validity and quality of, 82–83
- Qualitative *vs.* quantitative analysis, 18
- Quantitative data, 103, 108, 133, 205, 226, 270, 381, 396, 426, 542, 568, 869  
 method, 227, 383, 395, 396, 812, 857 (*see specific entries*)
- Quantitative methodology, 55–73, 104  
 defining research and, 56–58
- Quantitative research, 11–12, 80, 104, 277, 395, 430, 542, 673  
 methods, 93, 501–518 (*see also specific entries*)
- Quartimax, 443
- Questionnaire-based research, 277
- Questionnaires, 91, 107, 269–272, 274, 275, 281, 282, 423, 425, 427, 428, 431, 454, 543, 563n1, 750, 796, 800, 857  
 Cronbach's alpha coefficient of, 431
- Questionnaires-based approach, 731–732
- R**
- Random sampling, 272
- Rapport, 239
- Rasch Item Response Theory (IRT), 455
- Rasch model, 856, 857
- Rater attitude instrument, 861
- Raters, 543, 544, 550–555, 561, 575, 847, 849, 856, 861
- Rationale, 112, 138, 183, 237, 238, 341, 364, 502
- Reaction time (RT), 290, 297, 299
- Receptive' bilingual, 683–684
- Reflexivity, 82, 263
- Regions of interest (ROI), 290, 291, 299
- Regression, 403, 425, 430, 446, 451, 501–518, 530, 531, 584, 605  
 coefficients, 411
- Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment*, 855
- Relationality thesis, 42
- Relative/incremental fit, 473  
 indices, 474
- Reliability, 114, 231, 257, 258, 299, 396, 416, 417, 434, 446, 454, 550, 554, 804, 834, 847, 853, 855–857  
 coefficient, 430  
 instrument, 547  
 types of, 543
- Repeated measures ANOVA, 67–68
- Replication, 532  
 norm, 156
- Replication research, 145–156  
 bias issues in, 154  
 current issues in, 146–147  
 design overview, 148–150  
 future directions, 155–156  
 history of, 145–146  
 qualitative, 152–154  
 reporting and data sharing issues in, 154–155  
 result interpretation of, 151–152
- Reporting verbs, 131
- Representativeness, 127
- Research, 108, 269, 287, 359, 375, 395, 423, 523, 541, 569, 719, 741, 793, 811, 829, 845, 876, 887

- data analysis, 110
  - instruments/techniques, 110
  - setting and participants, 110 (*see specific entries*)
  - type of, 110
  - Research analysis, 17
  - Research article, 430, 742, 749
    - abstracts, 201, 202
    - implications for writing, 212–215
    - introductions, 202–204
    - macro-structure of, 200, 201, 208
    - methods, 205
    - results, discussions, and conclusions, 205–208
    - rhetorical features, 208–212
    - writing, 199–215
  - Research designs/types, 84
  - Research ethics, 384
  - Research gaps, 110, 111
  - Research instrument, 229, 269, 275, 427, 541, 838, 839
    - techniques and data sources, 14–17
  - Research interview, 226
  - Research interviewing traditions, 227–229
  - Research issues, 830–835
  - Research methods, 748–750
  - Research Methods in Language Policy and Planning*, 817
  - Research priorities, 745
  - Research problem, 113
  - Research process, 12–14
  - Research proposal, 183–196
    - criteria for assessing, 194–195
    - tools for developing, 185–193
  - Research questions, 126
  - Research report writing, 138
  - Research synthesis, 124, 131, 132, 138–141
    - meta-analysis, 675
  - Respecification, of model, 475, 489–490
  - Response theory (IRT), 856
  - Revised Hierarchical Model, 686–687
  - Rhetoric, 37–40
  - RMR, *see* Root mean squared residual
  - RMSEA, *see* Root mean squared error of approximation
  - Role-plays, 832, 833, 835, 837–838
  - Root mean squared error of approximation (RMSEA), 473
  - Root mean squared residual (RMR), 473
  - Rotated factor matrix, 445, 446
    - based on Varimax method, 447–451
  - Rotation dialog box, 445
  - Rotation method, 442–446
  - Routledge Frequency Dictionaries, 372n1
  - Routledge Handbook of Language and Identity*, 778
  - “R Square,” 408
    - computation and significance testing of, 409
- S
- Saccades, 290, 299
  - Sample analysis, 578–587
  - Sample size, 137, 269, 276–279, 411, 427, 430, 454, 524, 525, 527–529, 531, 536, 557, 731
  - Sampling, 205, 236, 269, 271–273, 276, 279, 291, 360, 381, 434, 523, 531, 538, 651, 848, 873
    - strategies, 272, 273
  - Scatterplot, 400–402
  - Screen environment, 378
  - Scree plot*, 436, 438–439
  - SEANCE, 577–578
  - Secondary research, 10
  - Second/foreign language, 84
  - Second language (L2), 45, 46, 148–150, 201, 282, 290, 293, 294, 297–301, 303–306, 313–317, 326, 328, 329, 331,

- Second language (L2) (*cont.*)  
 332, 332n1, 341, 342, 346,  
 348, 350–354, 363, 364, 369,  
 370, 401, 414, 418, 419, 424,  
 525, 547, 549, 570, 572, 573,  
 577, 586, 720, 731, 732, 753,  
 793, 800, 848, 855
- ANOVA and, 501–518
- instructed second language  
 acquisition (ISLA) and,  
 663–675
- knowledge, 664, 669–671, 674, 676
- learning, 690
- pragmatics, 829–840  
 research problems/gaps, 838  
 assessment, 834–835  
 challenges, 839–840  
 data collection approach,  
 835–838  
 defining, 829, 830  
 instruction, 832–833  
 learning, 830–832  
 research methods, 838
- Syntactic Complexity Analyzer  
 (SCA), 573
- Second language acquisition (SLA), 36,  
 57, 60, 61, 72, 93, 105, 109,  
 113, 132, 164, 204, 250, 264,  
 269, 270, 273, 289, 300, 305,  
 307, 339, 343, 345–347, 349,  
 352, 353, 365, 368, 418, 501,  
 525, 542, 570, 731, 796–799,  
 805, 845
- elicitation tasks, 313–331
- instructed, 663–675
- and gesture studies  
 analyses, 801–804  
 core issues, 796–799  
 data collection, 800–801  
 languages, proficiency, and age,  
 800  
 methods, 799–801
- Second language learning (SLL), 570,  
 588
- Second Language Research Forum  
 (SLRF), 7
- Second language writing, 204
- SEM, *see* Structural equation modeling
- Semi-artificial languages, 304–305
- Semiotic potential, 646
- Semiotics, 21
- Semi-structured interviews, 233, 234
- SenticNet, 578
- Sequential bilingualism, 690
- Sequential designs, 109
- Sequential embedded design, 112
- Shape statistics*, 397, 398
- Shapiro-Wilk test, 505
- Shared variance ( $r^2$ ), 402
- Significant contribution, 190, 191
- Signs, 795
- Simple frequencies, 526–527
- Simple linear regression model,  
 403–408
- Simultaneous bilingualism, 690
- Simultaneous/early bilingual, 684
- Singapore Colloquial English (SCE),  
 727
- Skewness, 397, 430
- SLA, *see* Second language acquisition
- SLRF, *see* Second Language Research  
 Forum
- Small-scale narrative interviews, 228
- Snowball sampling, 279
- Social bilingualism, 683, 695
- Social context, 891–892  
 definition of, 889
- Social-networking sites, 277
- Social practice, 230
- Social realism, 90
- Social realities, 84
- Social semiotics, 640, 644, 646, 647,  
 649, 653, 869, 873
- Sociocultural theory, 90

- Sociolinguistic and social psychological approaches, 693–694
- Sociolinguistic competence, 853
- Sociolinguistic interview, 228
- Sociopragmatic awareness, 831
- Sociopragmatic competence, 830
- Sociopragmatics, 21
- Spanish, 287, 300, 301
- Spanish Learner Language Oral Corpora (SPLLOC), 370
- Speaking practice, 33, 34
- Spearman rank-order correlation coefficient, 546, 548, 553
- Speech gesture, 796
- Speech-linked gestures, 795
- Speech segmentation paradigm, 303–304
- Speech transcription, 802
- Sphericity, Bartlett's test, 434
- Spoken discourse, 200, 587, 878  
analysis, 568
- Spontaneous conversation, 239
- SPSS dialog box, for EFA with PAF  
extraction method, 440, 441
- Standard error of measurement (SEM), 559
- “Std. Error of the Estimate,” 411
- Standardised residuals, 485  
distribution of, 486, 493
- Standardised solution, 462
- Standardized interview, 227, 231, 232
- Standardized variables, 417
- Standard language teaching, 207
- Standard priming paradigm, 297
- Standards for Educational and Psychological Testing*, 849
- Standards for Educational Testing*, 858
- Statistical distribution analysis, 878
- Statistical Literacy Assessment for Second Language Acquisition instrument, 675
- Statistical Package for Social Sciences (SPSS), 400, 406–409, 411, 412, 415, 423, 425, 426, 428, 431, 432, 439, 440, 443, 449, 451, 528, 557, 563n3, 583
- data sheet, 451
- factor scores in, 452
- descriptive dialog box, 433
- descriptive statistics options in, 430
- extraction dialog box, 436
- parallel analysis syntax in, 440
- rotation methods in, 443
- Statistical programme, 425
- Statistical significance, 454
- Statistical techniques, 538
- Statistics, 56, 57, 60–62, 65, 67, 68, 70, 72, 137, 277, 395–397, 400, 408, 415, 419, 427, 434, 523–527, 529–531, 536, 556, 559, 561, 563n3, 587, 804
- shape, 397
- tests, 526
- Status planning, 813
- Stigma, 523
- Stimulated recall, 84, 91, 339–354  
reactivity and veridicality with, 346–347  
interviews, 114
- Storytelling, 228, 241, 595, 596, 598, 599, 603, 608
- Strategic competence, 853
- Structural equation modeling (SEM), 22, 395, 454, 461, 464–468, 856  
software for, 469  
steps in, 469–477
- Structured interview, 231–233
- Structure of Scientific Revolutions, The*, 46
- Students' language learning, 108
- Studies in Second Language Acquisition* (journal), 105, 155, 204, 418
- Study notes, 128

- Substantive analytic insights, 622  
 Suite of Automatic Linguistic Analysis  
   Tools (SALAT), 574–575  
 Survey interviews, 227  
*SurveyMonkey*, 270  
 Surveys-based approach, 731–732  
 Survey usage, 726–727  
 Symbiotic gesture, 39  
 Symbolic value, 878  
 Syntactic priming, 298, 299  
 Syntactic sophistication indices, 577  
 Synthetic notes, 128, 129  
*System* (journal), 8  
 Systemic Functional Linguistics, 644,  
   645, 653  
 Systemic Functional Multimodal  
   Discourse Analysis, 640,  
   644–646, 650, 653
- T
- TAACO, 575–576  
 TAALES, 576–577, 579, 580, 582,  
   584, 587  
 TAASSC, 577  
 Talk-in-interaction, 38, 39, 87, 832  
 Target group, 276  
 Teaching, 84  
 Teaching English to speakers of other  
   languages (TESOL), 250, 598,  
   602  
 Techno-biographic interviews, 787  
 Techno-biography, 787  
 Technological developments, 21–22  
 Technologies, 60, 64, 65, 72, 202, 263,  
   294, 375, 551, 563n2, 567,  
   568, 572, 642, 733, 743, 840,  
   849, 877  
   for L2 learning, 674  
*TESOL Quarterly* (journal), 7, 80, 166,  
   602  
 Test of English as a Foreign Language  
   (TOEFL), 396, 847
- Tests, 137, 281, 287, 396, 423, 430,  
   524, 574, 617, 756, 845,  
   889  
   Bartlett's, 433, 441  
 Test-taker, 558  
 Textuality, 209, 870–872, 887  
 Thematic analysis, 386  
 Thematic coding, 604  
 “*Theoretical Bases of Communicative  
   Approaches to Second Language  
   Teaching and Testing*,” 853  
 Theoretical frameworks, 90  
 Theoretical linguistics, 42  
 Therapeutic intervention, 241  
 Think-alouds, 84, 339–354, 554, 888,  
   889  
   reactivity with, 343–345  
   verdicity with, 345  
 Thinking for speaking, 796–798,  
   801  
 Thinking stage, of empirical research,  
   13, 14  
 Third-order factor CFA model, 462  
 Thomson, Pat, 184, 185  
 Thought-collective, 35, 36  
   epistemology and, 40–45  
   incommensurability and, 45–48  
   rhetoric and, 37–40  
 Thought-style, 31–50  
 Three-circle model, 720  
 Time-locked event, 294  
 Topic ethnography, 730  
 Total variance, 438, 442  
 Traditional literature reviews, 124–126,  
   138–141  
 Transana, 92, 264, 624  
 Transcription, 87, 252, 272, 380, 383,  
   606, 618, 621, 622, 645, 802,  
   803  
   computer programs, 624  
   conventions, 618, 621  
   methods, standardization, 623  
   online training for, 624

Transferability, 82  
 Transgender people, 277  
 Translanguaging, 633  
 Translation and interpreting,  
   761–771  
   challenges and controversial issues  
     in, 766–768  
   current and core issues in, 762–766  
   limitations and future directions,  
     769–771  
 Translingual practice, 619  
 Transmodality, 629  
 Transparent reporting, 138  
 Triangulation, 104  
 Triumph of functional, communicative  
   tests, 853  
 Truth-value judgments, 327  
*t*-tests, 60, 273, 277, 524, 532, 536,  
   804  
   as special case of ANOVA, 514  
   effect sizes for, 514–515  
 Tukey's honestly significant difference,  
   510, 511  
 Turn-taking, 87, 230, 622, 623, 629  
 Twitch.tv, 379, 380, 384  
 Two-pager document, 193  
 Two-tailed hypothesis, 532

**U**

UG, *see* Universal grammar  
 Under-identified model, 471  
 Unique variance, 426  
 Unitary system hypothesis, 689, 690  
 Unitary Trait Hypothesis, 851–852  
 Univariate normality, 427, 469  
 Universal grammar (UG), 689, 690,  
   692  
 Universal human activity, 595  
 Unpredictable complexity, 207  
 Unstandardised factor solution, 462  
 Urban language, 869  
 URL, for survey, 271

**V**

Validity, 82, 83, 114, 127, 135, 139,  
   141, 257, 258, 269, 305, 396,  
   454, 542, 545, 548, 555, 556,  
   559, 561, 833, 834, 847,  
   849–852, 855–857  
   ecological, 271  
 VARBRUL, 36  
 Varimax method, 443  
   rotated factor matrix based on,  
     447–451  
 Varimax rotation, SPSS dialog box for,  
   444  
 Verbalization, *see* Verbal reports  
 Verbal reports, 339–354  
   challenges in, 349–351  
   controversial issues in, 351–352  
   core issues in, 343  
   historical development of, 340–341  
   limitations and future directions,  
     353–354  
   origin and use in applied linguistics,  
     341–342  
   research types and, 347–349  
 Video recording, 236  
 Vignette, 260, 261  
 Virtual environment, 378  
 Virtual linguistic landscapes, 877–880  
 Visualizations, 382  
 Visual word (VW) paradigm,  
   299–301  
*Vocabulary in Use*, 360  
 Voice onset time (VOT), 290, 292  
 Vote-counting, 132  
   synthesis, 133  
 Vulnerable group, 276–277

**W**

Waikato Environment for Knowledge  
   Analysis (WEKA), 580, 584  
 Wald Test, 482  
 Warrant, of argument, 130

- Web-based research, 269  
 disadvantages of, 270
- Wilcoxon rank-sum test, 535
- Wilcoxon signed-rank test, 535
- Word learning, 304  
 paradigms, 303
- World Englishes* (journal), 861
- World Englishes (WE), 528, 529, 728,  
 730, 733, 860–861  
 future directions in, 732–733  
 methodologies, 721–727  
 streams of, 720–721
- Writing, 124, 125, 129–131, 138, 256,  
 258, 262, 359, 363, 366, 367,  
 375, 379, 543, 549, 550, 553,  
 567, 571–575, 577, 578, 597,  
 603, 630, 732, 845, 846,  
 848, 849, 851, 852, 854,  
 870, 873, 879, 888, 889,  
 893–896  
 research article, 199–215
- Written discourse, analysis, 568–570
- Y**
- YouTube, 282, 380, 381, 383
- Z**
- Zone of proximal development (ZPD),  
 799
- z* statistic, 484, 488