Betty Lanteigne
Christine Coombe
James Dean Brown *Editors*

Challenges in Language Testing Around the World

Insights for language test users

Springer

Challenges in Language Testing Around the World

Betty Lanteigne · Christine Coombe ·
James Dean Brown
Editors

Challenges in Language Testing Around the World

Insights for language test users

Springer

Editors
Betty Lanteigne
LCC International University
Klaipeda, Lithuania

James Dean Brown
University of Hawai'i at Mānoa
Honolulu, HI, USA

Christine Coombe
Higher Colleges of Technology Dubai
Dubai Men's College
Dubai, United Arab Emirates

ISBN 978-981-33-4231-6 ISBN 978-981-33-4232-3 (eBook)
https://doi.org/10.1007/978-981-33-4232-3

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

The themes of most edited collections in language assessment address traditional areas of test development, scoring models, statistical analyses, score interpretation, research, and policy. In addition, they typically provide guidelines for idealized situations with near perfect solutions. There is almost never mention of things that could get overlooked, misapplied, misinterpreted, misused, and are erroneous in test development, practice, research, and policy. This is where this collection comes in—it reveals human infallibilities and misinterpretations in the development, practice, and research in a variety of contexts from many countries. It then offers reflective chapters on how to resolve or reconcile these matters. This is an untold story in the field of language assessment.

The thirty-seven chapters in six parts in this collection are devoted to experience-based as well as data-based issues but are centered on reflecting on mistakes, issues of washback, fairness, and construct-irrelevance. Chapters reflect on failing to consider the effects of descriptive statistics in interpreting other important testing statistics like reliability estimates, correlations, criterion-related validity, and failing to account for the assumptions that underlie higher-order statistics. Other chapters reflect on important challenges related to fairness of score use, washback, and consequences as well as the use of local test development, high-stakes assessments, and placement and classroom assessments.

One of the travesties in the field has been that insights into test development, practice, research, and policy have depended largely on insights from Western English-speaking world countries. Such a perspective is limiting in vision and lacking meaningfulness and applicability to other contexts. Thus, it is refreshing for a collection to include chapters on assessing language varieties (British, Australasian, and North American), in different languages (Arabic, Hebrew, and Slovenian) from a variety of countries (Japan, Ukraine, France, Iran, Mexico as well as the USA). Furthermore, it is encouraging to have chapters that discuss assessing traditional skill areas (listening, speaking, reading, and writing) from the perspective of local high-stakes placement tests and classroom tests. Of course, the backbone of assessment in schools, colleges and universities is language teachers. It is critical therefore that they have the requisite understanding and methods to deal with routine as well as challenging high-stakes contexts. This burgeoning area, now referred to as assessment literacy, is also included in this collection with perspectives from teachers from Turkey, Sri Lanka, Bangladesh, Malta, Iran, the United Arab Emirates, and the USA.

Another delightful aspect of this collection is the inclusion of authors who are probably first-time authors or less published researchers along with well-known and famous authors. This inclusion is an intelligent choice as the former group of authors could offer a view that is new and eye-opening while the latter group might seek to consolidate their well-known positions one more time. This idea of mixing authors from different regions can also be seen in the choice of editors of this collection: Betty Lanteigne (from Europe), Christine Coombe (from the Asia/Middle-East), and James Dean Brown (from the USA). Such collaboration can surely bring to the fore multiple perspectives in choosing authors, themes, and chapters. We see an outstanding example of this in this extraordinary volume.

San Gabriel, CA, USA Antony John Kunnan
October 2020

Acknowledgments

First and foremost we would like to thank all of the contributors for sharing their observations about, experiences with and investigations of challenges in language testing in many parts of the world. We also appreciate their patience throughout the entire publication process. Gratitude is also extended to the Springer reviewers, whose insightful comments helped shape *Challenges in Language Testing Around the World*. Special thanks go to the Springer team: Lawrence Liu, Senior Editor, whose initiative was the spark that began this project; Lay Peng Ang, Senior Editorial Assistant; Sriram Srinivas, Project Coordinator; and Grace Liyan Ma, Editor, for their organized and practical assistance with producing this edited volume.

We dedicate this volume to our students and colleagues who have helped us work through assessment challenges and turned them into learning opportunities.

Contents

1 Introducing *Challenges in Language Testing Around the World* 1
Betty Lanteigne, Christine Coombe, and James Dean Brown

Part I Learning from Language Test Interpretation Problems, Negative Effects, or Misuse

2 Problems Caused by Ignoring Descriptive Statistics in Language Testing 15
James Dean Brown

3 Disregarding Data Due Diligence Versus Checking and Communicating Parametric Statistical Testing Procedure Assumptions 25
Phillip B. Rowles

4 Washback of the Reformed College English Test Band 4 (CET-4) in English Learning and Teaching in China, and Possible Solutions 35
Feifei Han

5 Fairness in College Entrance Exams in Japan and the Planned Use of External Tests in English 47
Yuko Goto Butler and Masakazu Iino

6 (Mis)Use of High-Stakes Standardized Tests for Multiple Purposes in Canada? A Call for an Evidence-Based Approach to Language Testing and Realignment of Instruction 57
Li-Shih Huang

7 Testing in ESP: Approaches and Challenges in Aviation and Maritime English 67
Neil Bullock and Carolyn Westbrook

8 A Conceptual Framework on the Power of Language Tests as Social Practice 79
Tuçe Öztürk Karataş and Zuhal Okan

9 The Washback Effect of the Vietnam Six-Levels of Foreign Language Proficiency Framework (KNLNNVN): The Case of the English Proficiency Graduation Benchmark in Vietnam 99
Phương Hoa Đinh Thị and Handoyo Puji Widodo

10 Avoiding Scoring Malpractice: Supporting Reliable Scoring of Constructed-Response Items in High-Stakes Exams 127
Kristina Leitner and Benjamin Kremmel

11 Score Changes with Repetition of Paper Version(s) of the TOEFL in an Arab Gulf State: A Natural Experiment 147
Betty Lanteigne and Hana Sulieman

Part II Learning from Tests of World Languages

12 Whose English(es) Are We Assessing and by Whom? 169
Liying Cheng, Gwan-Hyeok Im, and Rubaiyat Jabeen

13 Challenges in Developing Standardized Tests for Arabic Reading Comprehension for Secondary Education in the Netherlands 181
Anneke de Graaf

14 The Conflict and Consequences of Two Assessment Measures in Israel: Global PISA vs. the National *MEITZAV* 191
Ofra Inbar-Lourie and Elana Shohamy

15 How to Challenge Prejudice in Assessing the Productive Skills of Speakers of Closely Related Languages (the Case of Slovenia) 201
Ina Ferbežar and Petra Likar Stanovnik

Part III Learning from Program-Level Language Tests

16 EFL Placement Testing in Japan 223
James Carpenter and Sawako Matsugu

17 TEFL Test Practices at a Ukrainian University: Summative Test Design Through Teacher Collaboration 233
Olga Kvasova

18 Designing a Multilingual Large-Scale Placement Test with a Formative Perspective: A Case Study at the University of Grenoble Alpes 243
Cristiana Cervini and Monica Masperi

19 The Relationship Between English Placement Assessments and an Institution: From Challenge to Innovation for an Intensive English Program in the USA 255
Nicholas Santavicca

20 Placement Decisions in Private Language Schools in Iran 267
Kioumars Razavipour and Tahereh Firoozi

21 Perceptions of (Un)Successful PET Results at a Private University in Mexico 287
Luis Alejandro Figueroa and Krisztina Zimányi

Part IV Learning from Tests of Language Skills

22 Completing the Triangle of Reading Fluency Assessment: Accuracy, Speed, and Prosody 307
Jihye Shin

23 (Re)Creating Listening Source Texts for a High-Stakes Standardized English Test at a Vietnamese University: Abandoning the Search in Vain 317
Xuan Minh Ngo

24 The Oral Standardized English Proficiency Test: Opportunities Provided and Challenges Overcome in an Egyptian Context 327
Deena Boraie and Ramy Shabara

25 Opening the Black Box: Exploring Automated Speaking Evaluation 333
Nahal Khabbazbashi, Jing Xu, and Evelina D. Galaczi

26 Developing a Meaningful Measure of L2 Reading Comprehension for Graduate Programs at a USA Research University: The Role of Primary Stakeholders' Understanding of the Construct 345
Ahmet Dursun, Nicholas Swinehart, James McCormick, and Catherine Baumann

27 Challenging the Role of Rubrics: Perspectives from a Private University in Lebanon 357
Christine Sabieh

28 A Mixed-Methods Approach to Study the Effects of Rater Training on the Scoring Validity of Local University High-Stakes Writing Tests in Spain 383
Julia Zabala-Delgado

Part V Learning from Tests, Teachers, and Language Assessment Literacy

29 A Critical Evaluation of the Language Assessment Literacy of Turkish EFL Teachers: Suggestions for Policy Directions 411
Elçin Ölmezer-Öztürk, Gökhan Öztürk, and Belgin Aydın

30 Some Practical Consequences of Quality Issues in CEFR Translations: The Case of Arabic 421
Björn Norrbom and Jacob Zuboy

31 Assessment Literacy and Assessment Practices of Teachers of English in a South-Asian Context: Issues and Possible Washback 433
Radhika De Silva

32 English Language Testing Practices at the Secondary Level: A Case Study from Bangladesh 447
Arifa Rahman and Rubina Khan

33 A New Model for Assessing Classroom-Based English Language Proficiency in the UAE 465
Slim Khemakhem

34 Assessing Teacher Discourse in a Pre-Service Spoken English Proficiency Test in Malta 487
Odette Vassallo, Daniel Xerri, and Larissa Jonk

35 High-Stakes Test Preparation in Iran: The Interplay of Pedagogy, Test Content, and Context 497
Shahrzad Saif

36 Development of a Profile-Based Writing Scale: How Collaboration with Teachers Enhanced Assessment Practice in a Post-Admission ESL Writing Program at a USA University . . . 529
Xun Yan, Ha Ram Kim, and John Kotnarowski

Part VI Closing Thoughts

37 Reflecting on *Challenges in Language Testing Around the World* 549
Betty Lanteigne, Christine Coombe, and James Dean Brown

Editors and Contributors

About the Editors

Betty Lanteigne (Ph.D. in Linguistics, Indiana University of Pennsylvania, USA) is Associate Professor of Linguistics at LCC International University, Lithuania. She has taught ESL, EFL, EAP in the USA, Middle East, and Lithuania for 26 years, and linguistics/language assessment literacy for 14 years, supervising MATESOL theses and projects, and senior theses in linguistics. She has conducted language assessment literacy workshops for English language teachers in training in Africa, Asia, Europe, and North America. Her research interests focus on language assessment literacy, applied linguistics, and sociolinguistics, with publications about English language use, language teaching, and classroom-based language testing.

Christine Coombe (Ph.D. in Foreign/Second Language Education, The Ohio State University) is an Associate Professor of General Studies at Dubai Men's College, Higher Colleges of Technology, UAE. She has previously taught in Oman, France and the USA. Her research interests include assessment, language assessment literacy, teacher effectiveness, TBLT, research methods and teacher professionalism. Her many books, chapters and journal articles focus on these areas and others. Throughout her career, she has given presentations and trained teachers in 34 countries. Christine served as President of the TESOL International Association from 2011 to 2012 and received the James E. Alatis Award in 2018.

James Dean Brown (Ph.D. in Applied Linguistics, University of California at Los Angeles) is Emeritus Professor of Second Language Studies at the University of Hawai'i at Mānoa, USA. Before receiving his Ph.D., he taught ESL/EFL for seven years in France, the USA, and the PRC. His research focuses on language testing, curriculum development, research methods, and connected speech. He has published dozens of books and hundreds of articles on those topics (and others) and counts himself lucky to have presented at conferences, guest-lectured, and taught for 44 years in hundreds of places around the world ranging from Albuquerque to Zagreb.

Contributors

Belgin Aydın (Ph.D. in English Language Teaching) is a professor and the head of the department of English language teaching, TED University, Turkey. She has been training pre-service EFL teachers for more than thirty years. She has conducted several projects and published books and articles on English language teaching and teacher education. Her research interests are teacher education, assessment literacy and affective factors in language learning and teaching.

Catherine Baumann (Ph.D. in Second Languages and Cultures Education, University of Minnesota) is the Director of the Chicago Language Center at the University of Chicago in the USA. She has been teaching language for over 30 years, and has been directing the CLC since 2014. Her work focuses on assessment-driven pedagogy in foreign languages and language center structure and leadership.

Deena Boraie (Ph.D. in Education, University of Wolverhampton, UK), is Vice President for Student Life at the American University in Cairo, Egypt. She has three decades of experience in higher education administration, teaching and research. Her research has focused on language assessment, teacher beliefs and motivation. She has published ten articles, seven book chapters and co-edited a book entitled *Language Assessment in the Middle East and North Africa: Theory, Practice and Future Trends* (2017).

James Dean Brown (Ph.D. in Applied Linguistics, UCLA) is Emeritus Professor of Second Language Studies at the University of Hawai‘i at Mānoa, USA. His research focuses on language testing, curriculum development, research methods, and connected speech. He has published dozens of books and hundreds of articles on those topics (and others) and has been lucky enough to teach and lecture for 44 years in hundreds of places around the world ranging from Albuquerque to Zagreb.

Neil Bullock (M.A. Applied Linguistics, Birmingham University; M.A. Contemporary French studies, Northumbria University) is a specialist consultant in teaching, testing, and teacher training for English in the aviation world. Having started his professional career in aviation, he has been a language specialist for 18 years. His recent research has included specific purpose constructs and communicative skills, and he has written on teaching methodology and specific purpose operational knowledge for teachers, washback and matching constructs to real-life communication.

Yuko Goto Butler (Ph.D. in Educational Psychology, Stanford University) is Professor of Educational Linguistics in the Graduate School of Education at the University of Pennsylvania, USA. She is also the Director of the Teaching English to Speakers of Other Languages (TESOL) Program at Penn. Her research interests include language assessment and second and foreign language learning among children.

James Carpenter (M.A. in TESOL, M.Ed. in Educational Technology, Northern Arizona University) is a lecturer at Rikkyo University, Japan. He has taught EFL and ESL for 9 years. He researches learning in unique situations: currently EFL in a school for the visual impaired. He has published articles on learning through drama, project-based learning, and the assessment of inter-cultural understanding. He is pursuing a Ph.D. in Education with a concentration in Applied Linguistics at Temple University.

Cristiana Cervini (Ph.D. in Educational Linguistics, University of Macerata) is Senior Assistant Professor at the University of Bologna, Department of Interpreting and Translation, in Italy (https://www.unibo.it/sitoweb/cristiana.cervini/cv). She has been teaching applied linguistics and foreign language teaching for 15 years. She coordinated the development of the SELF placement test at the University of Grenoble Alpes. Her present studies focus on linguistics for conference interpreting, institutional communication and on assessment and evaluation.

Liying Cheng (Ph.D. in Language Testing, University of Hong Kong) is Professor in Language Education and Director of the Assessment and Evaluation Group (AEG) in the Faculty of Education, Queen’s University. Her research on washback focuses on the global impact of large-scale testing, the relationships between assessment and instruction, and the academic and professional acculturation of international and new immigrant students, workers, and professionals to Canada.

Anneke de Graaf (M.A. in Arabic, University of Utrecht), is a language tester and consultant in the field of (international) educational assessment at Cito, the Netherlands. She has worked as an exam developer for Arabic language for the Dutch national exams for the different levels in secondary education and has acted as consultant and project leader in international projects in the domain of (language) assessment for 25 years.

Radhika De Silva (Ph.D. in Language Education, University of Reading, UK) is Senior Lecturer in Language Studies at the Open University of Sri Lanka. She has taught psycholinguistics, language assessment, academic writing, language teaching methodology, and research methods, and been engaged in research supervision and thesis examination for twenty years. Her research interests include writing strategy instruction, language testing, open and distance learning, and teacher education. She has published in reputed international journals.

Phương Hoa Đinh Thị (Ph.D. in Educational Assessment and Evaluation, Vietnam National University, Hanoi; M.A. TESOL, Victoria University, Australia) is Acting Dean of the Department of Foreign Languages at Hanoi Law University. She has taught applied linguistics in the context of English studies for eighteen years and language assessment for three. Her research areas are TESOL, educational assessment, and evaluation. She has published more than twenty articles about English language teaching/learning, and washback in language testing.

Ahmet Dursun (Ph.D. candidate in Applied Linguistics and Technology, Iowa State University) is Director of the Office of Language Assessment, University of Chicago, USA. He has worked in researching, developing, and managing the language assessment programs for ten years. His work focuses on language assessment through technology, test design and development, test validation, and CALL. His publications focusing on CALL and testing research and practice have appeared in multiple peer-reviewed journals, edited volumes, and books.

Ina Ferbežar (Ph.D. in Linguistics, University of Ljubljana) is head of the Examination centre, Centre for Slovene as a second and foreign language, University of Ljubljana, Slovenia, and has worked in the area of Slovenian as L2 for twenty-five years. Her research explores interlanguage, comprehension and comprehensibility, ethical issues in language testing and language policy. She has published several articles on teaching and testing Slovenian as L2, and the book "Razumevanje in razumljivost besedil" (2012).

Luis Alejandro Figueroa (M.A. in Applied Linguistics, Universidad de Guanajuato) is a language teacher at the Tecnológico de Monterrey. Prior to his master's degree he worked as a translator and language teacher four years for a UNESCO Category 2 Centre. He has published three articles where he explores privileged settings, reliability and validity in standardized testing, and test-taking approaches.

Tahereh Firoozi (Ph.D. in Applied Linguistics, Shahid Chamran University of Ahvaz, Iran) is a Ph.D. student at the Centre for Research in Applied Measurement and Evaluation (CRAME), University of Alberta, Canada. Her research area of interest is educational assessment. She has published papers on Item Response Theory and language assessment literacy.

Evelina D. Galaczi (Ed.D. in Applied Linguistics, Columbia University) is Head of Research Strategy at Cambridge Assessment English, UK, and has worked in language education for over 25 years as a researcher, assessment specialist, teacher, teacher trainer and materials writer. Evelina's research lies in speaking assessment and the integration of digital technologies with assessment and learning. She has presented worldwide and published in academic journals, including most recently in *Language Assessment Quarterly* and *Language Testing*.

Feifei Han (Ph.D. in Education, The University of Sydney) is a Research Fellow at Griffith University, Australia. She has worked as an educational researcher for twelve years. One of her current research interests is in the area of language and literacy education. She has published more than 50 peer-reviewed articles in educational research.

Li-Shih Huang (Ph.D. in Applied Linguistics, OISE/UT) is Associate Professor of Applied Linguistics at the University of Victoria, Canada. Her interests in research and scholarly dissemination have included areas such as learner and program needs and outcomes assessment, reflective learning, and strategic behaviors in language-learning/-testing contexts. Her work has been supported by the Social Sciences and Humanities Research Council of Canada (SSHRC), the Education Testing Service (ETS), and the International English Language Testing System (IELTS).

Masakazu Iino (Ph.D., University of Pennsylvania) is Professor of Sociolinguistics at the School of International Liberal Studies and the Graduate School of International Culture and Communication Studies, Waseda University, Tokyo, Japan. His current research investigates ELF (English as a Lingua Franca), EMI (English-Medium Instruction) in higher education, and translanguaging in contact situations. His recent publication includes "Revisiting LPP (Language Policy and Planning) Frameworks from an ELF (English as a Lingua Franca) Perspective" (2020).

Gwan-Hyeok Im (Ph.D. in Education, Queen's University) is Research Professor in the Faculty of General Education at Chungwoon University, South Korea. His research interests include test validity/validation, social dimensions of testing, and Business English as a lingua franca. He has published six articles and four book chapters about linguistics, test validation models, and social aspects of language testing, including "Test of English for International Communication (TOEIC)" (2019).

Ofra Inbar-Lourie (Ph.D. in Language Education, Tel-Aviv University) is a lecturer in the Multilingual Education program in the School of Education, Tel-Aviv University, Israel. She specializes in the areas of language policy, language assessment with emphasis on language assessment literacy and teacher education, and has published widely on these topics. Her most recent publication (with M. E. Poehner) is *Towards a Reconceptualization of L2 Classroom Assessment: Praxis and Researcher- teacher Partnership* (published in 2020 by Springer).

Rubaiyat Jabeen (M.A. in TESL, Minnesota State University, Mankato, USA) is a Ph.D. student in Education at Queen's University, Kingston, Canada. She worked 10 years as an English language instructor and program coordinator in Bangladesh and the USA. Her Ph.D. research aims to explore the academic English skills important for international students' success at post-secondary education in Canada. She has two publications in the areas of education and English as an additional language.

Larissa Jonk (M.A. in English, University of Malta) is Assistant Lecturer of English Proficiency at the University of Malta, Malta and has taught English Language Proficiency for 8 years. Her research interest is English language proficiency testing for teachers, and she has published in areas of the English language teaching professional and creativity, as well as English language testing.

Nahal Khabbazbashi (D.Phil. in Education, University of Oxford) is Senior Lecturer in Language Testing at CRELLA, University of Bedfordshire, UK, and has worked in language assessment and education for 15 years. Her research interests include L2 speaking assessment, the use and impact of technology in assessment, and the effects of task and test-taker variables on performance. Nahal has published in refereed journals such as *Language Testing*, *Language and Education*, and *Linguistics and Education.*

Rubina Khan (Ph.D. in English Language Teaching, University of Warwick) is Professor of English at the University of Dhaka, Bangladesh and has taught ELT courses for twenty years. Her research explores testing, assessment and teacher development areas. She has entries on "Assessment & Evaluation" and "Assessing Large

Classes" in the *TESOL Encyclopedia of ELT* (2018) and contributions in the *An A to Z of Second Language Assessment: How Language Teachers Understand Assessment Concepts* (2018).

Slim Khemakhem (Ph.D. in Applied Linguistics, University of the West of England, Bristol) is Assistant Professor and Academic Chair of the Education Division, Higher Colleges of Technology, United Arab Emirates. He has taught different Education courses for fifteen years. His research interests cover assessment and testing, feedback, classroom interaction, advising at-risk students, and a few more. He has submitted many papers for national and international conferences, and he is getting his first chapter published in this volume.

Ha Ram Kim (Ph.D. in Linguistics, University of Illinois at Urbana-Champaign) is Lecturer and Academic Coordinator in the Program in Academic English at the University of California, Irvine, USA. She has taught courses in academic writing, reading, and oral communication for many years. Her main research interests include instructed second language acquisition, language testing and assessment, and English for Academic Purposes. Her work has appeared in major journals, including *TESOL Quarterly* and *Assessing Writing.*

John Kotnarowski (MATESL, University of Illinois at Urbana-Champaign) is Lecturer at the University of Illinois at Urbana-Champaign, USA where he has taught ESL academic writing and pronunciation classes for the last five years. Prior to working at UIUC, he was a U.S. State Department English Language Fellow in Russia and taught English at the Escuela Politécnica Nacional in Quito, Ecuador for three years. His primary professional interests are EAP writing instruction and materials development.

Benjamin Kremmel (Ph.D. in Applied Linguistics, University of Nottingham) is Head of the Language Testing Research Group Innsbruck (LTRGI), University of Innsbruck, Austria, where he teaches and researches language learning, teaching, and assessment. He has been involved in multiple language test development and research projects for 10 years. His research interests include vocabulary assessment, reading assessment, and language assessment literacy. He has published in *Language Testing*, *Language Assessment Quarterly*, *Applied Linguistics*, *Language Teaching*, and *TESOL Quarterly.*

Olga Kvasova (Ph.D. in Language Education, Kyiv National Linguistic University, Ukraine) is Associate Professor of Methods of Teaching Foreign Languages at Taras Shevchenko National University of Kyiv, Ukraine. She has worked in FL teaching for over 30 years. Her current research explores pre- and in-service teacher language assessment literacy. She has published a course book on fundamentals of language assessment for Ukrainian teacher students (2009), and numerous articles in national and international journals/peer-reviewed volumes.

Betty Lanteigne (Ph.D. in Linguistics, Indiana University of Pennsylvania, USA) is Associate Professor of Linguistics, LCC International University, Lithuania. She has taught ESL, EFL, EAP in the USA, Middle East, and Lithuania for 26 years;

linguistics/language assessment literacy for 14 years, conducting language assessment literacy workshops in Africa, Asia, Europe, and North America. Research interests and publications focus on language assessment literacy and sociolinguistics, particularly task-based language learning and testing, and real-world English language use.

Kristina Leitner (M.A. in Language Testing and Assessment, Lancaster University) is a language testing professional at the Federal Ministry of Education, Science and Research, Austria. She is involved in the development of the standardized matriculation exam for modern languages in Austria. She has worked in the area of language testing and test development for over ten years.

Petra Likar Stanovnik (degree in Slovenian Studies, University of Ljubljana) is employed at the Examination centre, Centre for Slovene as a second or foreign language, University of Ljubljana, Slovenia, and has worked as test developer, tester and tester trainer for six years. Prior to that, she worked as a teacher of Slovene as L1 and L2 and proof-reader. In her research she focuses on statistical analysis, especially on inter- and intra-rater reliability.

Monica Masperi (Ph.D. in Educational Linguistics, University Stendhal Grenoble 3) is a senior lecturer in linguistics and didactics at Université Grenoble Alpes. Her research focuses on Italian didactics, plurilingualism and the use of technology in language teaching and learning. She led numerous research projects in language training and headed the language department in Stendhal University for 12 years. She is currently the scientific director of the IDEFI Innovalangues project (ANR-11-IDFI-0024).

Sawako Matsugu (Ph.D. in Applied Linguistics, Northern Arizona University) is a Junior Associate Professor at Meiji University, Japan. She has taught EFL and ESL for about 10 years as well as coordinated various assessments in the IEP at NAU. Her research interests include language assessment, English medium instruction and project-based language learning and their assessment. She has published articles on assessment and PBLL-related issues in the IEP and EFL contexts.

James McCormick (Ph.D. in Germanic Studies and Social Thought, University of Chicago) is Director of Academic Affairs for the Humanities Division at the University of Chicago, USA and has previously worked with the University of Chicago Language Center on the design and implementation of graduate-level second-language reading assessments. In his current role, he assists departments with the recruitment and review of academic personnel.

Xuan Minh Ngo (Master of Applied Linguistics, University of Queensland, Australia) is a Ph.D. student at University of Queensland, Australia and a former TESOL lecturer at Vietnam National University, Hanoi. His research interests lie in the intersection of language assessment, language policy, and teacher education. His works have appeared in *System, English Today,* and *Asian EFL Journal* among others.

Björn Norrbom (M.A. in Applied Linguistics, Stockholm University) is a consultant in the Department of Language Testing at the Education and Training Evaluation Commission (ETEC), in Riyadh, Saudi Arabia. He has worked in language teaching and assessment for over a decade. His research interests include test validation and Arabic vocabulary.

Zuhal Okan (Ph.D. in Applied Linguistics, University of Kent at Canterbury) is working as Professor at the English Language Teaching Department at Çukurova University, Turkey, and has taught at both undergraduate and graduate levels for the last 30 years. Her research interests are teacher education, language and social justice, and educational technology, and she has recently published a book chapter on social justice and language (2019).

Elçin Ölmezer-Öztürk (Ph.D. in English Language Teaching, Anadolu University) is a faculty member at the Department of Foreign Language Education, Anadolu University, Turkey. She has been training pre-service EFL teachers for about four years, and prior to that, she taught English in various contexts for about ten years. Her research interests are assessment in foreign language teaching, language assessment literacy and second language teacher education.

Gökhan Öztürk (Ph.D. in English Language Teaching, Middle East Technical University) is an associate professor at the Department of Foreign Language Education, Anadolu University, Turkey. He taught English for several years prior to his Ph.D., and currently he has been training pre-service EFL teachers. His research focus is language teacher cognition, affective factors in language learning and teaching, and second language teacher education.

Tuçe Öztürk Karataş (Ph.D. in English Language Teaching, Çukurova University) is a Research Assistant in the English Language Teaching Department at Mersin University, and has taught at the undergraduate level. Her research explores language assessment and testing, foreign language teaching/learning and English language teacher education, and she has published a book chapter about assessment and testing for citizenship, "Language Assessment Policy within Citizenship Context: A Case of Canada and Turkey" (2020).

Arifa Rahman (Ph.D. in Languages in Education, Institute of Education, London University) is Professor of ELT and Teacher Education, at Dhaka University, Bangladesh. She has 35 years' experience in English language teaching, teacher education, materials design, assessment, testing and ELT research. Her research interests are context and culture in education, inequity in language education policy and classroom practices in low resourced settings. She has 42 publications in international and national journals and books.

Kioumars Razavipour (Ph.D. in Applied Linguistics, Shiraz University, Iran) is Assistant Professor at Shahid Chamran University of Ahvaz, Iran. He has worked in the field of English Language Teaching and Assessment for twenty years. His areas of interest include language assessment literacy, validity theory, critical language testing, and language policy. He has published on language assessment literacy and test washback.

Phillip B. Rowles (Ed.D. in Applied Linguistics, Temple University) is Junior Associate Professor of English at Tokyo University of Science, Japan, and has taught English as a Foreign Language in Japan for 27 years. His research explores Rasch model measurement methodology, language assessment, vocabulary acquisition, and enhancing assessment and measurement literacy. He has published various articles on vocabulary testing, surveying, measuring and assessment.

Christine Sabieh (Ph.D. in Psychology, Université du Saint-Esprit) is Professor at Notre Dame University, Lebanon. For over 30 years, she has taught graduate and undergraduate students and held administrative positions in the disciplines of Education, Psychology, and English Language Communication. Currently, her research interests include OER, assessment, problem-based learning (PBL), content-based instruction (CBI), blended and flipped classrooms, distance learning, ESP, and teacher training. Christine's 80 publications include book chapters, journal articles and proceedings.

Shahrzad Saif (Ph.D. in Applied Linguistics, University of Victoria) is a professor at the Departément de Langues, Linguistique et Traduction, Université Laval (Québec, Canada) where she teaches undergraduate and graduate courses in language testing and assessment. Her research focusses on the impact of high-stakes tests on teaching and learning practices, test consequences, language test development and validation, classroom assessment, language standards, and needs assessment.

Nicholas Santavicca (Ph.D. in ESL/Bilingual Education, Texas Tech University) is Associate Professor of English & Communication at the University of Massachusetts Dartmouth, USA, and has taught language education courses for more than 10 years. He has held teaching and administrative positions in both K-12 and higher educational settings in the United States and abroad. His current research reimagines assessment practices and curriculum design through the lens of diversity for multilingual students.

Ramy Shabara (Ph.D. in TEFL, Ain Shams University, Egypt) is the Test Development & Assessment Manager at the School of Continuing Education of the American University in Cairo, Egypt. He has been teaching courses on language and classroom assessment. His research interests include language assessment, research methods and, teacher beliefs and teacher education. He has published several articles on research methods and language assessment.

Jihye Shin (Ph.D. in Applied Linguistics, Northern Arizona University) is an academic consultant at Anaheim University. She has taught EAP, TESOL, and linguistics courses in the United States for the past few years. Prior to her Ph.D., she taught English in foreign language settings in South Korea. Her research explores second language reading instruction and assessment, psycholinguistics, and research methods. Her publications appear in *TESOL Quarterly* and *Applied Psycholinguistics.*

Elana Shohamy (Ph.D. in Language Education, Measurement and Evaluation, University of Minnesota) is a Professor of ML education, Tel Aviv University, Israel. Her research interests are in critical language testing, language policy, immigration, linguistic landscape and language rights. Her books on testing include *The Power of Tests* (2001), and she is the editor of two volumes of *Encyclopaedia of Language and Education*, Volume 7, *Language Testing and Assessment* (2009 and 2018, with Iair G. Or), Springer.

Hana Sulieman (Ph.D. in Applied Statistics, Queen's University, Canada) is Professor of Statistics and Associate Dean of the College of Arts and Sciences at the American University of Sharjah, UAE. She has over 20 years of teaching experience. Her research interests include sensitivity analysis, feature selection in supervised learning, statistical design of experiments. She has published numerous articles, three book chapters and co-edited a special volume of the *Proceedings in Mathematics & Statistics* by Springer, 2017.

Nicholas Swinehart (M.A. in Applied Linguistics, Ohio University) is a Multimedia Pedagogy Specialist at the University of Chicago, USA. He has worked in foreign language instructional technology for six years. His research on autonomous language learning in the digital wilds has appeared in *CALICO Journal* and *TESL Canada Journal*, and his book *Teaching Languages in Blended Synchronous Classrooms: A Practical Guide* was published by Georgetown University Press in 2020.

Odette Vassallo (Ph.D. in Applied Linguistics, University of Nottingham) is Senior Lecturer of Applied Linguistics at the University of Malta and has taught applied linguistics for ten years. Her research prioritizes language testing, teacher discourse, and academic proficiency within a cross-cultural context. Her publications focus on assessment of pre-service teachers and L2 learners' reading strategies, including "Enhancing Responses to Literary Texts with L2 learners: An Empirically Derived Pedagogical Framework" (2016).

Carolyn Westbrook (M.A. in Applied Linguistics, University of Southampton) is a Test Development Researcher at the British Council, UK. She has been involved in teaching and testing ESP for over 25 years and was formerly an Associate Professor at Southampton Solent University in the UK. Her main research areas are testing and teaching ESP and EAP, and she has published and edited 5 books and 10 journal articles in these areas.

Handoyo Puji Widodo (Ph.D. in Applied Linguistics, University of Adelaide) is Research Professor at King Abdulaziz University, Saudi Arabia. He has taught pre-service language teachers for 17 years and trained teacher educators for 5 years. Widodo's research interests include language teaching methodology, language curriculum and materials development, systemic functional linguistics (SFL) in language education, and teacher professional development. He has published more than 100 papers in edited volumes, refereed journals, and proceedings.

Daniel Xerri (Ph.D. in Education, University of York) is a lecturer at the University of Malta and has taught TESOL for seven years. His research focuses on teacher development, creativity, and practitioner research. His most recent co-edited books are *English for 21st Century Skills* (2020, Express Publishing) and *ELT Research in Action: Bringing Together Two Communities of Practice* (2020, IATEFL).

Jing Xu (Ph.D. in Applied Linguistics and Technology, Iowa State University) is Principal Research Manager at Cambridge Assessment English, UK, and has worked in language assessment for 10 years. Jing's research focuses on L2 speaking assessment, automated scoring and feedback, and validity theory. He has published in refereed journals, such as *Language Testing* and *Language Assessment Quarterly*, and recently co-edited the book, *Language Test Validation in a Digital Age* (2021).

Xun Yan (Ph.D. in Second Language Studies, Purdue University) is Assistant Professor of Linguistics at the University of Illinois at Urbana-Champaign. His research interests include speaking and writing assessment, psycholinguistic approaches to language testing, and language assessment literacy. His work has been published in a number of journals such as *Language Testing, TESOL Quarterly, Assessing Writing, System, Journal of Second Language Writing*, and *Frontiers in Psychology.*

Julia Zabala-Delgado (Ph.D. in Language and Technology, Universitat Politècnica de Valencia; M.A. in English Philology, Universitat de València; M.A. in Language Testing, Lancaster University) is a Language Advisor at the Language Centre of the Universitat Politècnica de València, coordinating standardized exams, test development and rater training. She is an expert member of the Association of Languages Centres in Higher Education in Spain (ACLES).

Krisztina Zimányi (Ph.D. in Translation and Interpreting Studies) is a full-time lecturer at University of Guanajuato, Mexico, involved in teacher and translator education for six years. Her research interests include language use in L2 classrooms, translation in L2 learning/teaching, and discourse analysis. She has published five book chapters and a dozen articles, including "Gateways into Teaching Translation in the Language Classroom" (2017).

Jacob Zuboy (M.A. in Teaching, School for International Training, USA) is a consultant in the Department of Language Testing at the Education and Training Evaluation Commission (ETEC), in Riyadh, Saudi Arabia. He has worked in a range of capacities in education in a variety of countries and contexts for the past 20 years and maintains diverse research interests.

Chapter 1
Introducing *Challenges in Language Testing Around the World*

Betty Lanteigne, Christine Coombe, and James Dean Brown

1.1 Why There Is a Need for This Book

The inspiration for this book came about through years of observing language teachers in training and practicing language teachers who were seeking to learn more about effectiveness in language assessment, particularly language testing. Sometimes training in good testing practice was not enough, because teachers continued with practices they had been using for years—practices which went against the principles of language testing they were studying. The new ideas were simply added on top of existing practices. What was needed, in addition to instruction about *what to do* in good language testing, was real-life examples of *what not to do and why*.

Being *language assessment literate*, as described by Fulcher (2012), consists of having the required knowledge and skills, understanding of principles of language assessment, and awareness of the historical and social background of language testing. Recognizing the necessity for language teachers to have language assessment literacy appropriate for their contexts, numerous organizations are seeking to meet this need. International language testing organizations are implementing initiatives to train language test users in good testing practices, particularly promoting language assessment literacy for language classroom teachers. The International Language Testing Association (www.ilta.com) has a Language Assessment Literacy Special Interest Group, and the Association of Language Testers in Europe (www.

B. Lanteigne (✉)
LCC International University, Klaipeda, Lithuania
e-mail: blanteigne@lcc.lt

C. Coombe
Higher Colleges of Technology, Dubai, UAE
e-mail: ccoombe@hct.ac.ae

J. D. Brown
University of Hawai'i at Mānoa, Honolulu, HI, USA
e-mail: brownj@hawaii.edu

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_1

alte.org) has a Teacher Training Special Interest Group that focuses on language assessment literacy. In 2013, after a language assessment literacy symposium at the Language Testing Research Colloquium, the journal *Language Testing* devoted a special edition of the journal to this topic. Also, English teaching associations make available to English teachers resources about language assessment. For example, TESOL International (www.tesol.org) offers online courses, seminars, and other resources about language assessment, and the International Association of Teachers of English as a Foreign Language (www.iatefl.org) has its Testing, Evaluation, and Assessment Special Interest Group which provides opportunities for training in language assessment literacy.

However, language teachers are not the only people who would like to, indeed who need to know more about the appropriate use of tests. Going beyond assessment literacy for teachers, Stiggins (1994) points to negative effects of decisions made by policy makers who lack assessment literacy, such as school boards. Focusing on language testing, Kremmel and Harding (2020) indicate stakeholders other than language teachers need language assessment literacy: “However, the important role of language assessment in decision-making processes across a range of domains, and the diverse nature of stakeholder groups involved in assessment processes, demands a view of LAL [language assessment literacy] that extends beyond a focus on teachers” (p. 101). Such stakeholders include national exam boards, people involved in making policy, parents of test takers, and “the greater public” (Taylor 2009, p. 25), admissions personnel (O’Loughlin 2013), and TESOL faculty (Jeong 2013). Taylor (2013) and Inbar-Lourie (2017) also indicate that stakeholders outside of language teaching have a part to play in language assessment, including making decisions about language tests and language test results, and as a consequence, they also have need of language assessment literacy appropriate to their contexts.

Thus it is evident that not all language test users (such as language program directors, testing center directors, policy makers, employers, admissions officials, as well as language teachers) are cognizant of what testing practices can negatively affect appropriate test use and/or how they can do so, including effects on students and other test takers, institutions, and organizations, as well as societies.

Areas which look at the consequences of tests are washback, consequential validity, and critical language testing. Washback (Cheng and Curtis 2004; Messick 1996; Tsagari and Cheng 2017) looks at the effects (positive or negative) of tests on teaching and learning. Messick (1989) raised the issue that consequential aspects of tests, i.e., their impact beyond the immediate test use, are part of construct validity and thus are essential to consider in designing, administering, and scoring tests, as well as interpreting test results. McNamara and Roever (2006) described societal consequences of language tests, raising issues for language test developers and test users to be aware of, issues which can have drastic effects on test takers’ lives. Shohamy (2001) identified the problem of the power of language tests, calling for test developer responsibility, aiming to decrease negative effects on test takers. Shohamy (2017) reviewed developments in critical language testing, highlighting the power wielded by high-stakes language tests which can have negative effects on vulnerable populations such as immigrants—effects of which test users need to be aware. She proposed solutions, some of which include dynamic assessment, formative assessment, and alternative assessments, and she advocated the promotion of language assessment literacy as a means of decreasing negative consequences of tests.

In a similar vein, *Challenges in Language Testing Around the World* presents critical narratives and research about language testing in different regions around the world, describing real-world language testing situations which illustrate complications of not following language testing principles or a lack of language assessment literacy. A very crucial part of increasing language assessment literacy is helping language test users be aware of potential consequences of not following language testing principles when administering, scoring, and/or using and interpreting test results. Seeing the consequences of problematic practices can actually be enlightening, highlighting the importance of adhering to validity and reliability in language testing, including being socially responsible. Brown (2010, 2012, 2014) illustrates the value of learning from mistakes, and his 2014 critical self-reflection sums up a perspective similar to that of this volume:

> In this paper, I have not only admitted my research mistakes, but in fact, I have taken pride in my ability to use these mistakes as opportunities to learn. Somewhere along the line I learned to acknowledge that mistakes are inevitable; admit that I made them; correct subsequent behavior; and, discover something entirely new from the process. (p. 277)

1.2 Audience for the Book

One important audience for this book is language test users such as those working in testing centers and language program directors, but another important audience is language teachers and students in MA TESOL programs training for those roles in the future. The goal is to provide insights from mishaps in real-world contexts so that testing practices can be improved through seeing what to avoid and why and so that more reliable and valid testing practices will therefore be encouraged, including more socially responsible decision-making.

Unfortunately, very few test users of any kind (even well-trained ones) know what testing practices can negatively affect test reliability and validity. This lack of awareness may result from the fact that such stakeholders most often think about *language testing* as involving mostly (a) large-scale national or international high-stakes tests and (b) multiple-choice items. Such narrow views of language testing lead stakeholders to consider testing as only tangential to their work rather than as integral parts of almost everything they do. Certainly, language testing researchers and people who work in test centers, as well as many administrators, are fundamentally involved with (largely multiple-choice) national and international high-stakes tests. But language teachers and those in training to be language teachers need to recognize both the importance of testing and assessment to what they do, as well as how little they typically understand about the issues involved.

As explained in Brown (2013), the only thing that distinguishes assessment from ordinary classroom activities is *feedback*. Given that assessment can take "the form of a score or other information (for example, notes in the margin, written prose reactions, oral critiques, teacher conferences) that can enlighten the students and teachers about the effectiveness of the language learning and teaching involved" (p. x), assessment covers a broad range of everyday teacher activities. Brown (2019, p. 343) further argues that "…it is important to recognize that, if you are giving feedback, you are doing assessment." Thus, assessment not only covers a broad range of teacher activities but is also central to the act of teaching. Additionally, as pointed out by Coombe, Al-Hamly, and Troudi (2009, pp. 14–15), "Research shows that the typical teacher can spend as much as a third of their professional time involved in assessment or assessment-related activities (Cheng 2001, Herman and Dorr-Bremme 1982, Stiggins and Conklin 1992). Almost all do so without the benefit of having learned the principles of sound assessment (Stiggins 2007)."

Thus it is our view that all language teachers, administrators, and test-center personnel can benefit from examining and learning from the missteps, problems, and mistakes that their colleagues around the world have encountered in real-world settings and how those colleagues have dealt with these issues.

1.3 Structure of the Chapters

There are two types of chapters in this volume: experience-based chapters and data-based chapters. Experience-based chapters are grouped together first in each part of the volume and are founded on the authors' observations of challenges in language testing, including narrations, reflections, and analyses about language testing issues. All experience-based chapters include the following sections:

Introduction: Purpose and Testing Context
Testing Problems Encountered
Insights Gained
Solution/Resolution of the Problem
Conclusion: Implications for Test Users

Data-based chapters are listed together in the second group of chapters in each part. These chapters present data collected and analyzed by the researchers, and include the following sections:

Introduction: Purpose and Testing Context
Testing Problem Encountered
Review of Literature
Methodology
Findings
Insights Gained
Conclusions: Implications for Test Users

1.4 Themes Addressed Within the Volume

Challenges in Language Testing Around the World consists of five main parts, each highlighting challenges encountered/observed/investigated by the authors, challenges from which language test users can benefit through seeing the effects of many quite diverse challenges, issues, and problems and how they were addressed, resulting in greater insight into language testing and language test use.

Part I

Part I is about learning from language test problems, negative effects, or misuse, with experience-based papers in Chapters 1–7 and data-based in Chapters 8–11.

- Chapter 2 "Problems Caused by Ignoring Descriptive Statistics in Language Testing," by James Dean Brown, reports that the mistake the researcher made was not including the effects of descriptive statistics in interpreting higher order testing statistics. The solution is to understand that descriptive statistics are closely interrelated with and affect the magnitude of all other higher order testing statistics like correlation, reliability, criterion-related validity (both concurrent and predictive), etc., and include thinking about the effects of descriptive statistics in interpreting all other statistics.
- In Chapter 3 "Disregarding Data Due Diligence Versus Checking and Communicating Parametric Statistical Testing Procedure Assumptions," Phillip B. Rowles discusses the problem that some testers and researchers forget that there are important assumptions that underlie all statistical tests. The solution is to practice data due diligence, which means making sure that all assumptions for each and every statistical test are checked and met.
- Feifei Han's Chapter 4 "Washback of the Reformed College English Test Band 4 (CET-4) in English Learning and Teaching in China, and Possible Solutions" addresses the problem of negative washback effects. The author's solutions include expanding the number of success indicators, making the spoken test compulsory, decreasing the Chinese-to-English translation section, and using integrated skills formats.
- In Chapter 5 "Fairness in College Entrance Exams in Japan and the Planned Use of External Tests in English," Yuko Goto Butler and Masakazu Iino examine fairness issues related to the use of external examinations in Japan. Their solutions include re-examining the basic premises of the policy in terms of fairness, as well as the importance of English-speaking ability in today's global world.
- Li-Shih Huang's Chapter 6 "(Mis)Use of High-Stakes Standardized Tests for Multiple Purposes in Canada? A Call for an Evidence-Based Approach to Language Testing and Realignment of Instruction" identifies the problem of the use in Canada of standardized test scores for purposes other than those for which they were designed. The author advocates transparency, consultation with the learners involved, and critical evaluation of the language testing policies in practice.

- Chapter 7 “Testing in ESP: Approaches and Challenges in Aviation and Maritime English,” by Neil Bullock and Carolyn Westbrook, concerns the disconnect in certification tests for aviation and maritime English between what is being tested and the real-world communication involved. The suggested solutions center on relying on all stakeholders in a given ESP context to describe the appropriate technical vocabulary and authentic real-world tasks used in such tests.
- Chapter 8 “A Conceptual Framework on the Power of Tests as Social Practice,” by Tuçe Öztürk Karataş and Zuhal Okan, concerns the issue that high-stakes language tests are often viewed by the public as being unrelated to social, economic, and political realities. In response, the authors examine the roles of testers, the meaning of testing to the public, the views of examinees, and the functions of tests in order to suggest a framework for critical language testing policy, practice, and literacy.
- In Chapter 9 “Washback of the Vietnam Six-Levels of Foreign Language Proficiency Framework (KNLNNVN): The Case of the English Language Proficiency Graduation Benchmark,” Phương Hoa Đinh Thị and Handoyo Puji Widodo investigate the washback effect of a national English-as-a-foreign-language test in Vietnam. The authors argue that testing policy should be based on understanding how the test is related to teachers’ knowledge and abilities, as well as what are the effects of the test on instruction and learning, including assessment, curriculum, and materials.
- Chapter 10 “Avoiding Scoring Malpractice: Supporting Reliable Scoring of Constructed-Response Items in High-Stakes Exams,” by Kristina Leitner and Benjamin Kremmel, addresses problems in scoring reliability and fairness in constructed-response items in an Austrian school-leaving exam. The authors suggest improving and refining the scoring guides through a marker support system that they describe.
- Betty Lanteigne and Hana Sulieman in Chapter 11 “Score Changes with Repetition of Paper Version(s) of the TOEFL in an Arab Gulf State: A Natural Experiment” investigate the problem of construct-irrelevant effects of taking paper TOEFL tests repeatedly. The findings suggest that the fluctuations that resulted from repeated test taking were due to factors other than change in language ability—perhaps caused by factors like fatigue, affect, test familiarity, or test-wiseness—which raises for the authors a number of questions that need to be addressed about repeated test taking.

Part II

Part II is about learning from tests of languages other than British–Australian–North American English. Chapters 12–14 in Part II are experience-based papers, and Chapter 15 is a data-based chapter.

- Chapter 12 “Whose English(es) Are We Assessing and by Whom?” by Liying Cheng, Gwan-Hyeok Im, and Rubaiyat Jabeen, addresses issues surrounding the English language construct in terms of international contexts. The authors suggest a variety of new aspects of communication that test designers and users should consider.

- Anneke de Graaf, in Chapter 13 “Challenges in Developing Standardized Tests for Arabic Reading Comprehension for Secondary Education in the Netherlands,” addresses the challenges in testing Arabic language reading comprehension within a CEFR-type framework. She shows how the Dutch language tests’ specifications and test development strategies can be adapted to meet these challenges.
- In Chapter 14 “The Conflict and Consequences of Two Assessment Measures in Israel: Global PISA vs. The National *MEITZAV*,” Ofra Inbar-Lourie and Elana Shohamy cite research that found that two standardized tests used in Israel and the massive external testing policies surrounding them lead to a great deal of negative washback. The authors conclude by suggesting that a critical assessment literate view needs to be brought to bear on large-scale national and international tests and their use.
- Chapter 15 “How to Challenge Prejudice in Assessing the Productive Skills of Speakers of Closely Related Languages (the Case of Slovenia),” by Ina Ferbežar and Petra Likar Stanovnik, describes the problem of bias in rating Slovenian language writing samples of speakers of related South Slavic languages. The authors suggest accounting for concerns about foreignness and prejudice in thinking about fairness in language testing.

Part III

Chapters in Part III are about learning from tests related to curriculum and instruction, with experience-based papers in Chapters 16–19 and data-based papers in Chapters 20 and 21.

- Chapter 16 “EFL Placement Testing in Japan,” by James Carpenter and Sawako Matsugu, addresses *Filiopietism*, which is defined by the authors as the uncritical adherence to making decisions as they have been in the past. This chapter suggests the need for considering filiopietism and its role in determining reliability, validity, and practicality in program level language tests.
- In Chapter 17 “TEFL Test Practices at a Ukrainian University: Summative Test Design Through Teacher Collaboration,” Olga Kvasova addresses the many problems faced by teachers untrained in language testing when they are required to produce a summative test (of grammar and receptive language skills, but also of speaking performance) in a Ukrainian university. She suggests conducting a series of workshops and demonstrates how to do so.
- Cristiana Cervini and Monica Masperi, in Chapter 18 “Designing a Multilingual Large-Scale Placement Test with a Formative Perspective: A Case Study at the University of Grenoble Alpes,” describe how designing the SELF test in six languages presented a number of challenges, including maintaining the same communicative construct across languages, enhancing item writers’ abilities, dealing with logistical and technical issues, and coordinating varied teams over six years. Specific strategies were used to address these issues and are described in this chapter.
- Chapter 19 “The Relationship Between English Placement Assessments and an Institution: From Challenge to Innovation for an Intensive English Program in

the USA," by Nicholas Santavicca, discusses how stakeholder relations, student language skills, assessment development, and testing innovation all present challenges for institutional placement testing. The author describes assessment design and administration principles and procedures for dealing with such challenges.

- Chapter 20 "Placement Decisions in Private Language Schools in Iran," by Kioumars Razavipour and Tahereh Firoozi, examines placement testing issues in Iran especially with regard to the content areas tested, test taker characteristics, institutional issues, test users, and issues of power. To address these issues, the authors suggest raising stakeholders' awareness of the hidden agendas involved in English language testing and teaching; enhancing the assessment literacy among all stakeholders; and establishing national and local standards for language placement testing and decision-making.
- In Chapter 21 "Perceptions of (Un)Successful PET Results at a Private University in Mexico," Luis Alejandro Figueroa and Krisztina Zimányi examine difficulties faced by the B2-C2 high school students related to institutional test score requirements. The authors suggest re-evaluating institutional policies especially in test selection to account for the standard error of measurement as well as consequential validity.

Part IV

Chapters in Part IV are about learning from tests of language skills, with Chapters 22–26 being experience-based and Chapters 27 and 28 data-based chapters.

- Chapter 22 "Completing the Triangle of Reading Fluency Assessment: Accuracy, Speed, and Prosody," by Jihye Shin, addresses a mismatch between the definition of *reading fluency* and how it is assessed. The author argues against solely using *words correct per minute* as a reading fluency measure and for inclusion of prosody rating scales—some of which are described.
- Xuan Minh Ngo, in Chapter 23 "(Re)Creating Listening Source Texts for a High-Stakes Standardized English Test at a Vietnamese University: Abandoning the Search in Vain," confronts the problems faced in finding "perfect" authentic listening source texts for a high-stakes English test at a Vietnamese university. The author addresses this problem by demonstrating how to use considerable editing and even create new listening texts that match the testing needs involved, and discusses the implications of this strategy for item-writer training.
- In Chapter 24 "The Oral Standardized English Proficiency Test: Opportunities Provided and Challenges Overcome in an Egyptian Context," Deena Boraie and Ramy Shabara describe a testing problem encountered at the American University in Cairo in one task in a standardized oral test: communication breakdown between examinees of different proficiency levels, breakdown that affected their scores. The authors describe how they solved this problem by changing the design of the task.
- Chapter 25 "Opening the Black Box: Exploring Automated Speaking Assessment," by Nahal Khabbazbashi, Jing Xu, and Evelina D. Galaczi, explores the mysteries surrounding automated scoring of spoken language. In order to make

such scoring more transparent, the authors consider the benefits, problems, and caveats related to automated speaking assessment: theoretically in terms of score interpretation and construct representation and practically in terms of the equipment necessary for recording high quality audio and difficulties related to training data.

- Ahmet Dursun, Nicholas Swinehart, James McCormick, and Catherine Baumann, in Chapter 26 "Developing a Meaningful Measure of L2 Reading Comprehension for Graduate Programs at a USA Research University: The Role of Primary Stakeholders' Understanding of the Construct," address the problems posed by testing foreign language reading ability with a translation exam. The authors describe the creation of an alternative test by the University of Chicago Language Center in a number of steps: transferring responsibility to language testing specialists; meeting with various departments to discuss the reading construct; introducing the new exam format; persuading faculty/administration of the test's validity, and other follow-up steps. They thereby provide a model that can be replicated elsewhere in similar contexts.
- Chapter 27 "Challenging the Role of Rubrics: Perspectives from a Private University in Lebanon," action research by Christine Sabieh, confronts problems encountered in planning and using rubrics. The author concludes that it is important to clearly define how rubrics function in teaching, learning, and assessment; create rubrics that reflect detailed and immediate learning outcomes; design rubrics and create tasks that encourage self-monitoring and critical thinking; and vary rubric use during different phases of the learning process to enhance student learning.
- Julia Zabala-Delgado's Chapter 28 "A Mixed-Methods Approach to Study the Effects of Rater Training on the Scoring Validity of Local University High-Stakes Writing Tests in Spain," addresses the problem of international transferability of writing sample ratings by describing a mixed-methods longitudinal study designed to deal with the effects of rater training at a Spanish university, based on raters' memory of the rating process, reliability of the scores, and scale use. The author suggests identifying expert raters for controlling the testing context.

Part V

Part V is about learning from tests, teachers, and language assessment literacy, with Chapters 29–30 being experience-based papers and Chapters 31–36 data-based.

- Chapter 29 "A Critical Evaluation of the Language Assessment Literacy of Turkish EFL Teachers: Suggestions for Policy Directions," by Elçin Ölmezer-Öztürk, Gökhan Öztürk, and Belgin Aydın, explores problems in the language assessment practices in Turkish EFL teaching, especially regarding teachers' language assessment literacy. The authors suggest possible solutions including the need for involving teachers, trainers, and policy makers in addressing the assessment literacy problem and supporting those teachers who already have assessment background knowledge.
- In Chapter 30 "Some Practical Consequences of Quality Issues in CEFR Translations: The Case of Arabic," Björn Norrbom and Jacob Zuboy address problems

with the quality (in terms of terminology, level descriptors, and style) of the CEFR Arabic translation used by the Council of Europe. The authors discuss the implications and suggest that, in the short term, users exercise care in utilizing any teaching or assessment items based on that Arabic translation and in the long term an official Council of Europe Arabic CEFR translation be produced along with supporting documentation including a multilingual glossary.

- Chapter 31 "Assessment Literacy and Assessment Practices of Teachers of English in a South-Asian Context: Issues and Possible Washback," by Radhika De Silva, is a mixed-methods study which examined issues in assessment literacy among English teachers in Sri Lanka. The author found that some teachers were fairly knowledgeable about the basic principles of assessment, but most had problems applying such principles, and suggests the need for training and support for setting, administering, and scoring tests.
- Arifa Rahman and Rubina Khan's Chapter 32 "English Language Testing Practices at the Secondary Level: A Case Study from Bangladesh," is a study which explores issues raised by testing practices in Bangladeshi English language secondary schools and their washback on examinees and the educational system as a whole. The authors suggest a need for developing assessment literacy for teachers, test developers, and scorers so that these groups, as well as test users, can identify and correct harmful practices with consequences in teaching and learning.
- In Chapter 33 "A New Model for Assessing Classroom-Based English Language Proficiency," Slim Khemakhem presents a study which examines problems raised by an IELTS band 6 graduation requirement at the end of a B.Ed. program in the UAE designed to ensure that students have the minimum English language proficiency needed to teach English in schools. The researcher suggests bridging the gap between what is and what should be by using a new assessment tool (i.e., the Classroom-Based English Language Proficiency Rubric) that merges IELTS descriptors with the principal features of classroom interaction.
- Odette Vassallo, Daniel Xerri, and Larissa Jonk, in Chapter 34 "Assessing Teacher Discourse in a Pre-Service Spoken English Proficiency Test in Malta," address problems in the spoken English proficiency of pre-service teachers. The authors describe the design and use of a spoken proficiency test that included teacher discourse as the first of five criteria and explain how assessing teacher discourse is a suitable way to address the needs of pre-service teachers of English and the effects of doing so on English teaching in Malta.
- Chapter 35 "High-Stakes Test Preparation in Iran: The Interplay of Pedagogy, Test Content, and Context," by Shahrzad Saif, is a study which explores issues related to high-stakes test preparation in Iran. The author finds that the culture of the test center shapes the test preparation courses in terms of test demands, and that instruction goes well beyond test-related activities to include contextual factors like student goals/needs, teacher experience, and second language learning beliefs.
- In Chapter 36 "Development of a Profile-Based Writing Scale: How Collaboration with Teachers Enhanced Assessment Practice in a Post-Admission ESL Writing

Program at a USA University," Xun Yan, Ha Ram Kim, and John Kotnarowski describe a study which investigates issues that arose because raters of academic English writing samples in the USA approached the rating of argument development differently, which resulted in conflicting ratings on certain essays. The author showed how several rounds of discussion led to resolving these differences and creating separate criteria for argument development and lexico-grammar, which in turn led to a scale that more accurately reflected the examinees' range of writing performances.

1.5 Entering a World of Challenges

The 35 chapters in this volume, written about language testing in 22 countries worldwide or language testing in general, have described the effects of the many different issues, challenges, and problems that the authors have experienced and/or encountered in their own educational contexts and careers. We invite readers to enter this world of challenges encountered/observed/investigated in this book so they, too, can benefit from examining the effects of these different issues and challenges. It is our hope that readers will find many new insights here into the real world of language testing and language test use.

References

Brown, J. D. (2010). Adventures in language testing: How I learned from my mistakes over 35 years. In *English language testing: Issues and prospects.* Proceedings of the 2010 Annual GETA International Conference (pp. 18–28). Gwangju, Korea: Global English Teachers' Association.

Brown, J. D. (2012). The perils of language curriculum development: Mistakes were made, problems faced, and lessons learned. In H. Pillay & M. Yeo (Eds.), *Teaching language to learners of different age groups. Anthology Series 53* (pp. 174–193). Singapore: SEAMEO Regional Language Centre.

Brown, J. D. (Ed.) (2013). *New ways of classroom assessment* (revised). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Brown, J. D. (2014). Adventures in language research: How I learned from my mistakes over 35 years. In J. Settinieri, S. Demirkaya, A. Feldmeier, N. Gültekin-Karakoç, & C. Riemer (Eds.), *Empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache: Eine Einführung* (Empirical research methods for German as a foreign and second language: An introduction) (pp. 269–279). Paderborn, Germany: Ferdinand Schöningh UTB.

Brown, J. D. (2019). Assessment feedback. *The Journal of Asia TEFL, 16*(1), 334–344. Also available at: http://dx.doi.org/10.18823/asiatefl.2019.16.1.22.334.

Cheng, L. (2001). An investigation of ESL/EFL teachers' classroom assessment practices. *Language Testing Update, 29,* 53–83.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 3–17). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Coombe, C., Al-Hamly, M., & Troudi, S. (2009). Foreign and second language teacher assessment literacy: Issues, challenges and recommendations. *Cambridge ESOL: Research Notes, 38,* 14–18.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041.

Herman, J., & Dorr-Bremme, D. (1982). *Assessing students: Teachers' routine practices and reasoning*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment.* Encyclopedia of language and education series (3rd ed.) (pp. 257–270). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-02261-1.

Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers? *Language Testing, 30*(30), 345–362. https://doi.org/10.1177/0265532213480334.

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly, 17*(1), 100–120. https://doi.org/10.1080/15434303.2019.1674855.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA: Blackwell Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: MacMillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241–256.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363–380. https://doi.org/10.1177/0265532213480336.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests.* Harlow, England: Pearson Education.

Shohamy, E. (2017). Critical language testing. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment.* Encyclopedia of language and education series (3rd ed.) (pp. 441–454). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-02261-1.

Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York, NY: Macmillan College Publishing Company.

Stiggins, R. J. (2007). Conquering the formative assessment frontier. In J. McMillan (Ed.), *Formative classroom assessment* (pp. 8–28). New York, NY: Colombia University Teachers College.

Stiggins, R. J., & Conklin, N. (1992). *In teachers' hands: Investigating the practice of classroom assessment*. Albany, NY: SUNY.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29,* 21–36. https://doi.org/10.1017/S0267190509090035.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412. https://doi.org/10.1177/0265532213480338.

Tsagari, D., & Cheng, L. (2017). Washback, impact, and consequences revised. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment.* Encyclopedia of language and education series (3rd ed.) (pp. 360–372). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-02261-1.

Part I
Learning from Language Test Interpretation Problems, Negative Effects, or Misuse

Generally, this partis about language test interpretation problems, negative effects, or misuse. Six of the chapters are experience-based Chapters (2–7). This part also contains four data-based Chapters (8–11). The following are the chapters in this part:

- The part begins with Chapter 2, “Problems Caused by Ignoring Descriptive Statistics in Language Testing,” by Brown, which reflects on mistakes the researcher made based on failing to consider the effects of descriptive statistics in interpreting other important testing statistics like reliability estimates, correlations, criterion-related validity, etc. He shows how descriptive statistics are fundamentally related to the magnitude of all the other higher-order testing statistics.
- The following Chapter (3) “Disregarding Data Due Diligence Versus Checking and Communicating Parametric Statistical Testing Procedure Assumptions,” by Rowles, addresses the problems that result when testing researchers fail to account for the assumptions that underlie all higher-order statistics. He advocates ensuring that the assumptions for all statistics be checked and discussed.
- Chapter 4, entitled “Washback of the Reformed College English Test Band 4 (CET-4) in English Learning and Teaching in China and Possible Solutions,” by Han, addresses issues surrounding negative washback effects. The author argues for increasing the number of indicators of success, requiring the spoken test, reducing the translation section, and integrating the skills formats.
- Next, the fifth Chapter, “Fairness in College Entrance Exams in Japan and the Planned Use of External Tests in English,” by Butler and Iino, addresses issues of fairness as they are related to using international examinations from outside of Japan. They suggest re-assessing this policy in terms of fairness and the importance of speaking English in a global world.
- The sixth Chapter, “(Mis)Use of High-Stakes Standardized Tests for Multiple Purposes in Canada? A Call for an Evidence-Based Approach to Language Testing and Realignment of Instruction,” by Huang, addresses the problematic use of standardized test scores in Canada for purposes they were not designed for. The author argues for the importance of transparency, consultation with stakeholders, and critical evaluation of language testing policies.

- Then, Chapter 7, “Testing in ESP: Approaches and Challenges in Aviation and Maritime English,” by Bullock and Westbrook, examines the disconnect between certification tests for aviation and maritime English and the real-world communication required in those fields. They suggest developing such tests in collaboration with appropriate stakeholders to furnish the necessary technical vocabulary and authentic real-world tasks involved.
- The eighth Chapter, “A Conceptual Framework on the Power of Language Tests as Social Practice,” by Öztürk Karataş and Okan, addresses problems associated with the public view that high-stakes language tests are often seen as unrelated to social, economic, and political issues in the real-world. They consider the roles of testers, the meaning of such tests to the public, the opinions of examinees, and the purposes of tests, and then suggest a critical framework for language testing policy, practice, and literacy.
- The next Chapter (9) “Washback of the Vietnam Six-Levels of Foreign Language Proficiency Framework (KNLNNVN): The Case of the English Language Proficiency Graduation Benchmark,” by Đinh Thị and Widodo, studies the washback effect of a national test of EFL in Vietnam. They argue for changing the testing policy to one that appreciates the relationship between the knowledge and abilities of teachers and the effects of the test on learning and instruction (including curriculum, materials, and classroom level assessment).
- The tenth Chapter, “Avoiding Scoring Malpractice: Supporting Reliable Scoring of Constructed-Response Items in High-Stakes Exams,” by Leitner and Kremmel, investigates score reliability and fairness issues in the constructed-response items in an Austrian school-leaving exam. They advocate improving and refining the scoring guides by using a marker support system.
- The next Chapter (11), “Score Changes with Repetition of Paper Version(s) of the TOEFL in an Arab Gulf State: A Natural Experiment,” by Lanteigne and Sulieman, studies the construct-irrelevant effects of repeatedly taking paper-based TOEFL tests. They suggest that the resulting fluctuations (from repeated test taking) may be due to variables other than change in language ability (perhaps issues like fatigue, affect, test familiarity, or test-wiseness) and that a number of questions need to be explored with regard to repeated test taking.

All of the chapters in this part are connected to the section theme (and title) and with each other in that they deal with problems that arise in test use policy with regard to interpreting scores (2–4, 7, & 11), negative effects of tests (4 & 9), or misuse of tests (5, 6, 8, 10, & 11). While the chapters involve a wide variety of countries (an Arab Gulf State, Canada, China, Japan, & Vietnam), all chapters are similar to each other in that they are about high-stakes English language tests, and most of those are large-scale standardized tests (4–6, 8, 9, & 11).

Chapter 2
Problems Caused by Ignoring Descriptive Statistics in Language Testing

James Dean Brown

Abstract In 1980, I published a study on the relative merits of four cloze scoring methods [exact-answer (EX), acceptable-answer (AC), clozentropy (CLZNT), and multiple-choice (MC) scoring] analyzed in terms of item analysis, reliability, and validity statistics. My interpretation of the results was that AC was the best overall scoring method because AC produced the best item facility, item discrimination, and reliability estimates, and was tied with CLZNT in validity coefficients. I later realized that I made two important errors:

- I did not include my descriptive statistics in thinking about and interpreting the other testing statistics in my study.
- I forgot that all testing statistics are for scores based on performances of a certain group of examinees on one set of items under a particular set of conditions.

My solutions to these problems were based on learning from my mistakes: (1) reporting *and examining* the descriptive statistics (especially in relationship to any more advanced statistics) in all of my subsequent statistical studies and (2) stressing this important set of relationships to all of my students who have used statistics in their studies. Readers can learn from my explanations of these mistakes and from remembering in their own research and in reading published research that investigators should *include the descriptive statistics in thinking about and interpreting all other testing statistics* and *remember that testing statistics are only for scores based on performances of a certain group of examinees on one set of items under a particular set of conditions, period*.

2.1 Introduction: Purpose and Testing Context

In 1980 I published a study in the modern language journal (Brown 1980) based on my MA thesis (Brown 1978), which investigated differences in item analysis statistics (item facility and discrimination), reliability, validity, and usability for four methods

J. D. Brown (✉)
University of Hawai'i at Mānoa, Honolulu, HI, USA
e-mail: brownj@hawaii.edu

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_2

of scoring cloze tests [these statistical terms will be defined below]: exact-answer scoring (EX), acceptable-answer scoring (AC), clozentropy scoring (CLZNT), and multiple-choice scoring (MC). The cloze test was based on a passage from an intermediate ESL reader. The passage was long enough to contain a total of 50 blanks with an every-seventh word deletion pattern including two unmutilated (i.e., no blanks) sentences at the beginning and one at the end. Two versions of this cloze passage were created: (1) an open-ended version with numbered blanks and (2) a multiple-choice version with four options for each blank. The two main groups of participants in this study were ESL students who were taking the ESL Placement Examination (ESLPE) at UCLA in 1978. The cloze tests were administered at the end of the ESLPE to examinees who were randomly assigned to take the open-ended version ($n = 55$) or multiple-choice version ($n = 57$). The analyses compared the four scoring methods in terms of two item analysis statistics, reliability, validity, and usability.

2.1.1 Descriptive and Item Analysis Statistics

All during my four years at UCLA, I regularly took statistics and testing courses in the School of Education along with my Applied Linguistics courses, and I remember my statistics teachers repeatedly talking about the importance of reporting descriptive statistics in all statistics studies regardless of what other more advanced statistics were being reported, because they describe the distributions of numbers and those distributions may affect the other statistics in a study. Hence, in Brown (1980), I dutifully reported the descriptive statistics for the four scoring methods as shown in the top four rows of numbers in Table 2.1, including the sample size, mean (in this case the same as the arithmetic average), range (the distance from the lowest to highest score), and standard deviation (a sort of average of the distances of all scores from the mean). Unfortunately, I just reported these descriptive statistics without

Table 2.1 Descriptive and item statistics (Adapted from Brown, 1980, pp. 314-315 Tables 3 & 7)

Statistic	Exact-answer Scoring	Acceptable-answer scoring	Clozentropy scoring	Multiple-choice scoring
Sample size (*N*)	55	55	55	57
Mean (*M*)	15.00	25.58	33.40	31.84
Range	0-33	0-46	0-65	14-48
Standard Deviation (*SD*)	8.56	12.45	16.78	8.99
Mean Item Facility (*IF*)	.30	.51	NA	.64
Mean Item Discrimination (*ID*)	.44	.61	NA	.42

Table 2.2 Reliability and validity statistics (Adapted from Brown, 1980, p. 314 Tables 4 & 5)

Statistic	Exact-answer Scoring	Acceptable-answer scoring	Clozentropy scoring	Multiple-choice scoring
Reliability: K-R20	.90	.95	NA	.89
Reliability: Split-half adjusted	.90	.94	.93	.90
Correlation (*r*) with ESLPE	.88	.90	.91	.89
r^2 with ESLPE	.77	.81	.83	.79

interpreting them or giving them further thought in terms of how those distributions might affect the other statistics in the study.

The last two rows of Table 2.1 show mean item analysis values for three of the four scoring methods: the mean item facility (IF) (i.e., the average proportion of students who answered correctly) and item discrimination (ID) (i.e., the average degree to which items separated the highest and lowest thirds of students on the whole test). Since CLZNT scoring was weighted instead of right/wrong, I did not know at that time how to calculate either IF or ID for CLZNT scoring. My interpretation for IF was that AC scoring was the best centered and best at spreading examinees out, that is, with a mean IF of .51, I knew that 51% of the students had answered the items correctly on average, and, with the highest item discrimination value of .61 (higher than either .44. or .42 for EX and MC scoring), I knew that AC scoring was discriminating best (between the high scoring and low scoring examinees). I therefore concluded that the AC scoring had the best item statistics.

2.1.2 Reliability and Validity Statistics

The reliability statistics (i.e., the proportion of reliable or consistent variation in the test scores, ranging from .00 to 1.00) for the four scoring methods are shown (where applicable) in the first two rows of numbers in Table 2.2 including Kuder-Richardson formula 20 (K-R20) and split-half (adjusted for full-test) reliability. I interpreted these values as indicating that AC scoring was yet again the best scoring method because the reliability values (.95 and .94) were higher than any of the others.

In the third row, correlation coefficients (which indicate the degree to which any two sets of numbers go together) are reported for the four scoring methods in the last two rows of Table 2.2; these coefficients represent criterion-related validity coefficients, which indicate the degree to which the scores derived from each of the four scoring methods were correlated with the sum of the other ESLPE scores for these students, or the degree to which they were correlated with the ESLPE criterion measure. The correlation squared (r^2) values are a bit easier to understand than the correlation (r) because r^2 represents the proportion of overlapping variation between the cloze test scores in each case and the total scores on the ESLPE criterion measure. For example, the correlation of the AC scores with the total ESLPE scores was .90, and the squared value of .90 is .81, which indicates that 81 percent of the variance on the cloze measure was shared, that is, overlapped with the ESLPE scores. Taken together, I interpreted these criterion-related validity coefficients as indicating that CLZNT scoring was most valid because it correlated higher than the other scoring methods with the total ESLPE scores, but I also noted that AC scoring was second best.

Overall, my interpretation of the results of this study was that "the best overall scoring method is the AC methods" (p. 316) because AC scoring produced the best item facility and discrimination and reliability estimates, and was tied for best with CLZNT in terms of validity coefficients (note that I tested the significance of the differences between the three EX, AC, and CLZNT validity coefficients (of .88, .90, & .91, respectively) and found no significant difference at $p < .05$, meaning that the observed differences were probably due to chance alone).

When I submitted this article to the *Modern Language Journal*, the reviewer feedback was very positive with minor suggestions for improving the article, including a special request from the journal editor that I add some pedagogical implications, which I dutifully did. So the article was then published. I should also point out that I had received the very best advice available at UCLA on this project from three well-known faculty members in the TESL section and one very famous measurement specialist in the School of Education.

2.2 Testing Problems Encountered

When re-examining my 1980 study with hindsight after getting a bit more experience, I realized that I had made two errors. As a result, I learned from those errors, and I think other language testers and test users can learn from them, too:

- While I did report the descriptive statistics for the four scoring methods, I did not include them in thinking about and interpreting the other testing statistics in my study.
- Because I was viewing the descriptive, item statistics, reliability, and validity statistics as characteristics of the tests (and scoring methods) themselves, I forgot that all testing statistics are for scores based on performances of a certain group of examinees on a certain set of items under a certain set of conditions.

2.2.1 Not Including Descriptive Statistics in Thinking About and Interpreting Other Testing Statistics

Unfortunately at the time of the project that ended in Brown (1980), I did not understand why descriptive statistics were so important. That is, I did not really understand how and why interpreting all other higher order statistics in a study depend on the descriptive statistics. There are two factors that can affect the statistical results of any study: lack of normality and restrictions of range.

Most testing statistics assume *normality* which is important because, to function properly, testing statistics require that the underlying data be normally distributed. Indeed, if the distributions are not normal, that fact alone can dramatically affect the results of any testing project. Only by carefully examining and thinking about the distributions (i.e., the mean, mode, median, range, standard deviation, etc.) can testers understand what the distributions represent. Then, if a distribution is normal, the mean, mode, and median should all be about the same, and there should be enough room to fit at least two standard deviations above the mean and at least two below. For example, if a set of test scores has low and high scores of 3 and 50, and the mean, mode, and median are 25.13, 25, and 24.50, respectively, those are one set of indications of normality because, in a normal distribution they would be very similar, as is the case in this example. But it would also be important to look at the standard deviation. If the standard deviation turned out to be 10.11 in the above example, there would be room for two standard deviations below the mean (i.e., 2 x $SD = 2$ x $10.11 = 20.22$, and $M - 20.22 = 25.13 - 20.22 = 4.91$) and two standard deviations above the mean (i.e., $M + SD = M + 20.22 = 25.13 + 20.22 = 45.35$), and the low value of 4.91 and the high value of 45.35 both fit well within the range of scores from 3 to 50. Thus, this particular distribution appears to be approximately normal. If, on the other hand, the mean, mode, and median are 38.13, 30.25, and 24.50, and the standard deviation is 12.75, the distribution in all likelihood is not normal because the mean, mode, and median are so different from each other and because there is only room for one standard deviation above the mean (i.e., the mean plus *one* standard deviation $= 37.13 + 12.75 = 49.88$, which is almost as high as the highest score of 50; hence there is no room for a second standard deviation above the mean). There is much more that could be said about interpreting the normality of sets of scores, but let this suffice for the moment (for more see Brown 2005, pp. 89–113).

Another factor that can affect the functioning of testing statistics is called *restriction of range*. By and large, reliability estimates, and the various statistics used to study the degree to which a test is valid (e.g., correlation coefficients, factor analysis, analysis of variance, etc.) all work best (or are most likely to indicate that test scores are relatively valid) when the distribution of scores is not only normal but also showing a wide range of values (as indicated by a wide range of scores and relatively high standard deviation). Thus, if a sample of examinees has a narrow range of abilities, reliability and validity will tend to be low. For example, if the examinees all come from a single level of study where the range of abilities is narrow, the reliability estimates will tend to be low and validity statistics depressed as well. If in contrast a sample of examinees has a wide range of abilities, reliability and validity will tend to be high. For example, if the examinees come from all levels of study at a particular institution where the range of abilities is wide, the reliability estimates will tend to be high and validity statistics will have every chance of indicating relatively high validity for the scores involved.

2.2.2 *Forgetting that All Testing Statistics Are for Scores Based on Performances of a Certain Group of Examinees Under Certain Conditions*

In thinking about the problem discussed in the previous subsection, I noticed another major error that I was making as a young language tester/researcher: I was forgetting that all testing statistics are for the scores that represent the performances of a certain group of examines on a certain set of items under a certain set of conditions. This means that the statistics might have been very different, indeed, if any one of three things had been changed: a different set of items, or a different group of examinees, or a different set of conditions. If for example, two *different sets of items* were used for the same group under the same conditions, it would be unreasonable to expect the scores from the two sets of items to be the same on average. Indeed, if one set of items was known to be more difficult than the other, that set would quite reasonably produce lower scores. Similarly, if one set of items was administered to two groups of examinees with the first having much higher ability levels, it would be reasonable to expect the scores produced by the first group to be higher than those for the second group. Changing testing conditions can have similar effects even when the sets of items and group of examinees are held constant. Thus for all of these examples of different items, different groups, and changing conditions, the best we can say about the testing statistics that will result from any of these sets of scores is that they represent the distributions, item statistics, reliability, and validity of a certain set of items when administered to a certain group of examinees under a certain set of conditions. If we change any of these three variables, we can expect the statistics to change. In fact, the testing statistics can be expected to change quite a bit if big changes are made in items, examinees, or conditions.

2.3 Insights Gained

So, how do the problems explained in the previous main section apply to the Brown (1980) study? Recall that I interpreted the results of that study as indicating that the acceptable-answer scoring method was superior (in item statistics, reliability, and validity) to the other methods of scoring cloze tests. That interpretation takes the view that testing statistics are characteristics of the different tests produced by different scoring methods and fails to consider how those results might be affected by changes in the distributions of scores. That is, I interpreted my results without thinking about the relative normality and ranges of ability produced in the scores of the four scoring methods. Worse yet, I did so without considering what would have happened if the cloze test items had been easier or more difficult.

For example, the Fry readability analysis (see Fry 1977) indicated that the passage I used in the 1980 study was at the 7.8 readability level (i.e., it was suitable for seventh grade native-speaker students in their eighth month of school). Consider what would happen if I administered my original cloze test along with two others: one based on a passage of much lower difficulty (say suitable for 6th grade students) and another passage of higher difficulty (say suitable for 10th grade students on the Fry scale).

If my cloze test had been easier at the Fry 6th level that would probably result in distributions for all four scoring methods that were considerably higher. Let's say that all of the means were to go up by 10 points. Looking back at Table 2.1, consider AC scoring which was the best centered distribution and had a high standard deviation. With ten points added to everyone's score because of an easier passage, the AC mean would be 35 (and IF values would be .70), which is considerably less centered on the 0-50 range of possible scores. The AC standard deviation (and ID values) would also probably be lower, or even if it stayed at the same at 12.45, there would not be room for two standard deviations between the mean of 35 and the top possible score of 50. Thus the AC distribution would have been asymmetrical with scores scrunched up toward the top end. Under such circumstances, the reliability estimates and validity coefficients would likely have been lower as well.

The same problems would be even more true for the clozentropy and multiple-choice scores, which, with scores 10 points higher, they would have means (and IF values) much closer to the top of the distribution, and therefore much lower standard deviations (and ID values), or at least scores that are scrunched up near the top of their respective score ranges. This would have the effect of lowering both the reliability estimates and validity coefficients even more.

So while the distributions for AC, CLZNT, and MC would all be moved up above the middle of the range, and have lower standard deviations or skewed distributions, with scores 10 points higher, the EX scores would probably have been nicely centered around a mean of 25 out of 50 and would therefore likely have produced a higher standard deviation with correspondingly higher reliability and validity coefficients. The EX scores would therefore appear to produce more reliable and valid scores than the AC, CLZNT, or MC scores. In other words, the results and interpretations would have been completely different if I had based the study on an easier passage. All of the same would be equally true, though in the opposite direction, for a study based on a more difficult passage where the means all dropped by say 10 points, in which case CLZNT or MC scoring would probably be best centered with lower means and item IF, and produce the highest ID and standard deviation, as well as reliability and validity coefficients.

Thus the testing statistics results and interpretations in Brown (1980) can be said to be determined to a large degree by my initial choice of a passage and that passage's difficulty, and by extension, the results would probably have been quite different if the passage difficulty had been higher or lower. How could I have been so stupid?

2.4 Solution/Resolution of the Problem

Well, it would only be stupid if I had learned nothing from the experience. And, I can assure you that I learned to report *and examine* the descriptive statistics very carefully in *every* statistical study that I did thereafter, thinking all the while about the relationships among all of the statistics in each study, especially those relationships between all of the advanced statistics and the distributions depicted by the descriptive statistics (as discussed in Brown, 1988, 2013). I have also stressed this important set of relationships and interdependencies to each and every student I've had who was using statistics in a study. In short, the lesson I learned was that understanding the descriptive statistics in a particular study is crucial to the interpretation of all higher-level statistics. One solution to the problems discussed above is that I have used what I learned in Brown (1980) to improve my own studies and insisted that my students do so, too.

Specifically, in terms of cloze testing, I have included the distributions in my thinking, especially with regard to differences in the examinees' ranges of ability, and differences in passage difficulties in my thinking. For example, I explicitly investigated the effects of differences in ranges of ability on cloze test performance in Brown (1984), and we considered the effects of differences in passage difficulty by systematically equating the groups of examinees through random selection and systematically selecting five passages at *different* levels of difficulty in Brown, Yamashiro, and Ogane (2001), or randomly selecting both the examinees who took each cloze test and the conditions under which they were administered, and the 50 levels of passage difficulty (by randomly selecting passages from a US public library) (see, e.g., Brown 1993; Brown et al. 2016; Trace et al. 2017). So in my subsequent cloze studies, my solution to these problems has been to account for the fact that testing statistics are for scores based on performances of a certain group of examinees on a certain set of items under a certain set of conditions. In short, I dealt with the problems discussed in this chapter by specifically including the following as independent or control variables in my cloze studies: sets of items, groups of examinees, and testing conditions.

2.5 Conclusion: Implications for Test Users

In any language test development project or testing research project that readers may be involved in, or even in reading about or using tests for decision making, it is crucial to remember that reporting the descriptive statistics for all tests is important, but more importantly, *including the descriptive statistics (and the distributions they* *represent) in the interpretation of all other testing statistics is crucial.* In addition, instead of viewing all testing statistics as characteristics of whatever language test is involved, it is crucial that readers also remember that *all testing statistics are for scores based on performances of a certain group of examinees on a certain set of items under a certain set of conditions, period.*

In addition, language teachers or Ministry of Education officials who are developing and analyzing any end-of-year tests would be well advised to calculate and pay careful attention to the descriptive statistics for their tests and consider those statistics in interpreting their test scores and any other more elaborate reliability and validity statistics that are involved. It is equally true that all language teachers and their students would benefit from teachers calculating the relatively simple descriptive statistics discussed in this chapter for their classroom language tests. Granted the size of samples in most classes will tend to be small and the distributions of test scores might not be normal (making reliability and validity statistics hard to interpret). Nonetheless, calculation and commonsense analysis of descriptive statistics can lead teachers to interesting insights about their classroom tests with important pedagogical implications.

Following the suggestions in this paper as summarized in the previous paragraph should help language teachers and administrators to understand how all statistics tend to fit together, and to examine the complete statistical picture, as well as to develop better language tests, to do better language testing research, and to make better decisions with their language tests.

References

Brown, J.D. (1978). *Correlational study of four methods for scoring cloze tests.* MA thesis, University of California, Los Angeles.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal, 64*(3), 311–317.

Brown, J.D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R. Orem, & B. Taylor (Eds.), *On TESOL '83: The question of control.* Selected papers from the 17th Annual TESOL Convention, Toronto (pp. 109–119). Washington, DC: TESOL (also available from ERIC: ED275145).

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* Cambridge: Cambridge University.

Brown, J. D. (1993). What are the characteristics of *natural* cloze tests? *Language Testing, 10*(2), 93–116.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New ed.). New York: McGraw-Hill.

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what? *International Journal of Language Studies, 7*(1), 1–32.

Brown, J. D., Trace, J., Janssen, G., & Kozhevnikova, L. (2016). How well do cloze items work and why? In C. Gitsaki & C. Coombe (Eds.), *Current trends in language evaluation, assessment and testing: Research perspectives* (pp. 2–39). Newcastle Upon Tyne, UK: Cambridge Scholars.

Brown, J.D., Yamashiro, A.D., Ogane, E. (2001). The emperor's new cloze: Strategies for revising cloze tests. In T. Hudson & J.D. Brown (Eds.), *A focus on language test development: Expanding* *the language proficiency construct across a variety of tests* (pp. 143–161). Honolulu, HI: University of Hawai‘i Press.

Fry, E. (1977). *Graph for estimating readability–extended*. New Brunswick, NJ: Rutgers University Reading Center.

Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across learner backgrounds. *Language Testing, 33*(1), 151–174.

Chapter 3
Disregarding Data Due Diligence Versus Checking and Communicating Parametric Statistical Testing Procedure Assumptions

Phillip B. Rowles

Abstract A key principle in language testing best practices is data quality control. Unrefined data input can negatively impact so-called refined statistical output. The data quality control construct is characterized with an implicit warning that garbage in generates garbage out. Data due diligence is defined in this chapter as when a testing researcher initiates a careful investigation of the data's fundamental assumptions. This essential first step should occur prior to conducting parametric statistical testing procedures. Unfortunately, disregarding this precautionary check is a common yet fatal mistake in the current language testing environment. The hidden problem is that language test users are often unaware that inferences interpreted from unchecked statistical test results may be inaccurate. This is typically due to the user's lack of access to the researcher's documented evidence. The rationale behind not documenting the evidence is that it was never checked and therefore simply does not exist. A twofold solution to this problem is presented at both the local and global levels. At the local level, the ethical and moral solution for individual language testing researchers is to enact a maxim: Complete all preparatory data inspections beforehand and then clearly communicate that evidence to the audience. At the global level, language testing users need to create training and educational programs to raise awareness of these issues among current and future generations of language test users and researchers. A paradigm shift is long overdue to turn the tide from disregarding data due diligence into checking and communicating parametric statistical testing procedure assumptions.

3.1 Introduction: Purpose and Testing Context

Best practices in language testing design ideally involve the processing of numerical data. A commonly used collection of language testing methodologies is parametric statistical testing procedures. One prerequisite of these procedures requires stringent

P. B. Rowles (✉)
Tokyo University of Science, Tokyo, Japan
e-mail: rowles@rs.tus.ac.jp

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_3

checking of data quality control prior to data processing. If the data is not of high quality, the subsequent inferences derived from statistical results may be inaccurate. As research inferential validity is compromised, test user belief and trust could be broken. Furthermore, the consequences of this could negatively influence much broader issues involving test ethics and justice.

In this chapter, as there is a focus on data and more specifically data quality control, the word *data* is proposed to be conceptualized by the acronym DATA. Further dividing this acronym into two parts, this ordered sequence is represented as DA-TA, where DA refers to Due diligence of Assumptions and TA refers to Test Analyses. Therefore, even though the longer-range goal is to conduct statistical test analyses, individuals must initially concentrate on the shorter-range task of checking data due diligence of assumptions as a required first step.

In language testing research, as in life, finding solutions to problems may be one of our greatest teachers. This chapter addresses a fundamental problem and solution sequence faced by test users. The problem is that test users are more likely to arrive at a wrong conclusion about an issue when they falsely believe in, or just simply disregard, the underlying assumptions. A solution is for test reporters to initially take the time and care to diligently check and then later communicate these fundamental, although usually hidden, assumptions. This solution can also be described as doing research *data due diligence*. Investing the time, effort and resources to examine research data quality a priori is beneficial to the subsequent testing analyses and inferential quality.

One can also conceptualize this problem and solution in terms of statistical literacy assets and liabilities: Checking and reporting statistical assumptions is an asset, while not checking and reporting statistical assumptions is a liability. As the name suggests, making valid *inferences* is one of the important reasons for conducting *inferential* statistical tests. To achieve this end, the focus is on developing knowledge of one often overlooked part of language test user statistical literacy.

3.2 Testing Problem Encountered

Reporting the checking of statistical assumptions is a rare commodity in published language testing literature. In a second language acquisition (SLA) interaction meta-analysis study, Plonsky and Gass (2011) indicated that a mere 3% (5 of 174 studies) reported checking statistical assumptions. More specifically, these few reported statistical assumptions were just in "five studies, three by the same author, McDonough" (Plonsky and Gass 2011, p. 340). In another SLA research quality meta-analysis study, Plonsky (2013) reported that only 17% (101 of 606 studies) reported checking statistical assumptions.

This raises the question of why language testing researchers (a subset of SLA researchers) persist in neglecting their duties of reporting the checking of statistical assumptions. The answer may lie in the limited statistical literacy background of language test researchers (a subset of SLA researchers). In a statistical literacy survey study (Loewen et al. 2013) of 331 applied linguistics and SLA academic respondents (which includes the subset of language testers), 81% of respondents had completed at least one statistics course, with a median number of 2 statistics courses completed by both doctoral candidates and university professors. After examining statistical attitudes with a factor analysis, two factors were revealed, "statistics are important" and a "lack of statistical confidence." Only 13% of doctoral candidates and 30% of university professors felt their statistical training was adequate. Furthermore, 40% of doctoral candidates and 30% of university professors felt their statistical training was not adequate (Loewen et al. 2013). These attitudes might influence the lack of reporting about statistical assumptions.

3.3 Solution of the Problem

Instead of having only 3% (Plonsky and Gass 2011) or 17% (Plonsky 2013) of studies (although in different samples of second language research journals) reporting the checking of statistical assumptions, this reporting percentage should ideally be much higher. However, a change to a new consensus will take time. This time factor is compounded by the fact that academic journal editors in our field do not have specific submission guidelines for potential authors addressing this assumption reporting issue. A consensus needs to be reached on domain-wide editorial guideline revisions, otherwise, change will be slow and gradual. By reaching a threshold point, a bandwagon effect of rapidly changing popular opinion over a relatively shorter time could occur. This has been illustrated throughout human history with mass cognitive belief reversals, for example, from conceptualizing the model of the universe as geocentric to heliocentric and perceiving cigarette smoking as harmless to harmful.

By doing data due diligence, test reporters reveal hidden truths. This process deserves an awareness raising paradigm shift, that is, a scientific revolution (Kuhn 1996). In this chapter, the focus is on facing assumption denial head on by exposing this often-ignored reality. The current language research status quo still shrouds this issue in latent smoke and mirrors. Manifesting the checking and communicating of test assumptions is an essential prerequisite for conducting advanced quantitative methods in second language research (Brown 1992, 2015; Plonsky 2015a, b).

One of the major aims of inferential statistics (using parametric statistical testing procedures) is to facilitate valid inferences. Not reporting the checking of statistical test assumptions is the first step toward violating the validity of inferences. One of the first points to determine is which hierarchical level is the data scaled upon. If the data scale is at the interval or ratio level (refer to Table 3.1), then the researcher should check the assumptions for the proposed parametric statistical testing procedure. In contrast, if the data scale is at the nominal or ordinal level, then the researcher should check the proposed nonparametric statistical testing procedure. Central to this checking procedure is the concept of measurement. In Table 3.1, measurement is proposed to be at a higher quantitative level hierarchically than numerical coding on a lower qualitative level (Michell 1999).

Table 3.1 Statistical data testing analyses

Statistical Analysis	Data Level	Data Mode	Data Scales	Testing Procedure
Inferential	Measurement	Quantitative	Ratio	Parametric
			Interval	
	Numerical Coding	Qualitative	Ordinal	Nonparametric
			Nominal	
Descriptive	Measurement	Quantitative	Ratio	Parametric
			Interval	
	Numerical Coding	Qualitative	Ordinal	Nonparametric
			Nominal	

Nominal and ordinal scales of data typically use discrete variables. On the other hand, interval and ratio scales of data typically use continuous variables. Transforming raw scores (ordinal scale at a maximum) to measures (interval scale at a minimum) involves the action of constructing. An efficient methodology for constructing interval measures from ordinal raw scores is the Rasch model paradigm (Rasch 1960). Transforming nonlinear raw scores into linear measures scaled in logits (logarithmic odd units) via Rasch model parameter estimations has been recommended by testing researchers (Rasch 1960; Wright and Stone 1979; Brown 2015; Knoch and McNamara 2015).

This restricted conceptualization of modern measurement as displayed in Table 3.1 differentiates between the terms of *measurement* and *numerical coding* (Michell 1997, 1999). This contrasts with a more open-ended notion of measurement, for example, the one Stevens (1946) proposed defining and classifying measurement scales. Measurement was defined by Stevens (1946) as “the assignment of numerals to objects according to rules” (p. 677). However, Stevens’ (1946) definition was ambiguous. As a result of this vagueness, Stevens’ (1946) measurement definition and four levels are still widely quoted and adopted today in introductory statistical textbooks. Several problems with Stevens’ (1946) measurement levels and suggested statistical techniques are examined next.

Stevens (1946) created the measurement level terms of nominal, ordinal, interval and ratio. The nominal scale is the most basic form and allows “unrestricted assignment of numerals” (Stevens 1946, p. 678). Stevens (1946) was aware of the controversial inclusion of the nominal scale as *measurement*, stating “quite naturally there are many who will urge that it is absurd to attribute to this process of assigning numerals the dignity implied by the term measurement” (p. 679).

The ordinal scale represents a rank-ordering process. “As a matter of a fact, most of the scales used widely and effectively by psychologists are ordinal scales” (Stevens 1946, p. 679). Stevens (1946) then goes on to warn, “In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply a knowledge of something more than the relative rank-order of data” (p. 679). Unfortunately, after his caveat, Stevens (1946) reverses his stance on this issue, watering it down and making the issue ambiguous. Stevens (1946) states,

> On the other hand, for this ‘illegal’ statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results. While the outlawing of this procedure would probably serve no good purpose, it is proper to point out that means and standard deviations computed on an ordinal scale are in error to the extent that the excessive intervals on the scale are unequal in size. When only the rank-order of data is known, we should proceed cautiously with our statistics, and especially with the conclusions we draw from them. (p. 679)

The interval scale has an arbitrary zero point and the scale is made up of invariant units. “Almost all the usual statistical measures are applicable here, unless they are the kinds that imply a knowledge of a ‘true’ zero point” (Stevens 1946, p. 679). “Most psychological measurement aspires to create interval scales, and it sometimes succeeds. The problem usually is to devise operations for equalizing the units of the scales—a problem not always easy of solution but one for which there are several possible modes of attack” (p. 679). The responsibility is on the language test researcher to construct interval level measures after data collection.

The ratio level scale occurs when all four relations, “equality, rank-order, equality of intervals, and equality of ratios” exist (Stevens 1946, p. 679). A theoretical zero point is implied. Every type of statistical procedure can be used as, “(a)ll types of statistical measures are applicable to ratio scales” (p. 680).

For most language test researcher needs, constructing and then using an interval scale transformed from ordinal level raw scores is a necessary first step for conducting parametric statistical testing procedures.

3.4 Insights Gained

The insights gained were that the process of monitoring statistical test assumptions can be addressed even before data collection takes place. Prior to data collection, three methods (theory, empirical evidence and reason) can be evaluated (Wells and Hintze 2007). Within theoretical evidence, three domains are substantive, measurement and statistical theory. Within empirical evidence, two domains are prior research and pilot studies. Reason entails interpreting all the available evidence from theoretical and empirical evidence to make rational decisions about study design (Wells and Hintze 2007). In the planning stage, “using theoretical knowledge, prior empirical evidence, and reason allows one to address assumptions before the data are even collected” (Wells and Hintze 2007, p. 501). This alternative way examines the “assumptions while designing the study (i.e., prior to data collection) using established theoretical knowledge, prior empirical evidence, and reason to conclude which assumptions will likely be satisfied in the population. This has the advantage of being proactive and preventative” (Wells and Hintze 2007, p. 503). Therefore, one can preplan the checking of statistical test assumptions before data collection. Accordingly, a few key issues regarding the pre-checking of assumptions follows.

In the initial stage, test researchers should follow two preliminary steps: constructing and reporting. First, test researchers should construct interval level linear measures. The portal to parametric statistical testing procedures is using a continuous variable with interval level measurement data. The Rasch model (Rasch 1960) measurement paradigm facilitates this process of transforming ordinal level raw scores into interval level linear measures. "Data should be measured at least at the interval level. This assumption is tested by common sense" (Field 2009, p. 133). "To say that data are interval, we must be certain that equal intervals on the scale represent equal differences in the property being measured" (Field 2009, p. 9). However, test users must be careful as "(v)ariables like this that look interval (and are treated as interval) are often ordinal" (Field 2009, p. 9). Therefore, whether the data are interval or not is important to determine before proceeding with statistical analyses. Secondly, after constructing interval level data, descriptive statistics, including the arithmetic means and the standard deviations, can then be safely calculated and reported.

Next, in the parametric statistics main stage, the assumptions of the planned parametric statistical testing procedures should be checked and reported. However, "it is important to note that no assumptions will ever be *strictly* satisfied" (Wells and Hintze 2007, p. 502), "Therefore, we should design studies and select statistical analyses that are robust to assumption violations (e.g., groups of equal size, large sample sizes, etc.) whenever possible" (Wells and Hintze 2007, p. 502). Furthermore, "in the event we are not confident in a certain assumption, a statistical test should be selected that does not require that particular assumption; that is, we should err on the side of caution and thus choose a statistical test that is either robust to the assumption violation or has very few assumptions" (Wells and Hintze 2007, p. 502).

In the next section, the parametric statistical test assumptions will be outlined for descriptive and inferential statistical tests. For descriptive statistics, relevant to the particular study, measures of central tendency (the arithmetic means), measures of variability (the standard deviations) and measures of associations (the Pearson's Product-Moment Correlation coefficient), all assumptions should be checked and reported. For inferential statistics, relevant to the particular study, measures of means comparisons (the One-Way Independent Analysis of Variance and the Independent *t*-test), all assumptions should be checked and reported. The rationale behind focusing on ANOVAs, *t*-tests and correlations is that they are very common in the language testing published literature. Referring to the three time periods of 1980–1989, 1990–1999 and 2000–2009, "the two most commonly used means-based analyses-*t* tests and ANOVAs-increased steadily over time" (Plonsky and Gass 2011, p. 347). The three most commonly used statistical procedure analyses in L2 research were ANOVA (56%, or 341 of 606 studies), *t*-tests (43%, or 263 of 606 studies) and correlation (31%, or 189 of 606 studies) (Plonsky 2013). Language testing studies were a subset of these studies.

The assumptions that should be checked and reported for each statistical technique follow. Note that for all the following parametric statistical techniques, including both the descriptive and inferential statistics, interval or ratio level data is required.

3.4.1 Descriptive Statistics: Arithmetic Mean

The arithmetic mean, “the most commonly employed measure of central tendency, is the average score in a distribution” (Sheskin 2007, p. 7). The assumption of interval or ratio level data is required. If this assumption is not met, the nonparametric option, that is, the median, should be utilized (Sheskin 2007).

3.4.2 Descriptive Statistics: Standard Deviation

The standard deviation is one of the “most commonly employed measures of variability in both inferential and descriptive statistics” (Sheskin 2007, p. 11). The assumption of interval or ratio level data is required. If this assumption is not met, the nonparametric option, that is, the interdecile or interquartile range, should be utilized (Sheskin 2007).

3.4.3 Descriptive Statistics: Pearson’s Product-Moment Correlation Coefficient

Both the predictor variable and the criterion variable require interval or ratio level data (Frey 2016). There are five assumptions that should be checked and communicated (Sheskin 2007).

- Random selection of the sample from the representative population.
- Both variables have a bivariate normal distribution.
- Homoscedasticity refers to when the relationship between both the variables “is of equal strength across the whole range of both variables” (Sheskin 2007, p. 1223).
- Residuals independence is when the “degree of error with respect to prediction” (Sheskin 2007, p. 1223) are independent of each other.
- Linearity relationship shape of values on the scatterplot.

If the assumptions are not met, the nonparametric statistical alternative is Spearman’s rank-order correlation coefficient.

3.4.4 Inferential Statistics: One-Way Independent Analysis of Variance

The one-way independent ANOVA is also known as the single-factor between-subjects analysis of variance. The dependent variable requires interval or ratio level data but the independent variable requires nominal level data of two or more categories (Frey 2016). There are three assumptions that should be checked and communicated (Sheskin 2007).

- Random selection of each sample from the representative population.
- Normal distribution of the population data from which each sample is representative.
- Homogeneity of variances refers to when the variances of the samples are equal.

The assumption of normal distribution is fairly robust against violations. If any of the other assumptions are violated, "the reliability of the computed test statistic may be compromised" (Sheskin 2007, p. 869). If the assumptions are not met, the nonparametric statistical alternative is the Kruskal-Wallis one-way analysis of variance by ranks.

3.4.5 Inferential Statistics: Independent T-Test

The independent *t*-test is also known as, the *t*-test for two independent samples. The dependent variable requires interval or ratio level data but the independent variable requires nominal level data of two categories (Frey 2016). There are three assumptions that should be checked and communicated (Sheskin 2007).

- Random selection of each sample from the representative population.
- Normal distribution of the population data from which each sample is representative.
- Homogeneity of variance refers to when the variances of the samples are equal.

If any of these assumptions are violated, "the reliability of the *t* test statistic may be compromised" (Sheskin 2007, p. 427). If the assumptions are not met, the nonparametric statistical alternative is the Mann–Whitney *U* test.

3.5 Conclusion: Implications for Test Users

One purpose of increasing language testing statistical literacy is avoiding faulty statistical methods thinking. "(M)*ethodological thought disorder* is the sustained failure to cognize relatively obvious methodological facts. It is well known that many psychologists are ignorant of important methodological facts and their methodological thinking is often erroneous" (Michell 1997, p. 374). This is a point that not only psychologists should avoid but also language test researchers.

The implications for language test users are on two levels: local and global. At the local level, language testing researchers should increase their ongoing statistical literacy through individual professional development. This can be achieved by making efforts to learn more about parametric statistical procedure assumptions. Additionally, language test users should adopt a self-responsibility policy of checking and communicating the relevant assumptions for the statistical techniques they plan on using. Avoiding these local issues leads to liability consequences related to test fairness and justice.

At the global level, language testing trainers need to raise awareness and revise educational programs to educate current and future generations of language test users and researchers. The place to start is with education about the most commonly employed statistical techniques in language testing. In parametric statistical testing procedures, the *Big Three* is typically employed, that is, “nearly all quantitative studies employ *t* tests, ANOVAs, and/or correlations. In many cases, these tests are viable means to address the research questions at hand; however, problems associated with these techniques arise frequently (e.g., failing to meet statistical assumptions)” (Plonsky 2015a, p. 3).

At a minimum, if only one set of statistical technique assumptions is chosen, then education programs should concentrate on ANOVA. A statistical quality study proposed a call for reform that targeted six groups of stakeholders (Plonsky 2014). Specifically, one group of stakeholders was graduate curriculum committees and research trainers. One of the recommendations concerning analysis of variances was, “ANOVA’s status as the test of choice in L2 research is not likely to change soon and ANOVAs should and will continue to be used, so graduate students need to know how to test the assumptions of ANOVA” (Plonsky 2014, p. 466). This need appears so far to have been unfulfilled. Indeed, as previously illustrated in this chapter, the most commonly used statistical procedure analysis in L2 research was ANOVA (56%, or 341 of 606 studies), and yet overall only 17% (101 of 606 studies) reported checking statistical assumptions (Plonsky 2013).

The time for change is long overdue. Changing the consensus so that checking and communicating parametric statistical procedure assumptions occur in all studies is an important part of future language testing design best practices. If language test users continue to ignore this dilemma, there will be greater future consequences. Indeed, big global changes can start with just small local changes. Only a change in mindsets toward checking and communicating test researchers’ transparent actions will achieve these ends.

References

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, 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.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Frey, B. B. (2016). *There’s a stat for that! What to do & when to do it*. Thousand Oaks, CA: Sage.

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275–304). New York, NY: Routledge.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago, IL: University of Chicago Press.

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., et al. (2013). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly, 48*(2), 360–388.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355–383.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. 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. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *The Modern Language Journal, 98*(1), 450–470.

Plonsky, L. (2015a). Introduction. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 3–8). New York, NY: Routledge.

Plonsky, L. (2015b). Statistical power, *p* values, descriptive statistics, and effect sixes: 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., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61*(2), 325–366.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Expanded ed., 1980). Copenhagen and Chicago, IL: Danish Institute for Educational Research and University of Chicago Press.

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.

Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools, 44*(5), 495–502.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press.

Chapter 4
Washback of the Reformed College English Test Band 4 (CET-4) in English Learning and Teaching in China, and Possible Solutions

Feifei Han

Abstract This chapter reviews the key research on the washback of the College English Test Band 4 Test (CET-4), a compulsory high-stakes, large-scale, and nationwide English proficiency test to measure English proficiency of non-English major university students in China. The review results show a mixture of positive and negative effects of washback on both English learning and teaching. To reduce the negative effects of the washback, the following solutions are proposed: (1) The quality assurance body of College English teaching should use a wider range of indicators to gauge the quality of English teaching and learning rather than solely relying on the outcomes of the CET-4. (2) The Spoken English Test should be designed as a compulsory subset so that the English-speaking skill would no longer be devalued and ignored in College English teaching and learning. (3) The proportion of the Chinese to English translation section should be decreased to discourage using rote memorization of bilingual vocabulary lists as a main test preparation strategy. (4) The CET-4 should use the integrated format to replace the separate testing of listening, reading, and writing so that communicative English competence can be effectively assessed.

4.1 Introduction: Purpose and the Testing Context of the CET-4

China has the largest number of English language learners in the world. It is estimated that more than 400 million Chinese learn English as a foreign language across the nation (Wei and Su 2012). Among them, around one-tenth of the learners are university students, and College English is a compulsory subject for all freshmen and sophomores in China (Cheng and Curtis 2010). Chinese university students take a variety of high-stakes international and national English tests to satisfy their different purposes and needs. In order to study overseas in an English-speaking country, they

F. Han (🖂)
Griffith University, Brisbane, Australia
e-mail: feifei.han@griffith.edu.au

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_4

take TOEFL, IELTS, and the Pearson Test of English. In order to be competitive in job markets, they take TOEIC and Cambridge English Qualifications: Business. Apart from these international English tests, the most important large-scale standardized English test for Chinese university students at the national level is the College English Test (CET). The CET has two levels: level 4 (known as CET-4) and level 6 (known as CET-6), both of which are administered by the Ministry of Education in China (CMoE). While both tests examine the English proficiency of Chinese non-English major university students, CET-4 is compulsory, whereas CET-6 is not (Yang and Weir 1998). The compulsory nature of the CET-4 renders it to have high-stakes status, with around 10 million test takers annually (Zheng and Cheng 2008). CET-4 is developed by the National College English Testing Committee (NCETC) and is held twice a year in June and December. Currently, the test is open only to the currently enrolled undergraduates, and the registration of the test must be completed through the test takers' universities.

First launched in 1987 following the implementation of the first College English Curriculum (CMoE 1985, 1986), the initial development of the CET-4 aimed to achieve two goals. The first goal was to provide an objective assessment and evaluation of a university student's overall English proficiency. The second goal was to direct and to unify the College English teaching nationwide (Ma 2014).

Throughout its 31 years of administration and development, the CET-4 has undergone two waves of major reforms in terms of content, format, and scoring system (Jin 2017). The first wave of reform was implemented throughout the 1990s, and this wave involved three noticeable changes. The initial test items in the CET-4 were predominantly multiple-choice questions, which accounted for as much as 85% of the test items, and only 15% were essay writing for testing English writing skills. Although having a high percentage of multiple-choice questions has potential advantages in terms of objectivity in testing and efficiency for scoring for a large-scale test, and are appropriate for testing receptive skills, such as listening and reading skills, and the receptive knowledge of vocabulary and grammar, this format has apparent drawbacks for evaluating productive skills (Jin and Wu 2017). To solve this issue, the first change of the CET-4 was the inclusion of a variety of formats, including dictation of short English phrases and sentences, short answer questions, and translation from English to Chinese. The second major change concerned the scoring, with the revised scoring system emphasizing writing. A threshold was set for the writing subset, and not achieving the threshold in writing resulted in a penalty of the total score of the CET-4. If a student scored zero for writing, he/she had to retake the whole test, no matter what level of performance he/she had achieved for the remainder of the test. The third major change was the introduction of a spoken test in 1999, known as CET Spoken English Test (CET-SET) to assess students' English-speaking skills (Zhang 2005). However, the CET-SET was not open to all candidates, as the test takers had to achieve an overall 80 out of 100 points in the test to be eligible for the CET-SET. The CET-SET had three-level grades: A, B, and C. Grades lower than C did not produce a report, as this indicated not having a sufficient level of English-speaking skills.

The second wave of reform of the CET-4 was launched in 2006 to meet the needs of the transformation of the 1999 version to the 2007 version of the National College

English Teaching Curriculum and Syllabus and Requirements (CMoE 1999, 2007) as part of the large-scale project—Higher Education Undergraduate Level Teaching Quality and Teaching Reform by the CMoE (2006). The wave of the reform had two major purposes: (1) to meet "the pressing social need for college and university graduates with a stronger communicative competence in English" (Jin and Yang 2006, p. 21), and (2) to "maximize its positive backwash effect on teaching and beneficial impact on society" (p. 34). The reform was implemented by the NCETC, which constructed new tasks to replace some old ones in order to test contextualized English use, rather than the context-free of English language knowledge (NCETC 2006).

The reform included four major changes: (1) increasing the proportion of testing listening comprehension from 20 to 35%; (2) replacing the multiple-choice style of assessing English vocabulary and structure in single sentences with a contextualized cloze test; (3) adding fast reading to assess learners' skimming and scanning abilities in reading longer English texts; and (4) replacing translation from English to Chinese with translation from Chinese to English. The structure of the reformed CET-4, including subsets/skills, contents, formats, proportions, and time distribution, is displayed in Table 4.1.

Apart from the revision of the test format, the scoring system has also been dramatically changed. In the previous scoring system, the test takers only received a certificate indicating if they achieved a pass (60) or distinction (85) out of 100 points. The new scoring system is norm-referenced; hence the scores show how an individual test taker has performed relative to the whole group. The maximum achievable score is 710 points, and the test takers receive a report of the total score and the scores of the subsets of listening (maximum achievable score is 248.5), reading (maximum achievable score is 248.5), writing (maximum achievable score is 106.5),

Table 4.1 The structure of the reformed CET-4

Subsets	Contents	Formats	Number of questions	Proportions (%)	Time distribution
Listening	3 short news	multiple-choice	7	7	25 minutes
	2 long conversations	multiple-choice	8	8	
	3 short texts	multiple-choice	10	20	
Reading	vocabulary	cloze test	10	5	40 minutes
	fast reading	information matching	10	10	
	in-depth reading	multiple-choice	10	20	
Translation	Chinese to English	paragraph translation	1	15	30 minutes
Writing	composition	short essay	1	15	30 minutes
Total	–	–	57	100	125 minutes

and translation (maximum achievable score is 106.5). The new scoring system has also removed "pass" to prevent universities from using the CET-4 pass as a compulsory requirement for students to graduate and to prevent employers from using it as a mandatory selection criterion for job interviews.

These strategies aimed at reducing the pressure on the test takers and English teachers in the hope that the test can induce positive washback effects in College English learning and teaching (Jin 2010). Therefore, since the introduction of the second reformed CET-4, a few studies have been conducted to examine if the test has achieved its goal of stimulating positive effects. This review only covers the key studies on the washback of the reformed CET-4 (after 2006). The findings of the key studies are summarized, and the problems in terms of negative washback are outlined in the next section.

4.2 Testing Problems Encountered

In the language testing literature, washback is defined as "the effect of testing on teaching and learning" (Hughes 2003, p. 1). A test is able to produce both positive and negative washback on teaching and learning (Bachman and Palmer 2010). A search of the literature identified nine key studies, whose results showed a mixture of positive and negative washback of the CET-4 on Chinese students' English learning and the instruction of College English by teachers. This review excluded the studies which only focused on investigating the washback of one of the subsets tested in the CET-4 (e.g., the listening subset: Hou and Wang [2008], Shi [2010], Wang [2010]; the speaking subset: Zhuo [2017]). The results of the nine key studies can be broadly categorized as the washback of the CET-4 on learning and on teaching. For each category, both positive and negative effects were observed and will be discussed in turn. A detailed summary of the nine key studies, including types of research, participants and data collection methods, and key findings, is presented in Appendix 4.1.

Concerning the washback effect on English learning, the reformed CET-4 produced three major positive effects. First, the CET-4 test not only enabled students to put much time and effort in preparing for it, but also motivated students to make extra efforts in learning English (i.e., Li et al. 2012; Shao 2006; Sun 2016). The survey in these three studies out of the nine key studies reported that the majority of students felt the positive side of having to sit for the CET-4 was that they were motivated to learn English. Second, apart from putting much effort into English learning, some students also reported that they felt CET-4 made them more aware of the goals of English learning and kept them focused on the goals (i.e., Li et al. 2012). Third, CET-4 encouraged students to use cognitive strategies and test management strategies in the test (i.e., Xiao 2014).

There were also three prominent negative effects of the CET-4 on English learning. To start with, in five out of the nine key studies, the results showed that students adopted English learning strategies oriented toward passing the test rather than strategies enhancing language use competence and developing English communicative abilities, in the processes of test preparation (i.e., Ren 2011; Sun 2016; Xiao 2014; Xie and Andrews 2012; Zhan and Andrews 2014). These test preparation strategies encouraged rote memorization of linguistic forms, such as memorizing English–Chinese bilingual vocabulary lists; using grammar-translation methods to learn English and to prepare for the test; and developing test-wise skills through practicing the mock or past CET-4 papers, by means such as analyzing test papers, rehearsing test-taking strategies, practicing sample test papers intensely, and memorizing the model English essays. Students also used past tests and sample CET-4 papers as their main sources to learn English. Furthermore, students put much effort into and gave much weight to practicing reading and listening, which were two major skills tested in CET-4. On the other hand, they neglected writing and speaking, because writing only accounted for a small proportion and speaking was not compulsory (i.e., Li et al. 2012). Last, apart from the negative effects on cognition, CET-4 also produced psychological effects detrimental to students, as the test aroused pressure and anxiety in their English learning (i.e., Li et al. 2012).

In terms of the washback on English teaching and teachers, three major positive and negative effects were identified. In terms of positive washback on teaching, firstly, a greater importance was attached to English teaching due to the compulsory nature of CET-4 (i.e., Gu 2007). Secondly, CET-4 promoted the implementation of the new version of the National College English Teaching Curriculum and Syllabus and Requirements in the College English instruction (i.e., Gu 2007). Thirdly, English teaching shifted from grammar drilling to developing communicative competence in English, and teachers tended to integrate the four skills in their instruction (i.e., Chen 2007).

The first major negative impact of CET-4 on English teaching was its narrowing of the content of instruction (i.e., Gu 2007; Ren 2011; Shao 2006). The teaching strategies, teaching materials (e.g., the sample test papers are exclusively used in teaching), and teaching activities (e.g., teachers only practice the skills tested in CET-4 but ignore the skills not tested in CET-4, such as spoken English) tended to be exclusively focused on when CET-4 was approaching. As a result, teachers were unwilling to develop students' communicative competence, and the teaching lacked creativity. In the time period before CET-4, the normal schedule of English teaching was always disrupted, and teachers found it difficult to follow the syllabus and to complete the contents in the English textbooks.

Furthermore, CET-4 also negatively affected the content of classroom assessment (i.e., Ren 2011). The assessment tasks were predominantly designed to resemble those in CET-4, resulting in a lack of formative assessment of English learning. Last but not least, CET-4 cast a negative influence on teachers and teaching (Chen 2007; Gu 2007; Ren 2011). The results of students' performance on CET-4 were used heavily to benchmark the quality of English teaching and were given much weight for English teachers' promotion, which created considerable pressure for teachers.

4.3 Solution/Resolution of the Problems

Some possible solutions are proposed in order to mitigate the negative washback produced by the CET-4. (1) The quality assurance body of the universities should use a wider range of indicators to gauge the quality of College English teaching and learning rather than solely relying on the outcomes of the CET-4. (2) The CET-SET should be integrated as a compulsory subset for all the test takers so that English-speaking skills are no longer devalued in College English teaching and learning. (3) The weights of some subsets of the tests should be modified. In particular, the writing subset should be given more weight, and the proportion of the Chinese to English translation section should be reduced to discourage using rote memorization of vocabulary lists. (4) The NCETC should consider developing a test format which integrates the assessment of the four skills, so that communicative competence can be emphasized in English learning and teaching.

4.4 Insights Gained

To mitigate the negative effects generated by CET-4, the National College English Testing Committee (NCETC) may need to consider the following aspects. Most importantly, due to the importance of communicative competence in English learning, testing the speaking and listening proficiency of learners cannot simply be ignored. While the current CET-SET is optional, how to properly integrate it into the CET-4 and make it compulsory remains a critical issue for the NCETC to solve in order for the speaking and listening skills to be appropriately assessed. Accordingly, such implementation may direct the curriculum of college English teaching to place much more emphasis on training students' communicative competence rather than predominantly focusing on reading and writing per se. Secondly, since the current use of the CET-4 results are closely linked with benchmarking teaching quality, the contents and formats of teaching tend to be greatly impacted by the contents and format of CET-4. Hence, the National Educational Examinations Authority should consider making a policy to regulate how the test results should be used in the universities to minimize the negative influence of CET-4 on normal English teaching.

4.5 Conclusion: Implications for Test Users

This review indicates that there is a lack of nationwide research on the washback of the reformed CET-4 since Gu (2007). Considering the high-stakes status of the test, it is suggested that the NCETC should conduct large-scale studies on the washback of CET-4 so that the intended use, formats and contents of the test may be modified to address negative effects. The NCETC should also make the interpretation of the test scores transparent by possibly including qualitative and descriptive references as to what a test taker “can do” in English and/or the subskills of English for a particular range of scores and/or sub-scores by a test taker (Jin 2011).

This review also has important implications for stakeholders of the CET-4, including government, employers, and universities. It is risky to rely solely on the scores of the CET-4 as a means to gauge English learning and teaching. Students and teachers should develop strategies to develop English competence in the long run, which may in turn facilitate achieving a desirable level of performance in CET-4.

Appendix 4.1: Summary of the Key Studies on the Washback of the Reformed CET-4

Studies	Research method	Participants and data collection methods	Positive effects	Negative effects
Shao (2006)	mixed-methods	45 teachers from three teachers’ colleges were surveyed 356 students from the three teachers’ colleges were surveyed Of 45 teachers, four were interviewed Each interviewed teacher was observed for four hours of English teaching	70% of the students reported that they were motivated to learn English due to the CET-4	The CET-4 did not exert much influence on freshmen However, the CET-4 produced negative effect for sophomores, who would sit the CET-4 soon The teaching strategies, teaching materials (i.e., The sample papers were exclusively used in teaching), and teaching activities (e.g., Only practicing the skills tested in the CET-4 but ignoring the skills which were not tested in the CET-4, such as the spoken English) were exclusively focused on the CET-4, as the test was approaching

(continued)

(continued)

Studies	Research method	Participants and data collection methods	Positive effects	Negative effects
Chen (2007)	mixed-methods	154 teachers were surveyed Of the surveyed teachers, 112 were interviewed	Teachers felt that the revised CET-4 had positive effects on their curricular planning and instruction: (1) 100% of teachers reported that they were motivated to integrate listening and speaking skills in the teaching activities rather than only targeted reading and writing as in their previous teaching (2) Over 87% teachers reported that the teaching focus was shifted from grammar drilling to developing communicative competence	A positive correlation was found between teachers' perceptions of the importance of the CET-4 and the perceptions of the pressure in their teaching
Gu (2007)	mixed-methods	2609 students nationwide were surveyed 1220 teachers nationwide were surveyed English teaching of 38 teachers at three universities was observed Focus-group interviews were conducted with teachers, students, and administrators	(1) The majority of the teachers, students, and administrators felt positively about the CET-4 because the university attached greater importance to English teaching and learning due to the compulsory nature of the test (2) The CET-4 also promoted the implementation of the new version of the *National College English Teaching Curriculum and Syllabus and Requirements*	(1) When the CET-4 approached, teachers used the CET-4 preparation materials predominantly in teaching. As a result, they could not cover all the contents in the English textbooks, and the pace of English teaching was also accelerated (2) The administrators used the CET-4 results as the sole indicator to evaluate the quality of English teaching quality and students' English proficiency

(continued)

(continued)

Studies	Research method	Participants and data collection methods	Positive effects	Negative effects
Ren (2011)	mixed-methods	35 teachers were surveyed 210 students were surveyed Among 35 teachers, 22 were interviewed Among 210 students, 30 were interviewed		(1) The CET-4 encouraged rote memorization of linguistic forms (e.g., memorizing vocabulary lists) and practicing past CET-4 papers to prepare for the test. As a result, students were unable to use English in authentic situations (2) Classroom assessments highly resembled the formats and the contents of the CET-4 (3) Teachers were unwilling to train students' communicative competence in English (4) Teachers' promotion was linked to the CET-4 success rate achieved by their students (5) It suppressed teachers' creativity in teaching
Li et al. (2012)	quantitative	150 students at a university were surveyed	Most of the students surveyed felt the CET motivated them to invest greater effort to learn English (2) The CET-4 enabled students to set a clearer goal for English learning	(1) The students felt pressure and anxiety in English learning (2) The students put much effort in practicing reading and listening skills, which were given more weight in the CET-4, and they tended to neglect the writing and speaking skills, as the two skills were either given little weight or not tested

(continued)

(continued)

Studies	Research method	Participants and data collection methods	Positive effects	Negative effects
Xiao (2014)	quantitative	284 students were surveyed	The CET-4 moderately promoted cognitive strategy use and weakly promoted test management strategy use	The students' English learning strategies were test-oriented (e.g., using test-wise strategies) rather than developing competence of language use
Xie and Andrews (2012)	quantitative	870 students were surveyed		The students' perceptions of test design and test use affected their test preparation strategies, which were dominated by analyzing past test papers, rehearsing test-taking skills, practicing sample test papers intensively, and memorizing model writing essays in a rote manner
Zhan and Andrews (2014)	qualitative	24 students were required to keep diaries and were also interviewed following the diary entries		(1) The students favored learning strategies of rote memorization over developing their English communicative competence (2) The students used the past CET-4 papers as the only learning materials
Sun (2016)	mixed-methods	A CET-4 test developer from the development committee was interviewed Eight CET-4 test users from educational and social contexts (two deans of the two universities, two directors of the government employment offices; and four human resources managers of four companies) were interviewed 416 students from two universities were surveyed and they also took a retired CET-4	The students' perceptions of the high demands of the CET-4 led them to spend more effort in preparing for it	The student had high perception of instrumental uses of the CET-4. Such perception had stronger association with use of rehearsing and cramming strategy (e.g., memorizing model essays to prepare for the CET-4 writing) than use of long-term skill development strategies

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Chen, F. (2007). The washback effect of College English Test Band 4 on curricular planning and instruction. *CELEA Journal, 30*(1), 19–29.

Cheng, L., & Curtis, A. (2010). The realities of English language assessment and the Chinese learner in China and beyond: English language assessment and the Chinese learner. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 3–12). New York: Routledge.

CMoE. (1985). *National College English curriculum syllabus for science and technology students*. Shanghai: Foreign Language Education Press.

CMoE. (1986). *National College English curriculum syllabus for liberal arts students*. Shanghai: Foreign Language Education Press.

CMoE. (1999). *College English curriculum syllabus* (revised version). Shanghai: Shanghai Foreign Language Education Press.

CMoE. (2006). *Quality and reform of college English education*. http://www.moe.gov.cn/publicfiles/business/htmlfiles/moe/s3857/201011/xxgk_110823.html. Accessed 13 Feb 2019.

CMoE. (2007). *College English curriculum requirements*. http://old.moe.gov.cn/publicfiles/business/htmlfiles/moe/moe_1846/200711/28924.html. Accessed 13 Feb 2019.

Gu, X. (2007). *Positive or negative: An empirical study of CET washback*. Chongqing: Chongqing University Press.

Hou, X., & Wang, W. (2008). A study of the college students' attitudes towards the new CET-4 listening subset: Washback effect of the new CET-4 listening subtest on students. *Journal of Xi'an International Studies University, 16*(3), 91–94.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Jin, Y. (2010). The National College English Testing Committee. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 44–59). New York: Routledge.

Jin, Y. (2011). Fundamental concerns in high-stakes language testing: The case of the College English Test. *Journal of Pan-Pacific Association of Applied Linguistics, 15*(2), 71–83.

Jin, Y. (2017). Construct and content in context: Implications for language learning, teaching and assessment in China. *Language Testing in Asia, 7*(1). https://doi.org/10.1186/s40468-017-0044-1.

Jin, Y., & Wu, E. (2017). An argument-based approach to test fairness: The case of multiple-form equating in the College English Test. *International Journal of Computer-Assisted Language Learning and Teaching, 7*(3), 58–72.

Jin, Y., & Yang, H. (2006). The English proficiency of college and university students in China: As reflected in the CET. *Language, Culture and Curriculum, 19*(1), 21–36.

Li, H., Zhong, Q., & Suen, H. (2012). Students' perceptions of the impact of the College English Test. *Language Testing in Asia, 2*(3), 77–94.

Ma, F. (2014). College English Test: To be abolished or to be polished. *Journal of Language Teaching and Research, 5*(5), 1176–1184.

NCETC. (2006). *CET-4 Test syllabus and sample test paper* (2006 revised version). Shanghai: Shanghai Foreign Language Education Press.

Ren, Y. (2011). A study of washback effects of College English Test (Band 4) on teaching and learning English at tertiary level in China. *International Journal of Pedagogies and Learning, 6*(3), 243–259.

Shao, H. (2006). An empirical study of washback from CET-4 on college English teaching and learning. *CELEA Journal, 29*(1), 54–59.

Shi, X. (2010). A longitudinal study on the washback effect of the new CET-4 and CET-6 listening subset. *Foreign Language World, 177*(3), 80–86.

Sun, Y. (2016). *Context, construct, and consequences: Washback of the College English Test in China* (Unpublished doctoral dissertation). Queens' University, Kingston, Canada.

Wang, W. (2010). Investigating the washback effect of the new CET-4 listening comprehension subtest on language learners. *Chinese Journal of Applied Linguistics, 33,* 28–39.

Wei, R., & Su, J. (2012). The statistics of English in China. *English Today, 28*(3), 10–14.

Xiao, W. (2014). The intensity and direction of CET washback on Chinese college students' test-taking strategy use. *Theory and Practice in Language Studies, 4*(6), 1171–1177.

Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation: Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49–70.

Yang, H., & Weir, C. J. (1998). *The CET validation study*. Shanghai: Shanghai Foreign Language Education Press.

Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: Insights from possible self-theories. *Assessment in Education: Principles, Policy & Practice, 21*(1), 71–89.

Zhang, Z. (2005). A historical review of CET 4/6. *Modern Language Journal, 18*(2), 100–103.

Zheng, Y., & Cheng, L. (2008). College English Test (CET) in China. *Language Testing, 25*(3), 408–417.

Zhuo, W. (2017). The Washback effect of CET Spoken English Test upon college English teaching. *Canada Social Science, 13*(1), 62–68.

Chapter 5
Fairness in College Entrance Exams in Japan and the Planned Use of External Tests in English

Yuko Goto Butler and Masakazu Iino

Abstract The Japanese government recently decided to replace the English section of the nationwide college entrance exam with external proficiency tests. This policy was motivated by the desire to improve the speaking proficiency of students by directly assessing it in college entrance examinations. However, in Japan, an English-as-a-foreign-language context, students' English proficiency, and speaking ability in particular, is greatly influenced by socioeconomic status (SES) because students need to seek greater opportunities to develop English-speaking skills. The accessibility and affordability of taking external tests are also influenced by students' SES. Issues regarding the fairness of this policy need to be carefully examined. In this paper, we consider a series of potential rebuttals that would weaken the fairness of assessment in the validity arguments regarding the use of external tests in this policy. We also identify fairness issues that are critical for major stakeholders in this reform. And finally, we raise questions concerning the basic premises underlying this policy, including arguments for a positive washback effect caused by the speaking tests on primary and secondary school English education and the importance of English-speaking abilities for a globalizing world.

5.1 Introduction

Beginning in 2020, the Japanese government will test high school students' English-speaking skills as part of the nationwide college entrance exam (referred to as the *Common Test*) using external standardized proficiency tests. To satisfy this requirement, students can choose from among eight external assessments: TOEFL iBT, TOEIC, IELTS, Cambridge English tests, Eiken tests, GTEC, TEAP-PBT, and TEAP-CBT (Ministry of Education, Culture, Sports, Science and technology: MEXT

Y. G. Butler (✉)
University of Pennsylvania, Philadelphia, PA, USA
e-mail: ybutler@upenn.edu

M. Iino
Waseda University, Tokyo, Japan
e-mail: iino@waseda.jp

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_5

2018).[1] The first four tests are international proficiency tests, and the rest are domestic exams. This new policy was motivated by the desire to improve students' English-speaking skills by directly assessing these skills in college entrance examinations. Relying on external tests was a solution to the logistical challenges of measuring the speaking performance of a large number of students in a single day as part of the Common Test (MEXT 2017). Japanese national universities make admission decisions based on a two-step selection procedure that involves the results of both the nationwide college entrance exam (administered once a year) and in-house exams that are developed by individual universities. As part of the new policy, the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) asked national universities to accept any of these designated external tests either as a stand-alone screening for taking the universities' in-house tests or as part of a combined score with the Common Test, or both. A growing number of private universities are following this policy as well. All institutions following this policy need to decide which external tests to accept, how to determine the cut-off score for each designated test, and how to use the test results. Without reliable information about interpreting and using the scores of these external tests, however, universities are experiencing difficulty in making these decisions. Even more troubling, many universities appear to have made their decisions almost arbitrarily, without relying on any justifiable criteria (Kim and Mizuto 2019).

This policy is a great example of what McNamara and Roever (2006) described as "the manipulation of test consequences in the service of political goals, such as accountability or systematic reform, and the unintended fallout from the test" (p. 203). To illustrate their point, McNamara and Roever described Akiyama's 2004 study in which Japanese high school teachers resisted a proposal to introduce external English-speaking proficiency tests as part of high-stakes high school entrance exams. According to Akiyama, at the time the Japanese believed strongly in meritocracy and egalitarianism, and they expected tests to function purely on merit and apply equally to everybody. Moreover, it was believed that test scores should reflect one's diligence and effort, which are valued highly in Japanese society as characteristics possessed by everybody regardless of innate talent and background. High school teachers' resistance to the policy proposal in Akiyama's study stemmed, in part, from their perception that speaking performance is not a sign of diligence and effort because, for example, students can acquire English-speaking skills without making much effort if they have a chance to live in an English-speaking country. Given that English was treated as an academic subject rather than a practical subject in Japanese schools, testing students' English-speaking skills did not meet teachers' expectations for what should be tested by an entrance exam.

While there is still some expectation that entrance exam scores should reflect Japanese students' diligence and effort, in the 15 years since Akiyama's study there

[1] After we finished writing this chapter, the Japanese government announced on November 1, 2019 that they postponed the implementation of this policy (MEXT 2019). In the meantime, universities are still allowed to use these tests at their discretion. MEXT indicated that they will make a final decision on implementing this policy in a year.

have been substantial changes in educational practices and social perceptions. In schools, English teachers have gradually spent more time on “practical English” instead of focusing solely on grammar and reading instruction (Miyamoto 2018). In the broader society, there is a growing recognition that the widely held belief that Japan is an egalitarian and homogenous society is a myth. In fact, Japan has a high relative poverty rate (the 10th highest in 2015) (OECD 2015), and socioeconomic disparities are frequent topics of public discourse (Moriguchi 2017). Honda (2005) identified a new type of meritocracy—*hyper-meritocracy*—in Japanese postmodern society. In hyper-meritocracy, abilities that are viewed as highly desirable are unfortunately deeply rooted in one’s upbringing. Finally, Kariya (2008) empirically showed that students’ effort-making is not independent of their socioeconomic status (SES) in Japan.

It is with this background in mind that we argue that Japan’s new policy of testing English-speaking skills as part of college entrance exams further imperils *fairness* in Japan’s rapidly changing society. The new policy will most likely produce unintended fallouts because the external tests reflect students’ SES more than their diligence and effort; it will also likely contribute to widening social disparities. Moreover, the types of speaking abilities that are measured in the external tests are mostly irrelevant to the actual needs of the majority of Japanese students. In this chapter, we examine issues that potentially threaten the fairness and validity of the external assessments being used as part of this new policy. We also raise questions about a basic premise underlying this policy: that there is a universal, measurable (via a single test) oral communicative ability in our globalizing world.

5.2 Testing Problem Encountered

This new policy was implemented as part of a larger reform of college entrance examinations to make them more problem-solving oriented. For English, however, the main goal was to shift to measuring all four skill domains, based on the assumption that incorporating speaking into the exam will lead to greater emphasis on oral communicative skills in English education. There was also strong pressure from the business and political communities to take radical action to improve citizens’ English-speaking skills, which are viewed by many as necessary for the nation to be competitive in the global economy (Abe 2017). MEXT plans to completely replace the English portion of the Common Test (which currently mostly assesses receptive skills) with external tests starting in 2024.

The eight proficiency tests were chosen in 2018, but how they were selected is unclear. In 2017, a series of criteria for selecting external tests were released by the government, but the chosen external tests do not meet many of their criteria. For example, one of the critical requirements is that a test should be aligned with Japan’s national high school curriculum, but none of the international tests meet this requirement. Even for the four tests developed in Japan, the degree of alignment with the national curriculum is largely unclear because the test developers have not provided sufficient validity information.

Under this new policy, students can take any test(s) of their choice, but only up to two scores from the same test obtained during the 12th grade can be used for admission purposes. One could argue that having multiple opportunities to take tests will create less anxiety than having only one chance (which is the policy under the current format). But in reality, students already feel pressured to start preparing for tests early because they can practice taking tests an unlimited number of times before Grade 12 (Miyamoto 2018). Since universities can use any of these eight tests, students have to be strategic about which tests to prepare for in order to maximize their chance of being accepted by universities of their choice.

One of the biggest challenges for test users is to identify how to compare the results of the multiple tests, which vary substantially in terms of the test formats and goals as well as the targeted domains, abilities, and proficiency levels. MEXT made a conversion table based on the Common European Framework of Reference (CEFR) for test users (MEXT 2018). Critically, the table was not based on MEXT's own validation efforts; instead, MEXT simply put together information reported by the test developers, but the credibility of some of that information (i.e., validity evidence) is questionable. Curiously, MEXT modified the table a couple of times without clearly explaining the changes. For example, TOEIC has a listening and reading test (TOEIC L&R, 990 points in total) and a speaking and writing test (TOEIC S&W, 400 points in total), and the sum of the scores of these two tests (1390 points) was used in the table released by MEXT in July 2017. In the version released in March 2018, however, the TOEIC speaking and writing score was multiplied by 2.5 (1000 points) and added to the TOEIC L&R score, resulting in a total of 1990. Moreover, MEXT simply replaced the old numbers with the new aggregated scores without verifying their compatibility with CEFR (Hato 2018). Unexplained changes were made in all four domestic tests as well.

The problems discussed above are firmly rooted in the fairness of this new policy. Fairness, or the absence of bias, is a complicated notion, yielding multiple interpretations and definitions. Traditionally, fairness in Japan has often been discussed in the collectivist cultural framework, in which fairness means ensuring equal treatment of *all* members of society. This approach to fairness is often contrasted with Western-oriented conceptions of fairness, which frequently focus on equal treatment of the *individual*. However, empirical investigations do not necessarily support such dichotomous conceptualizations of fairness. For example, Kobayashi and Viswat (2007) compared Japanese and American students' perceptions of fairness in educational settings and reported "diverse viewpoints" (p. 1) in the respective groups. In any event, under either a collective or an individual view of fairness this new policy can be considered "unfair"—both because it does not ensure equal access to test takers (due to regional and socioeconomic differences) and because it does not ensure an evidence-based comparison of the scores of different tests.

In language assessment, test fairness is often discussed in relation to validity, but the relationship between fairness and validity can be conceptualized differently depending on how one defines fairness and validity (Kane 2010). For example, for

Kunnan (2004), validity is part of the quality of fairness, whereas Xi (2010) discusses fairness issues within an argument-based validity framework. Kane (2010) takes the position that fairness and validity are closely related (they essentially concern the same question) and that either one is part of the other. In this chapter, we subscribe to Kane's position because his broad conceptualization of fairness appears to fit the current complex policy context where multiple tests are involved.

According to Kane (2010), fairness can be conceived of as a combination of *procedural fairness* and *substantive fairness*. Procedural fairness stands on a core notion of fairness—"everybody should be treated in the same way"—and concerns "a lack of bias for or against any individual or group" in testing (p. 178). Substantive fairness demands that "score interpretation and any test-based decision rule be reasonable and appropriate" and, most importantly, that "they be equally appropriate for all test takers" (pp. 178–179). Procedural fairness is a necessary condition for fair and valid assessment but does not sufficiently ensure it. For test developers, procedural fairness is largely controllable, but substantive fairness is not entirely controllable.

Many of the problems with the new college entrance exam policy can be organized according to the procedural and substantive fairness frameworks. With respect to procedural fairness, first of all, basic validity and fairness information—including the results of differential item functioning (DIF, a statistical analysis detecting unexpected behaviors for certain subgroups at the item level)—is not fully available for all eight tests. For some tests, insufficient validation/fairness analyses have been carried out or reported. Second, there are a number of concerns related to test accessibility and administration. Some tests have a small number of test locations, which tend to be concentrated in large cities. This means that the accessibility of testing locations differs according to students' place of residence. Moreover, it is not uncommon for some domestic tests to have school-wide administration (students in a given school take the test together at their school). But such administration is a potential threat to fairness/validity of the tests if they are used for high-stakes admission purposes. Therefore, the test agencies need to secure sufficient locations and proctors outside high schools. Another threat to procedural fairness concerns test examination fees, which students are responsible for paying and which can vary substantially, ranging from 5800 to 26000 yen (approximately from US$52 to US$235). Such fees can be a potential hurdle for lower SES students and may influence which test they take and how many times they take it. In addition to paying test fees, students in rural areas far from testing sites might have to pay transportation and accommodation costs. In response to such concerns, some wealthier local governments are considering covering the examination fees for their residents, but this, in turn, can yield an additional potential bias by region. Finally, there are also fairness/validity concerns with respect to test scoring. For some domestic speaking tests, a large number of high school English teachers and college instructors have served as raters. Such practices are no longer acceptable from a fairness point of view, and the testing agencies must secure well-trained raters in a short period of time, which will likely be a tremendous challenge.

As pointed out by Kane (2010), resolving procedural fairness issues such as those described above is not sufficient for achieving a fair and valid assessment practice. Even if these problems are fixed, serious issues concerning substantive fairness remain. One such issue is differing access to test preparation materials and practices. The targeted test domains and proficiency levels in some of the external tests deviate substantially from national curriculum targets; if students want to perform well on those tests, they will likely need to obtain additional materials and learning opportunities beyond normal practices at school. High schools and families have varying capacities to offer additional support to these students. Even for the tests that are more or less aligned with the national curriculum, access to test preparation materials—which are often published by test agencies as well as other private entities and available for a fee—and opportunities to practice speaking English for the test within and outside of school likely differ by SES. Moreover, the misuse of test results, such as conducting inappropriate score aggregations, using invalidated and inappropriate cut scores for admission decisions, and comparing multiple test scores based on CEFR (also see Green 2018), are all serious issues of substantive fairness.

5.3 Unsolved Problems

Due to a chaotic rollout process, stakeholders have experienced tremendous frustration and confusion even before the policy implementation date. As mentioned, universities are having a difficult time deciding how to use the external test results for their admissions procedures (Kim and Mizuto 2019). Notably, a few top national universities, including the University of Tokyo, announced that they would not make the external test scores obligatory for applicants (Ujioka 2018). In Japan's highly centralized educational system, it is very unusual for schools not to follow MEXT's decisions; the fact that some of the most prestigious institutions are not falling in line indicates their strong opposition to the policy. Social network sites are full of students' remarks expressing their confusion and frustration about conflicting or insufficient information about the policy. A recent survey shows that high percentages of high schools in Japan have already started providing special instruction to help students prepare for the external tests (i.e., 68.6% for Eiken tests, 58.1% for GTEC) (Shibasaki 2018), although the nature of that instruction is not known. Meanwhile, select local boards of education have started offering workshops for English teachers at public high schools to provide the educators with information on the external tests as well as instructional tips for helping their students prepare for the tests. Again, the details of such workshops are unknown, but this could be a sign of a potentially undesirable test-driven washback effect. Finally, as of January 2019, MEXT has not proposed any explicit guidelines for accommodations or special considerations for students with disabilities or special needs; instead, all such considerations are left up to the individual test agencies, whose practices differ substantially.

5.4 Insights Gained

After observing preliminary unintended fallout from the use of external tests in the college entrance exam system in Japan, we can see that a number of the fairness issues addressed above appear to originate in the very assumptions that underlie this policy. Such assumptions include that (1) English-speaking skills, as an important global competence, should be used as a gatekeeper for everybody who seeks higher education; and (2) such skills should be understood and measured uniformly against a global framework or standard such as CEFR. But what particular English-speaking skills does MEXT expect students to develop? Are these particular skills really a global competence that Japanese students need? Should we evaluate Japanese students' English-speaking skills using a global standard and, if so, should CEFR be that standard?

Communicative competence—one's ability to use language appropriately in social situations—was originally proposed by Hymes (1972) and has had a tremendous influence on language teaching and assessment. There are various models for communicative competence, but many models conceptualize it as a composition of some sort of linguistic and social/pragmatic knowledge and the ability to use such knowledge in performance. In assessment theory, communicative competence has largely been conceived of as an individual's capability that can be inferred from his or her independent performance on tasks that are representative of language use in the target domain. In assessment practice, "the ability to use" component in the original Hymes model has not been seriously discussed due to its complexity, which goes beyond linguistic elements (various cognitive, social, and affective elements are also involved) (McNamara 1996). In many standardized proficiency tests, the knowledge components in communicative competence are organized into four skill domains and assessed separately. The "appropriateness" aspect of communicative competence has largely been judged based on the performance of "native speakers." In the context of Japan, the speaking skill domain is often considered the ultimate manifestation of communicative competence (Abe 2017).

In the past decade or two, however, there has been growing interest in socio-interactional approaches to conceptualizing language abilities. In those approaches, language abilities are considered to be embedded in social contexts and constructed in fluid and dynamic interaction. The field of English-as-a-lingua-franca (ELF) challenges the very notion of native-speaker norms and questions the static view of language ability that has been conventionally accepted in the assessment community (e.g., Canagarajah 2009; Harding and McNamara 2017; Jenkins 2006). ELF's emphasis on communicative effectiveness, rather than correctness and appropriateness, highlights the role of "the ability for use" in language abilities, which presumably varies substantially in communication in people's first language as well as their second language. Reflecting such a fluid conceptualization of language abilities in assessment is not easy, especially in standardized tests (Harding and McNamara 2017), but this new conceptualization of language abilities better fits the realistic needs of Japanese students who largely interact in English-as-a-lingua-franca contexts in the globalizing world.

What MEXT promotes and tries to measure through standardized tests, therefore, are the knowledge-based components of communicative competence that are sliced into skills under the old static view of competence based on native norms, even though that is not the kind of ability Japanese students need in a globalizing world. Measuring English-speaking skills and using those measurements as a qualification for higher education is particularly problematic because these are the skills where students' SES and regional backgrounds are most likely manifested, no matter how hard test developers work to control procedural fairness issues. In a society where Japanese is used almost exclusively, students must make a special effort to create opportunities to speak English and get feedback to develop their speaking skills, and those opportunities usually require financial and regional resources. Foreign language learning is a huge and fast-growing business in Japan, with an 867-billion-yen market in 2017 (Yano Economic Research Institute 2018). Parents with higher educational backgrounds and who reside in larger cities invest significantly more in their children's English-speaking practice and do so earlier in their children's lives (Benesse General Research Center for Education 2014).

Meanwhile, the assessment community has yet to develop an unbiased strategy for capturing the kinds of language abilities needed for a globalizing world (the newly conceptualized language abilities). One may even wonder if such abilities are measurable through a standardized test. Similarly, it is not clear if they can even be evaluated and compared against some sort of universal framework (besides the fact that CEFR was not developed for such purposes in the first place). Perhaps the language abilities necessary for a globalizing world are not competencies that can be made uniform or standardized across the globe. Because such language abilities are highly context dependent, fluid, and complex, quantification based on any uniform standards or frameworks is misleading, whether or not it is done through a standardized test. Until the assessment community can come up with a fair and valid remedy, quantified evaluation of such language abilities should not be implemented for high-stakes purposes such as college admission.

5.5 Conclusion: Implications for Test Users

We have discussed Japan's new policy decision to use external English proficiency tests for college admission, and argued that the central problem is one of fairness. Based on Kane's (2010) distinction between procedural fairness and substantive fairness, we examined a series of potential issues that appear to weaken the fairness and validity of these assessments. If MEXT wants to implement this policy, it must address these procedural fairness problems. However, there remain a number of serious substantive fairness issues as well. These substantive fairness issues are difficult to solve, even if test users could gain sufficient assessment literacy, because the premise underlying the MEXT policy not only rests on a misperception of the language abilities needed for Japanese students in a globalizing world but also structurally works against students with lower SES. Without a fair and valid assessment that captures the language abilities that students really need, making high-stakes college admission decisions based on existing quantification methods is highly misleading and potentially contributes to widening socioeconomic disparities in Japan.

References

Abe, M. (2017). *Shijosaiaku-no eigoseisaku [The worst English education policy in history].* Tokyo: Hitsuji Shobo.

Benesse General Research Center for Education. (2014). *Eigo-kyoiku dainikai* [English education Part 2]. https://berd.benesse.jp/berd/data/dataclip/clip0014/index2.html. Accessed 18 January 2019.

Canagarajah, S. (2009). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 16,* 229–242.

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly, 15*(1), 59–74.

Harding, L., & McNamara, T. (2017). Language assessment. In J. Jenkins, W. Baker, M. Dewey (Eds.), *The Routledge handbook of English as a lingua franca.* https://doi.org/10.4324/9781315717173.ch45

Hato, Y. (2018). Minkanshiken-no nani-ga mondai nanoka [What is the problem with external tests?]. In T. Haebara (Ed.), *Kenho—meiso-suru eigo nyushi [Verification: Troubled college entrance examinations in English]* (pp. 41–68). Tokyo: Iwanami.

Honda, Y. (2005). *Tagenka-suru nouryoku-to nihon shakai [Diversifying competencies and Japanese society].* Tokyo: NTT Publications.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Harmondsworth: Penguin Books.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELF Journal, 60*(1), 42–50.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182.

Kariya, T. (2008). *Gakuryoku-to kaiso [Academic achievement and social class].* Tokyo: Asahi Shinbun Publications.

Kim, S., & Mizuto, K. (2019, January 7). Eigo minkanshiken katsuyo-ni annun [English external tests in trouble]. *Mainichi Newspaper*, p. 14.

Kobayashi, J., & Viswat, L. (2007). An exploratory study of "fairness" in educational settings: American and Japanese university students. *Journal of Intercultural Communication, 14*. https://www.immi.se/intercultural/nr14/kobayashi.htm. Accessed 27 July 2019.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. J. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge, UK: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA: Blackwell.

MEXT. (2017). *Koudai setsuzoku kaikaku-no shinchoku jokyo-ni tsuite* [Progress report on the reform connecting high-school and college education]. http://www.mext.go.jp/b_menu/houdou/29/05/1385793.htm. Accessed 17 January 2019.

MEXT. (2018). *Daigaku nyushi kaikaku* [College entrance examination reform]. http://www.mext.go.jp/a_menu/koutou/koudai/detail/1408564.htm. Accessed 17 January 2019.

MEXT. (2019). Daijin meseji eigo minkan shiken-ni tsuite [A message from the Minister regarding the external English tests]. https://www.mext.go.jp/a_menu/other/1422381.htm. Accessed 20 March 2020.

Miyamoto, H. (2018). Koko-kara-mita eigo nyushi kaikaku-no mondaiten [Problems with the reform of college entrance examinations from a perspective of high schools]. In T. Haebara (Ed.), *Kensho – meiso-suru eigo nyushi [Verification: Troubled college entrance examinations in English]* (pp. 26–40). Tokyo: Iwanami.

Moriguchi, C. (2017). Nihon-wa kakusya syakai-ni nattanoka? [Has Japan had a disparity in wealth?]. *Discussion Paper Series A*, No. 666. http://www.ier.hit-u.ac.jp/Common/publication/DP/DPS-A666.pdf. Accessed 17 January 2019.

OECD. (2015). *Poverty rate*. https://data.oecd.org/inequality/poverty-rate.htm. Accessed 18 January 2019.

Shibasaki, O. (2018, November 6). Daigaku nyugakusya senbatsu kaikaku-ni kansuru anketo cyosa hokokusyo [A survey report on the college entrance exam reform]. http://company.sanpou-s.net/press/pdf/181106.pdf. Accessed 18 January 2019.

Ujioka, M. (2018, September 25). *Tokyodai, eigo minkanshiken-no seiseki teisyutsu hissyutosezu*, shindaigaku nyushi [The University of Tokyo, not requiring English external test scores, the new entrance exam]. Asahi Newspaper. https://www.asahi.com/articles/ASL9T5T8HL9TUTIL031.html. Accessed 20 January 2019.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170.

Yano Economic Research Institute. (2018). *Gogaku bijinesu tettei chosa repoto* [Report on extensive research of language business]. https://www.yano.co.jp/press-release/show/press_id/2013. Accessed 18 January 2019.

Chapter 6
(Mis)Use of High-Stakes Standardized Tests for Multiple Purposes in Canada? A Call for an Evidence-Based Approach to Language Testing and Realignment of Instruction

Li-Shih Huang

Abstract Drawing on interviews from a study identifying the language-learning needs of Syrian refugees in Canada and how these relate to their integration into Canadian society, this chapter reflects on key issues related to the appropriateness of using standardized language test scores for purposes for which they were never designed. The study's follow-up interviews featured the frustrating experiences of learners in the Language Instruction for Newcomers to Canada (LINC) program in trying to reach a required band score of the IELTS standardized language test for purposes of study, immigration and citizenship, and professional certification. The literature further shows a glaring research gap regarding the IELTS's purpose, functions, and intended or unintended effects within Canada. By discussing the problems of language testing when standardized tests are used for other than their original purposes, and by sharing the perceptions and insights of instructors and learners alike, the chapter seeks to provoke critical evaluation of language testing policy and practices in order to find solutions for the issues it raises.

6.1 Introduction: Purpose and Testing Context

Under demographic, geopolitical, and academic and professional pressures, Canada's culturally and linguistically diverse population of international students, workers, and migrants continues to increase sharply. The most recent Statistics Canada's figures in 2016 highlighted Canada's increasing diversity, with 21.9% of the population consisting of immigrants, the highest in 85 years (Grenier 2017). Meanwhile, over 36 cities across Canada have been facing pressing language-training issues for such key purposes as citizenship, academics, and work (see Lowrie 2017). Yet interviews from a multi-year research in progress (Huang 2020) designed to identify the unique language-learning needs of Syrian refugees in Canada, and how these relate to their

L.-S. Huang (✉)
University of Victoria, Victoria, BC, Canada
e-mail: lshuang@uvic.ca

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_6

integration into Canadian society and the workforce, revealed the learners' frustrating experiences in trying to reach a required band score for the IELTS (International English Language Testing System) standardized language test for purposes of study, immigration and citizenship, or professional certification (Schulman 2019). These refugees were enrolled in Canada's Language Instruction for Newcomers to Canada (LINC) program, funded by Immigration, Refugees and Citizenship Canada (IRCC).

The requirement that applicants for Canadian citizenship provide objective evidence of language ability came into force in November 2012 (see King 2012). A minimum language threshold and mandatory language testing were prioritized, with the stated rationales of "making language assessment more objective, while improving language outcomes for newcomers …"; facilitating "improved employability and earnings for permanent residents by providing an incentive to enhance their language skills before applying for citizenship"; and providing "an increased pool of available employees with good language proficiency" to benefit Canadian employers (Citizenship and Immigration Canada [hereafter CIC] 2013, p. 31).

Prior to LINC, language training in the 1950s was geared toward preparing for Canadian citizenship. Starting from the 1960s, these programs shifted to training for specific skills and satisfying domestic labor market demands. In the early 1990s, the nationally recognized standard known as the Canadian Language Benchmarks, which describes English language ability for second language learners via 12 benchmarks ranging from basic to advanced, was developed (Centre for Canadian Language Benchmarks 2012) and continues to be used for curriculum development, instruction, and assessment. LINC itself was designed to help immigrants and refugees successfully integrate into Canada by providing learners basic language training in either English or French, as well as supplying newcomers with knowledge about Canada; it has thereby become an essential component of the federal immigrant integration strategy.

6.2 Testing Problems Encountered

Within Canada, the number of users required to take the IELTS test has increased sharply. The IELTS is a high-stakes, gatekeeping standardized test originally created in 1989 in response to the intake of tertiary-level international students in Australia. In Canada it is used to assess test-takers' English language proficiency for purposes of immigration,[1] education, and professional certification. In 2016 it was taken by more than 3 million test-takers across more than 140 countries; additionally, over 10,000

[1] The IRCC also accepts the Canadian English Language Proficiency Index Program (CELPIP) as an alternative proof of language ability, but discussion of CELPIP tests was not a focus of this chapter. The tests were also never mentioned in the questionnaires completed or interviews conducted with LINC instructors and learners. Also worth noting is that the CELPIP is a computer-delivered test; however, technology can present a major barrier for test-takers in relation to issues of fairness and justice in language testing McNamara and Ryan (2011). This is illuminated by the study participants' numerous comments about their aversion and strong resistance to technology use (e.g., "I told you,

organizations (e.g., universities, schools, employers, and immigration bodies) accept IELTS globally, 350 in Canada alone (IELTS Official Test Centre 2017). Because of IELTS's global reach and numerous roles, its far-reaching impacts on users for different purposes thus merit close scrutiny.

6.2.1 Test Use

According to Messick (1989), use of a test should be supported by evidence demonstrating that the ability measured is relevant to the intended decision, and that the test score is useful for making that decision. Similarly, the most widely accepted standards for test design and use state that "when a test is designed or used to serve multiple purposes, evidence of validity, reliability/precision, and fairness should be provided for each intended use" (American Educational Research Association [AERA] et al. 2014, p. 195). At issue is test validity, which refers not only to "the accuracy of score inferences, but also to evaluation of [their] appropriateness, meaningfulness, and usefulness" (Messick 1989, p. 41). The IELTS, however, is commonly used for purposes other than those for which it was originally intended (e.g., Fulcher 2018; Muller 2017), even though Ingram (2015), one of the test's original designers, indicated that its "content and design do not meet the needs of tests to assess proficiency for vocational purposes or for general survival purposes" (p. 2). As Muller (2017) pointed out, "The IELTS organisation has not officially disapproved of the use of the test beyond its original purpose. It comments on recommendation test scores for study, but is quiet on its use for migration or work purposes" (para. 15).

A central issue concerns using tests to determine the quality of performance vis-à-vis criteria that dictate what constitutes a relevant and successful target performance and the nature of the communicative demands. In other words, does the IELTS test attend to situations and tasks in which the target language will actually be used? Yet another major concern is the insufficiency of any single measure in large-scale, high-stakes tests to provide reliable evidence or to adequately cover abilities with multiple components that depend on context. Also, necessary to ask is whether the instruction received by LINC learners in Canada aligns with the standardized tests required for immigration, citizenship, or professional certification. A survey of the literature, however, shows little to no research regarding the IELTS test's appropriateness for determining Canadian citizenship and workplace readiness.

I don't think it's a good idea. I don't like it" [L003]; "... for a lot of them who've never had school let alone technology" [T007]).

6.2.2 Test Fairness and Justice

A key concern of language testing is that conclusions drawn about test-takers can lead to inaccurate inferences and, in turn, unfair decisions. The claims or inferences made about their linguistic knowledge or abilities as language users need to be fair and relevant to decisions about the candidates. As Ingram (2015) pointed out:

> The inappropriateness of IELTS for migration purposes imposes a punitive load on visa applicants, many of whom take the test many times without success. Yet many who have been tested on a test that better reflects their language background and needs have shown that they have high proficiency at a level that would enable them to perform satisfactorily in their chosen vocation. (p. 2)

As tests such as IELTS are used to control access to residency, citizenship, and employment, these issues of test fairness and justice (referring to the defensibility of the values embodied in the test) become more serious (McNamara 2018; see also McNamara and Ryan 2011 for a critical discussion of fairness and justice in language testing). It is questionable whether standardized tests such as IELTS are even capable of yielding meaningful distinctions among test-taker groups for immigration vis-à-vis their readiness for immigration, citizenship, or professional certification. Moreover, although Canada, like the United States and Australia, requires applicants for citizenship to demonstrate a specified level of competence in English, yet, "test constructs are increasingly dictated as a function of policy . . . and can be cloaked by a scientific concern for fairness" (McNamara and Ryan 2011, p. 175).

6.2.3 Test Consequences

Extensive discussion in the literature has also shown (e.g., Bachman and Palmer 2010; Moeller et al. 2016; Shohamy 2014) that high-stakes tests such as the IELTS are acting as gatekeepers, resulting in critical decisions that directly affect the lives of learners, instructors, and other stakeholders. In addition to whether a test is being used for its intended purposes, another area of test validity is so-called *consequential validity*, referring to a test's positive or negative social consequences (Messick 2003). Research into such consequences has been increasing, as evidenced by studies focusing on, for example, school principals' perceptions of IELTS vis-à-vis teachers entering the profession via the test (Murray et al. 2014); university admission staff's perceptions of IELTS as an entry requirement in the UK (e.g., Hyatt 2012); Malaysian tertiary students' perceptions of IELTS (e.g., Zahari and Dhayaalan 2016); and students' and staff members' perceptions and attitudes toward IELTS in Australian, UK, and Chinese tertiary institutions (Coleman et al. 2003). Recent media headlines also reflect the issue, such as "Tough language tests blamed for drop in EU nurses: Recruitment firm warns 'inappropriate' £150 exams have led to a 96% fall in numbers" (Borland 2017) (see also Pym 2017; Tapper 2017). Stories have also surfaced about fully qualified native-English-speaking professionals failing IELTS for immigration purposes in Australia (e.g., Collins 2015).

These concerns about the consequences of language testing and its gatekeeping function are further highlighted in the following interview excerpts drawn from the study noted earlier on the language-learning needs of Syrian refugees in Canada (Huang 2020), which involved surveys and interviews with both learners and instructors ($N = 48$) in the LINC program. The interviews revealed both learners' and instructors' perceptions of the IELTS regarding the formers' needs that diverge from the test's requirements and are negatively affected or driven by the test.

According to one instructor, "For a lot of them I think, especially the women, they're interested solely in citizenship and then they're not really, you know, they're going to be staying home" (T007). As another observed, "They want to graduate from 4 [in speaking and listening] and start working and take the test—the citizenship test" (T016). Another instructor noted in particular:

> What we're seeing because of the citizenship requiring the 4 in the speaking and the listening . . . that's what they want. They want speaking and listening. Last week I said, "We're focusing on reading and writing for the next two weeks," and some of them were just not interested in that at all. We do have class beginning, not just for Syrian students but any students for citizenship, and we're focusing on listening and speaking, and that's all they want…. (T012)

This attitude is confirmed by a learner who stated: "For the citizenship, I want to reach level 4 and not continue. بدي اكثف ساعات العمل صفى اجباري" (I want to have extra working hours. It's a must!) (L008).

For LINC learners seeking professional certification in particular, the mismatch between what they are learning and what they need to attain on the IELTS is glaring. One stated: "They did ask IELTS for us, but what I need are special English courses and classes, preparing us to the Canadian Dentist Board examination. You have five years" (L019). Another said: "I come to Canada in eleven months, I study English, what your address, what your name, what's ahh, how many have children, what name your children, how old are your son, where is this work? Not work, it's not really for test or work related" (L013). As yet another learner commented: "So first when we had to pass the first qualification exam, and the second we have to . . . what I'm saying is Canadian style or Canadian pharmaceutical science . . . we have to apply to pass IELTS 6.5 maybe, or 6" (L017). Still another noted:

قطع السيارات كلياتها، قطع السيارات لازم احفظها شو هي بالانجليزي...

(All the car parts, the car parts I have to memorize them in English) . . . I need to remember all the technical words for trucks, the inside of the truck . . . in order to pass the test" (L006). One instructor highlighted another difficulty:

> The trouble is that there are so many students who even if they have . . . like the psychiatrist for example, even if he has a psychiatry degree. . . back home, but because he is not credited here, he has no choice but to take an entry level job. I don't know of any Syrian students who are able to go right into something. Even if they have lots of language experience, they still have to pass the language test or take another degree here. (T032)

Because of the IRCC's requirement of a language proficiency threshold, which aims to encourage immigrants to take LINC courses and improve their language skills (CIC 2012), demand for these courses is acute and has led to long waiting lists (Rolfsen 2016), even as costs to deliver them have soared across Canada (see Purdy 2017). These issues have further worked against those wishing to apply for settlement or citizenship, as statements by the Syrian refugee learners and their instructors also make clear. For example, one learner commented: "When we came to Victoria, at the very first, they tested us, and I don't remember the level that I was in, like one maybe, and then they said there is no place for you, we are very sorry but the classes are completed [i.e., full], and you have to wait, so I was waiting for months" (L003). Even when classes are available, many would-be citizens have to work, often in entry level or low-paying jobs, where they find it challenging to balance work shifts and classes; according to one learner, "It's difficult for me because I have to work, I have family" (L019). The difficulties are particularly highlighted by one instructor's observation that

> they gotta work and we have a waitlist at our school and, unfortunately—so, if students are missing too much time regardless of the reason, they are removed from class and then a student that can come gets their seat. So, they struggle with this "What do I do? Do I work or come to school?" and I don't know what the answer is for them because I get it, you know? . . . They have children as well. (T007)

These issues are compounded by the lack of fit between the language courses offered and the learners' perceived needs. One instructor noted, "I think it [IELTS] measures a very particular aspect of linguistic competency, but it's not necessarily a predictor of how they'll do in the other aspects. There's so much more to it" (T105). Another stated: "We had program evaluations today, actually, and . . . it got quite heated in my class. There were students who were very upset. . . they said at this rate, I'm never gonna pass the certification or get a job" (T016).

The current language policy requiring immigrants to reach a certain threshold on the IELTS as a condition for obtaining residency, citizenship, study, or certification thus has real impacts, especially since residency and citizenship in turn determine one's access to rights and benefits related to employment, training, health, education, and welfare.

6.3 Solutions to the Problems

The overarching concern regarding the issues discussed above has no quick solution. But certain steps can be taken in finding a resolution. At a minimum, it is incumbent upon stakeholders (e.g., federal agencies, researchers, testing specialists, LINC providers, instructors, and program developers) to question how language policies have led to current practices in language assessment for migration and citizenship. Why is the specified level of proficiency in English a requirement for obtaining citizenship and a prerequisite for social integration? Is the established threshold placing an undue burden on non-English-speaking immigrants and creating a barrier to acquiring citizenship? Can individuals residing in Canada participate in their communities and work without the required level of English proficiency? How should language proficiency or skills be measured? What is the IELTS's validity for its intended purposes in the Canadian context—immigration and employment—and how should the impact of current language policies be evaluated? What is an appropriate role for language testing specialists in this discussion? How should the mediation among key stakeholders (e.g., policymakers, test developers, and practitioners) occur? These are all questions that must be raised and addressed in the search for solutions. The implications of "retrofitting" the test for multiple purposes also merits scrutiny (Fulcher and Davidson 2009).

In the meantime, LINC instructors need to figure out ways to create personal relevance by supporting the needs of learners for immigration, citizenship, or work certification—needs that clearly have been sources of dissatisfaction and frustration for learners and instructors alike. In particular, to address the stark mismatch between learners' perceptions of the LINC instruction they receive and their language-learning goals, instructors might reconsider how they approach needs assessment in order to better identify learners' target situation needs (Huang 2020, forthcoming). Through such assessment, instructors could better align their instruction with those needs while navigating both internal and external testing demands for different purposes.

6.4 Insights Gained

The testimony from the interviews with learners and instructors in the LINC program provides crucial insights into how well, or rather, how not so well, the use of the IELTS test has been meeting the needs of refugees and, by extension, other immigrants to achieve their goals regarding immigration and citizenship, study, or professional certification. Rather than helping them reach these goals, the test has instead become a hurdle to overcome. As Bachman (1990) stated nearly three decades ago: "The single most important consideration in both the development of language tests and the interpretation of their results is the purpose or purposes which the particular tests are intended to serve" (p. 55). In other words, how these tests are used lies at the very heart of language assessment. Bachman further urged stakeholders to "provide as complete evidence as possible that the tests that are used are valid indicators of the abilities of interest and that these abilities are appropriate to the intended use, and then to insist that this evidence be used in the determination of test use" (p. 285). In line with Bachman's call, research is urgently needed on the wider social and educational impact of assessment that will articulate the relation between stakeholders' perceptions and attitudes toward the IELTS and their consequences, as well as on test validity within Canada. Finding solutions to these issues can happen only through empirical test design, development, use, and validation. As Messick (1996) observed, "what is to be validated is not the test or observation device per se but rather the inferences derived from test scores *or other indicators—inferences* *about score meaning or interpretation and about the implications for action that the interpretation entails*" (p. 235, emphasis added). Engaging with stakeholders' voices and understanding their perceptions is key to evaluating those "other indicators" when assessing the social, professional, and educational impacts of the use of language tests in Canada.

6.5 Conclusion: Implications for Test Users

Drawing on both the literature on language tests for multiple purposes (e.g., Extra et al. 2009; Shohamy 2014; Shohamy and McNamara 2009) and the challenges shared by learners and instructors in the LINC program, this chapter has critically considered key issues related to the appropriateness of using standardized test scores—in this case those of the IELTS—for purposes other than those intended in their original design. In doing so, it hopes to alert readers and stakeholders to the need for critical assessment and discussion about these vital issues. The growing concerns about the IELTS and its validity and the resulting social consequences have critical implications for Canada's situation, as reflected by recent media stories (e.g., Schulman 2019). Ultimately, those of us involved in language testing must heed the call of researchers and other testing experts for ways to encourage test users to engage with the IELTS test (Murray et al. 2014) by clarifying how high-stakes standardized tests can potentially be (mis)used for problematic purposes so that we can instead develop evidence-based recommendations for test users in a rapidly changing Canadian population.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (7th ed.). Washington, DC: AERA.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Borland, S. (2017, June 13). Tough language tests blamed for drop in EU nurses: Recruitment firm warn "inappropriate" £150 exams have led to 96% fall in numbers. *Daily Mail*. http://www.dailymail.co.uk/news/article-4601928/Tough-language-tests-blamed-drop-EUnurses.html. Retrieved October 15 2019.

Centre for Canadian Language Benchmarks. (2012). Canadian language benchmarks: English as a second language for adults. https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/language-benchmarks.pdf. Retrieved October 15 2018.

Citizenship and Immigration Canada. (2012). *Departmental performance report*. http://publications.gc.ca/collections/collection_2012/cic/Ci1-15-2012-eng.pdf. Retrieved October 15 2018.

Citizenship and Immigration Canada. (2013). *Annual report to Parliament on immigration*. https://www.amssa.org/wp-content/uploads/2015/05/CIC-annual-report-2013-to-Parliament.pdf. Retrieved October 15 2018.

Coleman, D., Starfield, S., & Hagan, A. (2003). The attitudes of IELTS stakeholders: Student and staff perceptions of IELTS in Australian, UK and Chinese tertiary institutions. *IELTS Research Reports, 4,* 1–76.

Collins, P. (2015, September 27). English test derailing Irish dream of Australian citizenship. *The Irish Times*. https://www.irishtimes.com/news/world/asia-pacific/english-test-derailing-irish-dream-of-australian-citizenship-1.2368668. Retrieved October 15 2018.

Extra, G., Spotti, M., & Van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives*. London: Continuum.

Fulcher, G. (2018). *Language testing review of the year 2017*. http://languagetesting.info/features/2017/review.html. Retrieved October 15 2018.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123–133.

Grenier, E. (2017, October 25). *21.9% of Canadian are immigrants, the highest share in 85 years: StatsCan*. http://www.cbc.ca/news/politics/census-2016-immigration-1.4368970. Retrieved October 15 2018.

Huang, L.-S. (2019). Analysing open-ended questions and semi-structured interviews on language-learning needs of Syrian refugees. *Sage Research Methods Datasets*. https://doi.org/10.4135/9781526483843.

Huang, L.-S. (2020, February). *(Mis)using standardized tests for multiple purposes in Canada: A call for criticality, creativity, and collaboration in language testing and realignment of instruction [Plenary]*. BC TEAL Vancouver Island Regional Conference, Victoria, BC, Canada.

Huang, L.-S. (forthcoming). Needs analysis. In H. Mohebbi & C. Coombe, (Eds.), *Research questions in language education and applied linguistics*. New York, NY: Springer.

Hyatt, D. (2012). Stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *Journal of Further and Higher Education, 37*(6), 844–863.

IELTS Official Test Centre. (2017). https://www.ieltscanada.ca. Retrieved Feburary 20 2019.

Ingram, D. E. (2015). *Submission to the inquiry by the Productivity Commission into the migrant intake into Australia*. https://www.pc.gov.au/__data/assets/pdf_file/0011/190379/sub016-migrant-intake.pdf. Retrieved Feburary 20 2019.

King, R. L. (2012, May 24). Language skills the target of Canadian citizenship reform. *National Post*. https://nationalpost.com/news/canada/language-skills-the-target-of-canadian-citizenship-reform. Retrieved October 15 2019.

Lowrie, M. (2017, March 16). Integration still a challenge for Syrian refugees one year later: Researchers. *CTV News*. https://www.ctvnews.ca/canada/integration-still-a-challengefor-syrian-refugees-one-year-later-researchers-1.3328739. Retrieved Feburary 20 2019.

McNamara, T. (2018, January 22). *Language testing for immigration and citizenship: Issues of fairness and justice*. Meet an AAAL Scholar Webinar.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8*(2), 161–178.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education/Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241–256.

Messick, S. (2003). Test validity and the ethics of assessment. In D. N. Bersoff (Ed.), *Ethical conflicts in psychology* (4th ed., pp. 263–265). Washington, DC: American Psychological Association.

Moeller, A. J., Creswell, J. W., & Saville, N. (Eds.). (2016). *Second language assessment and mixed methods research. Studies in Language Testing 43*. Cambridge, UK: Cambridge University Press.

Muller, A. (2017, November 30). Using university language tests for migration and professional registration is problematic. *The Conversation*. https://theconversation.com/usinguniversity-language-tests-for-migration-and-professional-registration-is-problematic-87666. Retrieved Feburary 20 2019.

Murray, J. C., Cross, J. L., Cruickshank, K. (2014). Stakeholder perceptions of IELTS as a gateway to the professional workplace: The case of employers of overseas trained teachers. *IELTS Research Report Series, 1*, 1–78.

Purdy, B. (2017). Newcomers learning English call on government to stop proposed cuts to Winnipeg EAL program. *CBC News*. https://www.cbc.ca/news/canada/manitoba/newcomers-government-stop-eal-cuts-1.4009544. Retrieved Feburary 20 2019.

Pym, H. (2017, August 22). Do I have to understand jam-making to be a nurse? *BBC News*. http://www.bbc.com/news/health-41010021. Retrieved Feburary 20 2019.

Rolfsen, C. (2016, January 8). Syrian refugees in B.C. face long waits for English courses. *CBC News*. https://www.cbc.ca/news/canada/british-columbia/syrian-refugees-in-b-c-face-long-waits-for-English-courses-1.3394425. Retrieved Feburary 20 2019.

Schulman, S. (2019, February 3). Refugees hoping to become citizens face high bar to achieve language benchmarks. *The Globe and Mail*. https://www.theglobeandmail.com/canada/article-refugees-hoping-to-become-citizens-face-high-bar-to-achieve-language-2//. Retrieved Feburary 3 2019.

Shohamy, E. (2014). *The power of tests: A critical perspective on the uses of language tests*. London, UK: Routledge.

Shohamy, E., & McNamara, T. (2009). Language tests for citizenship, immigration, and asylum. *Language Assessment Quarterly, 6*(1), 1–5.

Tapper, J. (2017, June 24). Difficulty of NHS language test “worsens nurse crisis,” say recruiters. *The Guardian*. https://www.theguardian.com/society/2017/jun/24/english-speaking-ovserseas-nurses-fail-nhs-too-tough-language-test. Retrieved Feburary 20 2019.

Zahari, D. A., & Dhayaalan, J. D. J. (2016). The perception of Malaysian ESL tertiary level students on the IELTS test. *Asian Journal of Education and Training, 2*(1), 1–6.

Chapter 7
Testing in ESP: Approaches and Challenges in Aviation and Maritime English

Neil Bullock and Carolyn Westbrook

Abstract Few would doubt that proficient language in air and maritime transport communication is a de facto precursor to safety. The introduction of internationally agreed standards on the formal testing of language proficiency in aeronautical communications is clear evidence. Having such an obvious common goal in both specific purpose language domains, however, has not been without its challenges. Still prevalent are serious concerns over the validity, reliability and quality of test instruments. Research has shown a wide disconnect between those being tested and those providing the tests. Research also indicates a lack of professional experience and limited collaboration between linguists and domain specialists, resulting in testing instruments that bear little or no resemblance to the real-world communication. Certification through proficiency standards has been seen as "box ticking" for certain stakeholders. "Test tourism" where operators actively seek the easiest way to attain a certificate of proficiency, brings with it the serious issue of having personnel deemed "language proficient" without actually being in a position to communicate adequately in their field. This chapter highlights and addresses such challenges in both domains and discusses possible solutions which could offer a way forward to provide valid and appropriate testing for such safety-related communication domains.

N. Bullock (✉)
Lancaster University, Lancaster, UK
e-mail: n.bullock@lancaster.ac.uk

C. Westbrook
British Council, London, UK
e-mail: Carolyn.westbrook@britishcouncil.org

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_7

7.1 Introduction: Purpose and Testing Context

As language proficiency has taken on a prominent role for channels of communication in both the aviation and maritime industries on the back of research-based safety recommendations from international bodies in recent years, the challenges of implementing appropriate and valid testing instruments to assess minimum acceptable levels of proficiency have been widespread.

Bachman and Palmer (1996) alluded to the importance of including the Target Language Use (TLU) in tests if they were to be seen as useful. Shaw and Weir (2007, p. 17) took this further, noting that tests should be judged on how they evaluate "real-life" performance. Douglas (2000) highlighted that authenticity in specific purpose domains comes from the interaction of knowledge and language in that domain. It can thus be advocated that working together, both Subject Matter Experts (SMEs) and English Language Experts (ELEs) can learn much more about the construct and language used in the communicative process (Bullock 2015). Similarly, Knoch (2014, p. 85) posits that "close co-operation between language-testing specialists and industry professionals is crucial" in ESP test development. However, while the importance of ELE/SME collaboration is well documented, what is less well advocated, and perhaps even more critical, is that among the language specialists, the inclusion of those experienced and qualified in language testing, and not simply those with teaching experience, is primordial. Hughes' (2003) observation that much of language testing falls short of measuring what it is intended for may start to explain some of the challenges mentioned above.

7.2 Testing Problems Encountered

7.2.1 *Validity, Reliability and Quality—How Global Are the Standards?*

7.2.1.1 Aviation

Despite a rather obvious link between effective communication and safety, and recommendations from the International Civil Aviation Organisation (ICAO) that language proficiency testing should follow good testing practices (ICAO 2010), early research showed that, in the aviation world, heeding such advice was often the exception rather than the rule. Alderson (2009) posited that language test developers in the aviation world must realize the importance of constructing such tests professionally and called for clear proof that tests are useful. It is not hard to disagree when he warned that the consequences of poor-quality language tests are "potentially very serious" (Alderson 2010, p. 51).

Further evidence regarding negligible validity of language tests in this domain was supplied by Kim (2013). She noted that a group of pilots in Korea seriously doubted whether the proficiency test they were obliged to take corresponded with the real tasks and demands of their job. Huhta (2009) noted that, while a reading task was included in an aviation language proficiency test in Finland, there was a concerning lack of any relation to real-life interaction. There is no requirement to include reading tasks when assessing language proficiency for pilots and Air Traffic Controllers (ATCOs), yet interactions using radiotelephony and plain language must be included if a test is to have any useful purpose in assessing the actual proficiency of pilots and ATCOs.

One attempt to address these issues was made by ICAO, in 2011, when they set up a method of language proficiency test endorsement. Test Service Providers (TSPs) could (pay and) apply to have their test endorsed as fulfilling the recommendations of the system. Since the introduction of this service, over 20 TSPs have applied for test endorsement, yet only one test is currently endorsed. Research and evidence as to why so many tests did not succeed in the endorsement process do not seem to be available. If the concerns raised in this paper are to be rightly addressed, it would be of particular value to all stakeholders to see why certain tests had *not* been endorsed.

7.2.1.2 Maritime

In the maritime industry, the International Maritime Organisation (IMO) is the United Nations' agency with overall responsibility for shipping. It is tasked with developing Conventions to ensure global standards of safety by laying down international standards for qualifications of Masters, Officers, and Officers of the Watch on merchant vessels.

First adopted in 1978 with major revisions in 1995 and 2010, the *1978 International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW Convention)* (IMO 1978) established "basic requirements on training, certification and watchkeeping for seafarers on an international level" (IMO 2019, paragraph 1). Prior to this, individual governments set their own safety standards, which often varied across countries (IMO 2019, paragraph 1). The *STCW, as amended* (IMO 1996 Table A-II/1 p. 7) states that officers of the watch must "[u]se the Standard Marine Navigational Vocabulary as replaced by the IMO Standard Marine Communication Phrases and use English in written and oral form." However, the level required to do this is not clearly defined, leading to ambiguity as to what constitutes effective communication and resulting in safety issues on board (Noble 2017).

Despite the lack of clarity regarding proficiency levels, the *STCW Convention* dictates the content of Maritime English training in the form of *Model Course 3.17: Maritime English* (IMO 2015). This document details training content for Maritime English courses, comprising both General and Specific Maritime English. While it includes detailed lesson plans on how to *teach* the content, there are only three pages giving brief information about possible *assessment* techniques. It also states that an officer of a navigational watch should "prove to be a communicatively competent seafarer" (IMO 2015) yet there is no information about what constitutes competence in communication.

IMO Model Course 3.12: Assessment, Examination and Certification of Seafarers (IMO 2017) provides information about selecting an assessment methodology but does not appear to cover gathering evidence for test validation purposes. While advice about writing and scoring tests is no doubt useful for Maritime English teachers, test validation requires test developers to provide both a priori and a posteriori validation evidence[1] (Weir 2005) to ensure that the decisions made on the basis of the test scores are justified.

A number of TSPs offer Maritime English tests, however, with little evidence to support claims that, by passing their tests, mariners will be appropriately qualified. However, for any test results to be meaningful, test instruments and assessment scales which measure language proficiency, must be developed through internationally recognized and approved standards. As Cole and Trenkner (2008) noted:

> [a] result expressed as a number of marks out of a maximum total, or as a percentage, is simple to read but often lacks any true meaning when read by an outsider with little or no knowledge of the subject and/or the difficulties involved in achieving the result. (p. 167)

Investigating how equitable maritime training is across institutions in the UK, Singapore and the Philippines, Sampson (2003, p. 44) found that, while Maritime English training centers are required to demonstrate compliance with STCW requirements externally, this does not necessarily happen within the individual training institutions. As Sampson points out, "in the case of the IMO there is an additional pressure to relax regulations with regard to specific administrations" (p. 44), possibly due to "the importance of particular nations to the international supply of seafaring labour" (p. 44). It would appear, therefore, that there are "no recognised international or European standards for the assessment of the English Language skills of seafarers" (MarTEL n.d., para. 1).

In an attempt to provide the Maritime English community with a framework for assessing Maritime English, Cole and Trenkner (2008) proposed a "Yardstick for Maritime English," offering benchmark criteria in line with IMO STCW requirements. While this has become the internationally accepted benchmark for Maritime English, it has not been officially recognized by organizations such as the IMO. Although clearly a step in the right direction, there are no benchmark exemplars of what constitutes performance at a given level, thereby making it difficult to link test performances to the Yardstick.

Given the above, it is perhaps not surprising that validation evidence for the different tests of Maritime English is not easy to source.

[1] See Weir (2005) for detailed information regarding the type of validation evidence required.

7.2.2 Test Takers and Test Developers—the Great Disconnect

7.2.2.1 Aviation

If real expertise in good testing practice is so clearly lacking, one can hardly be surprised if acceptance by those to be tested (face validity) is difficult to achieve.

In April 2017 a pre-conference survey conducted by the International Civil Aviation English Association (ICAEA) among conference delegates showed notable disparities in the perceptions of how successful or not the Language Proficiency Requirements (LPRs) system had been in addressing the issues of learning and testing to improve aeronautical communications (Bullock and Kay 2017). Participants represented a cross-section of the industry including pilots, ATCOs, language trainers, test developers, legislators and Air Navigation Service Providers (ANSPs).

The survey was organized to source opinions in the 10 years since the testing system was set up by ICAO. It included 22 questions divided into four key themes. Responses were taken from a 5-part Likert scale, which ranged from "completely agree" with a given statement to "completely disagree." The participation rate was 70%, which was seen as very encouraging.

Of particular interest to this paper is the disparity shown between testing and training service providers and those being trained or tested. If we look at the statement: "ICAO LPR language tests in your region adequately assess the communication needs of pilots and controllers in air-ground communication contexts," only 25% of non-native-English-speaking pilots and ATCOs, i.e., those likely to be affected the most by the LPRs, agreed with the statement, whereas 50% disagreed. This response seems to indicate that those actually being tested do not believe that test developers and trainers are doing a good job, which may well support earlier criticisms of the system (Alderson 2009; Kim and Elder 2009).

If test instruments therefore demonstrate only a limited connection with target language use, and little tangible learning of real-world communication skills is actually taking place as a result, further work is clearly needed to address this disconnect.

7.2.2.2 Maritime

Maritime English testing suffers similar problems as Aviation English testing in terms of a lack of construct validity. In addition to the need for test tasks to reflect the real-life domain, test takers should engage in the same cognitive processes as would be expected of them in a real-world situation (Weir 2005). Some Maritime English tests only include discrete multiple-choice items, while others include a variety of task types but do not test productive skills.

The SeaTALK project (SeaTALK n.d.), a three-year, European Union-funded project which ran from 2012 to 2015, aimed to create online Maritime English training materials. As the starting point for the project, the project team carried out a survey among maritime institutions to "collect information concerning the current Maritime English language training courses offered at Maritime Universities /Institutions /Training centres across Europe" (Toncheva and Zlateva 2014, p. 204). The survey attracted responses from 24 participants from 17 countries around Europe. One area that the survey investigated was the range of Maritime English assessments used. The findings revealed that assessment was very varied, including formative, summative and continuous assessment, with the frequency of assessment ranging from "every lesson to once a … semester" (p. 212). Regarding the tests used:

> [t]he majority of respondents state that tests are usually teacher-developed and are thus exclusive to the institution in question. Only two respondents refer to commercial tests being used, namely Headway and TOEIC. (p. 212)

Given that these tests are the ones used to "certify" the cadets as having the required level of competence in Maritime English, it is disturbing that many of the tests are written by teachers who, at least for the most part, may only have a limited understanding of testing theory. In addition, the fact that the tests are exclusive to the institutions in question, not only makes transparency questionable, but also makes comparison with, and adherence to, a worldwide standard rather difficult. On the other hand, the use of Headway and TOEIC is also dubious, given that these are not tests of Maritime English.

7.2.3 A Non-Collaborative Approach to ESP Testing—More Than just Language

7.2.3.1 Aviation

The ICAO documentation for the Language Proficiency Requirements (Doc9835) (ICAO 2010) shows numerous ambiguities in the manual and an overemphasis on plain language in isolation as the conduit by which communication is facilitated. Kim (2018) wonders whether in fact the interactive skills of the ICAO LPRs (*Pronunciation*, *Comprehension* and *Interactions*) are actually more significant than the linguistic skills (*Fluency*, *Vocabulary* and *Structure*), and she questions the fairness and validity of focussing on linguistic factors alone.

The ICAO documentation lacks real-world concrete examples of pilot/controller communication, with examples restricted largely to listed grammatical structures and language functions devoid of real contextual references. Breul (2013) discusses how such specific purpose communication very often relies on close referential language with common understanding operating within a dedicated speech community. This may well produce the paradox of creating elliptical language, where, to someone unfamiliar with the complexities of such communication, it may appear that more explicit meaning is necessary.

The combination of transaction and interaction in such spoken communication was underlined by Harmer (2007), while Hedge (2000) extended discourse interest to pragmatic and strategic skills to effect appropriate communication, rather than a sole reliance on discrete lexical and structural forms. If testing is to be valid and useful, then inclusion of the many communicative elements that make up the target language of both the pilot and the controller must be included when testing language proficiency in aviation.

Spoken communication in any domain is multi-disciplinary, based on a combination of context, knowledge and socio-cultural influences (Bullock 2018; Fan et al. 2015). Language is not produced in isolation, but is fundamentally determined by multiple factors working interdependently, which are, as Bullock (2018) suggests, both *manageable* (internal) and *influential* (external).

7.2.3.2 Maritime

Widdowson (1978) distinguishes between "usage" and "use," stating that the former relates to one's ability to produce grammatically correct sentences while the latter refers to the ability to apply these rules in order to communicate effectively.

Yongliang (2015, p. 315) makes the important point that being competent in language structure does not equate to communicative competence and that there is often a sizable gap between linguistic competence and communicative competence. The concept of communicative competence[2] has been addressed by a number of researchers over the years (Canale and Swain 1980; Bachman and Palmer 1996). Communicative competence requires a combination of linguistic, sociolinguistic, discourse and pragmatic competences (Canale and Swain 1980) in order to use the appropriate language functions in a given context (Zhang 2018, p. 28).

For Maritime English, the context is that of a seafarer, for example, a rating, an engineer, a deck officer or a master. Yongliang (2015) refers to the 2010 Manila amendments of the *STCW Convention*, pointing out that prior to these amendments, seafarers were required to "use English in oral and written form" (IMO 1996 Table A-II/1, p 7) but the 2010 amendments are now more clearly linked to the duties of a deck officer, requiring them to "use charts and other nautical publications, understand meteorological information and message in concern (sic), and perform the officer's duties in English" (*STCW, as amended* in Yongliang 2015, p. 313). Thus, officers need to have not only linguistic competence, but also communicative competence.

Without the inclusion of both language and content knowledge tasks (Douglas 2000), it is possible that one would be reduced to testing General English in a maritime context, which may meet the General Maritime English requirements of Core Sect. 1 of *Model Course 3.17*, but is unlikely to test Specific Maritime English as required by Core Sect. 2 of *Model Course 3.17* (IMO 2015).

[2] For an accessible, brief overview of models of communicative competence, see Zhang (2018), Chapter 2.

7.3 Solution/Resolution of the Problem

7.3.1 A Collaborative Effort—Communication in a Very Specific Domain

7.3.1.1 Aviation and Maritime

As explained in the previous section, it is clear that a system based solely on linguistic competence provides challenges in terms of validating whether a test can be defined as useful in assessing a required communicative competence in specific purpose domains.

Additionally, the perceived lack of inclusion of experienced testers in the test development teams is an issue. As with any test of languages for specific purposes (LSP), it is essential a test designer does not work alone. ALTE (2018, p. 6) warns that, while some LSP teachers may have some insights into the TLU domain, they should not be the only people involved in a test development project. Instead, the development of an LSP test is very much a collaborative effort between testing experts, LSP teachers and SMEs. In the aviation and maritime environment an SME would typically be an Air Traffic Controller, or a ship's captain. In addition, test developers must enlist the help of all appropriate stakeholder groups (ALTE 2018, p. 5) to find out not only what the day-to-day tasks of the job entail but to ensure that the construct of the test is representative of the real-world communication (p. 5).

As with all test design projects, a detailed item review process will identify any unsuitable items; however, collecting post-test feedback from test takers provides valuable insights into areas for improvement and contributes to the test's validation evidence.

It is therefore a requirement for all stakeholders to assist in conceptualizing and contextualizing authentic real-world communicative tasks, including "technical vocabulary" (Paltridge and Starfield 2013, p. 117) and "background knowledge" (Douglas 2000, p. 39). These both contribute to content and context validity allowing the language tester to apply recognized theoretical principles which will underpin the test. It can therefore be shown that appropriate foundations have been laid with which to make an assessment of the required communicative skills of test takers. This, in turn, will contribute greatly in allowing valid interpretations of the test scores to be made by stakeholders.

7.4 Insights Gained

Alderson (2011, p. 394) noted that certain claims about what available tests do fall well short of what could be described as "professional scrutiny and is, in our view, irresponsible." He also drew attention to the fact that Test Service Providers were making claims about using linguistic and operational experts, but failing to provide any evidence of such. Knoch (2014) suggests that studying the criteria that experienced professionals use in the relevant field when evaluating communicative skills "adds to the validity of the resulting assessment criteria as they will more closely reflect norms expected in the workplace" (p. 78). Perhaps most worrying is Knoch's (p. 78) observation that "very few studies have employed industry professionals as informants for post hoc validation of … the linguistic criteria of an LSP rating scale."

7.5 Conclusion: Implications for Test Users

More than ten years after language proficiency testing in aviation was introduced, challenges are still evident, but the foresight of some stakeholders means that a new appreciation of the real-world communicative tasks is starting to prevail over linguistic skills in isolation. ICAEA is also looking at further projects in the near future to help ensure the LPR system evolves and functions both appropriately and positively, while addressing the concerns of all stakeholders. The Association aims to explore new thoughts and ideas, focussing on air-ground communications between pilots and ATCOs in the real world and to address more of the emerging issues in a way that supports and reinforces the system of the LPRs. The Association is supported by its own research group which offers a theoretical approach toward broader communicative competence based on the facets of real-world communication as well as looking at how this can be built into learning pedagogy and assessment instruments.

In the maritime field, steps are being taken to improve Maritime English proficiency through amendments to the *STCW Convention* and Model Course 3.17. However, much still remains to be done to ensure that the tests used by maritime institutions and the shipping industry demonstrate validity and quality. Tests and performances at different levels must be linked to the Yardstick of Maritime English (Cole and Trenkner 2008) through a comprehensive linking process for ease of comparison across different maritime educational and professional contexts in the same way that General English tests are linked to the CEFR. This requires a series of validated exemplars of performance at the different levels that can be used to judge test performances and, ultimately, the communicative competence of seafarers in the English language.

We would therefore conclude by advocating that, not only are such ongoing projects crucially valid in helping to evolve LSP test design in the two domains, but that successful ESP test design projects are the result of close collaboration within a triumvirate of core stakeholders that includes SMEs, LSP teachers and language testing experts.

References

Alderson, C. J. (2009). Air safety, language assessment policy, and policy implementation: The case of aviation English. *Annual Review of Applied Linguistics, 29,* 168–187.

Alderson, C. J. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51–72.

Alderson, C. J. (2011). The politics of aviation English testing. *Language Assessment Quarterly, 8*(4), 386–403.

ALTE. (2018). *Guidelines for the development of languages for specific purposes tests—A supplement to the manual for language test development and examining.* https://www.alte.org/resources/Documents/6093%20LSP%20Supplement%20-%20WEB.pdf. Accessed 11 March 2019.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Breul, K. (2013). Language in aviation: The relevance of linguistics and relevance theory. *LSP Journal, 4*(1), 71–86.

Bullock, N. (2015). Wider considerations in teaching speaking of English in the context of aeronautical communications. *IATEFL ESPSIG Journal, 45,* 4–11.

Bullock, N. (2018). Evolving teacher training programmes through integrating contextual factors for language learning as part of aeronautical communication. *Conference proceedings from ICAEA Conference*, ERAU, Daytona Beach, 9–11 May 2018.

Bullock, N., & Kay, M. (2017). *Reviewing 10 years of the ICAO LPRs.* Paper presented at the ICAEA Annual Conference, Dubrovnik, 24–25 April 2017.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1–47.

Cole, C., & Trenkner, P. (2008). The yardstick for maritime English STCW assessment purposes. *IMLA2008.* http://web.deu.edu.tr/maritime/imla2008/Papers/19.pdf. Accessed 11 March 2019.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Fan, S. P., Liberman, Z., Keysar, B., & Kinzler, K. D. (2015). The exposure advantage: Early exposure to a multilingual environment promotes effective communication. *Psychological Science, 26*(7), 1090–1097.

Harmer, J. (2007). *The practice of English language teaching*. Harlow: Pearson Longman.

Hedge, T. (2000). *Teaching and learning in the language classroom.* Oxford: Oxford University Press.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Huhta, A. (2009). An analysis of the quality of English testing for aviation purposes in Finland. *Australian Review of Applied Linguistics*, 32 (3), 26.1-26.14.

International Civil Aviation Organisation. (2010). *Doc 9835, Manual on the implementation of ICAO language proficiency requirements* (2nd ed.). Montreal: ICAO.

International Maritime Organisation. (1978). *International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW Convention).* London: IMO.

International Maritime Organisation. (1996). *STCW 95: International Convention on Standards of Training, Certification, and Watchkeeping for Seafarers, 1978, as amended in 1995*. London: IMO.

International Maritime Organisation. (2015). *IMO Model Course 3.17: Maritime English* (2015 ed.). London: IMO.

International Maritime Organisation. (2017). *IMO Model Course 3.12: Assessment, examination and certification of seafarers.* London: IMO.

International Maritime Organisation. (2019). International Convention on Standards of Training, Certification and Watchkeeping for Seafarers. *International Maritime Organisation.* http://www.imo.org/en/About/Conventions/ListOfConventions/Pages/International-Convention-on-Standards-of-Training,-Certification-and-Watchkeeping-for-Seafarers-(STCW).aspx. Accessed 05 March 2019.

Kim, H. (2013). Exploring the construct of radiotelephony communication: A critique of the ICAO English testing policy from the perspective of Korean aviation experts. *Papers in Language Testing and Assessment, 2*(2), 103–110.

Kim, H. (2018). What constitutes professional communication in aviation: Is language proficiency enough for testing purposes? *Language Testing, 35*(3), 403–426.

Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca—Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics*, 32 (3), 23.1-23.17.

Knoch, U. (2014). Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes, 33,* 77–86.

MarTEL. (n.d.). Why MARTEL? *MarTEL Martime Tests of English Language.* http://www.martel.pro/. Accessed 05 March 2019.

Noble, A. (2017). *Maritime English put to the test! The feasibility and desirability of setting global standards for Maritime English: A survey-based study.* (Doctoral dissertation.) Antwerp, Belgium: UPA University Press Antwerp.

Paltridge, B., & Starfield, S. (2013). *The handbook of English for specific purposes*. Chichester: Wiley-Blackwell.

Sampson, H. (2003). Unequal training in an unequal world? An examination of global MET standards. Proceedings of SIRC's Third Symposium, Cardiff University, Cardiff, 19 September 2003.

SeaTALK. (n.d.). Executive summary SeaTALK survey report. http://seatalk.pro/images/reports/D5_Executive_Summary_Survey_Analysis_FinalV_131127_AMA.pdf. Accessed 08 March 2019.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Toncheva, S., & Zlateva, D. (2014). The SeaTALK project survey of maritime English—Current practices and challenges for the future. Proceedings of IMEC 26. Maritime Institute Willem Barentsz, Terschelling, Netherlands, 7-10 July 2014.

Weir, C. J. (2005). *Language testing and validation.* Basingstoke, UK: Palgrave Macmillan.

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.

Yongliang, C. (2015). From general maritime English to specific maritime English—Some thought (sic) on the revision to the 2009 edition of the IMO Maritime English Model Course 3.17. *Journal of Shipping and Ocean Engineering,* 5, 312–317. https://doi.org/10.17265/2159-5879/2015.06.004.

Zhang, L. (2018). *Metacognitive and cognitive strategy use in reading comprehension—A structural equation modelling approach.* http://www.springer.com/978-981-10-6324-4. Accessed 11 March 2019.

Chapter 8
A Conceptual Framework on the Power of Language Tests as Social Practice

Tuçe Öztürk Karataş and Zuhal Okan

Abstract In most societies, high stakes language tests as widespread practices symbolize success and achievement and generate power over individuals and society at large. However, in most cases, these tests are believed by the public to be independent of social, economic, and political contexts and the quality of the tests has been identified and ensured by their psychometric features. However, more recently, the dimensions of tests as social practice with an emphasis on their power and use have increasingly been emphasized. Here, in this chapter, we propose a conceptual framework on the power of tests as social practice. To do this, a scoping (A scoping review aims "to map rapidly the key concepts underpinning a research area and the main sources and types of evidence available, and can be undertaken as standalone projects in their own right, especially where an area is complex or has not been reviewed comprehensively before" (Mays et al. 2001, p. 194).) review of the literature on "the uses, meaning, roles, effects, consequences, contexts, discourses and power of assessment/testing/tests" was conducted. A total of 60 theoretical and empirical publications were inductively analyzed which produced four main themes: (1) the roles of testers, (2) the meaning of tests in public, (3) the feelings and meaning tests evoke in test takers, and (4) the functions of tests. Building upon these themes and the relevant literature, this study concludes with the theoretical contributions of the framework and implications for language testing policies and practices, and critical language testing literacy.

T. Öztürk Karataş (✉)
Mersin University, Mersin, Turkey
e-mail: ecut14@gmail.com

Z. Okan
Çukurova University, Adana, Turkey
e-mail: zuhalokan2@gmail.com

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_8

8.1 Introduction: Purpose and Testing Context

In most educational policy developments, high stakes language tests have become the focus of political, economic, social, and cultural expectations for change. Very often a new high stakes test is introduced to the system of education in order to bring about a change in language teaching policy (Luxia 2005; Shohamy et al. 1996). In fact, it has been commonly assumed that high stakes tests could exert a desirable influence on educational practices and thus promote curricular and pedagogical reform (Luxia 2005; Madaus 1988). However, it should be noted that the use of those tests inevitably generates power in a testing context, particularly when an educational system is based on those tests (Menken 2017).

Today, ample research addresses the use and power of tests not only in the contexts of asylum seekers and immigrants, in citizenship processes but also as tools for language and education policies in social and educational contexts. These publications offer valuable insights into social dimensions and the power of tests. One issue all these publications agree on is that tests do exercise power due to their high stakes effects (McNamara 2008). What is not certain is where test power originates and which factors in a specific testing context cause it.

Thus, the purpose of this chapter is to present, through a comprehensive review of the language testing literature, key concepts and approaches to the power of language tests as social practice. Therefore, the focus is on the following:

- identifying what has been exactly stated so far in the language testing and assessment literature on what the power of tests is and identifying the key concepts underpinning this power
- presenting a conceptual framework on the power of language tests as social practice by introducing what is meant by "social practice" and "tests as social practice"
- contributing to the development of testing literacy of language test users
- leading test users to look at language testing practices from a more critical perspective.

8.2 Testing Problem Encountered

Looking back over the last 40 years, it is seen that language testing practices are introduced in a top-down manner. Their social, political, economic consequences on individuals and society at large are seldom taken into consideration (McNamara 2009b). Menken (2008b) agrees and states that because most of the consequences of the tests are implicit in nature, the quality of those practices is identified and ensured by the contribution of their psychometric features, rather than careful consideration of their uses and functions in the society.

However, the current critical view of high stakes language tests emphasizes not only investigating their effects and consequences from a critical and wider perspective but also understanding the power they exercise in shaping changes at social, economic, and political levels (Shohamy 2001b, 2007a, 2008b). It seems obvious that a mere focus on psychometric features of a test would not yield such an understanding. As McNamara and Roever (2006) explain, "a psychometrically good test is not necessarily a socially good test" (p. 2).

Now there seems to be no doubt that the social dimensions of high stakes language tests with an emphasis on their power must be considered when their quality and effectiveness are concerned. However, the debate still seems to be far from reaching an agreement on which factors and dimensions might be influential in the generation of the power these tests exercise.

The present chapter attempts to contribute to this debate. It first presents a scoping review[1] of the uses, meaning, roles, effects, consequences, contexts, discourses and power of language tests in the current published literature. Then, it proposes a conceptual framework for the power of these tests as social practice, highlighting the social dimensions of language tests.

8.3 Review of the Literature

The past fifteen years have seen a growing awareness that language testing can move beyond concerns over reliability and validity issues. The recent critical lens identifies and examines the wider effects and consequences of tests in their own educational and social contexts (Taylor 2005). This line of research associates language tests with their own educational, political, economic, and social contexts while evaluating their uses and power (see McNamara 2005, 2008; McNamara and Roever 2006; Menken 2008b; Shohamy 2001b).

In early studies, the social nature and use of language tests found a space within the perspective of some validity theories (McNamara and Roever 2006). For example, while Cronbach (1988, 1989) does not directly acknowledge the importance of the social nature of testing practices, he foregrounds the roles of values and beliefs in test construction in terms of validity. Messick (1989), on the other hand, explicitly places social consequences of test use in his validity arguments. McNamara (2008) and McNamara and Roever (2006), however, criticize these earlier studies as they relied on validity theories when they foreground social aspects engrained in testing. They believe that the wider social uses and functions of language tests placed in social contexts are not adequately conceptualized. As McNamara and Roever (2006) state, "language testing research has a chance to move beyond the limits of validity

[1] A scoping review aims "to map rapidly the key concepts underpinning a research area and the main sources and types of evidence available, and can be undertaken as stand-alone projects in their own right, especially where an area is complex or has not been reviewed comprehensively before" (Mays et al. 2001, p. 194).

theory and make a proper contribution to the wider discussion of the general and specific functions of tests in contemporary society" (p. 40).

Therefore, what is required is "a theory of the social context in which tests have their function" which examines language testing practices as social phenomena (McNamara and Roever 2006, p. 149). Kunnan's (2004, 2005b) principles of test fairness framework is one of the first attempts in the literature. His model includes five test qualities: validity, absence of bias, access, administration, and social consequences. He has associated the social consequences of language tests with test fairness. The second framework has been developed by Shohamy. Based on the principles of Critical Language Testing, she (2001b) argues that the uses of tests and their consequences and roles in education and society should be questioned and examined critically. Another attempt can be traced in Lynch's (2001) study. Like Shohamy, he proposes some principles dealing with unfairness and unjust uses of tests in society. Bachman (2005) has also dealt with the issue of functions of tests in light of a code of ethics and a code of practice (McNamara and Roever 2006). We should also note Filer's (2000) model of "Discourse of Assessment" in which she deals with two broadly distinct discourses—technical and sociological discourse of assessment. Built on these theoretical attempts, McNamara and Roever (2006) criticize the tradition of highlighting the psychometric features of tests, but emphasize the social aspects and use of tests in language testing.

8.4 Methodology

As stated above, a scoping review of the language testing literature forms the backbone of this study. Based on Arksey and O'Malley's (2005) methodology for a scoping review, this study followed these stages as given below:

Stage 1: identifying the research question
Stage 2: identifying relevant studies
Stage 3: study selection
Stage 4: charting the data
Stage 5: collating, summarizing, and reporting the results

Thus, prior to conducting the literature review, one guiding question had to be established before deciding on the inclusion criteria: What is known from the existing literature on language testing and assessment about the uses, meaning, roles, effects, consequences, contexts, discourses, and power of assessment/testing/test? In line with this question, included in the review were publications addressing the following issues:

- the uses, meaning, roles, effects, or consequences of assessment
- the social, political, and economic contexts of assessment
- the issue of power of assessment
- the discourse of tests or discursive effects of assessment
- the issue of assessment and testing from critical perspectives.

After determining our priorities for the review, we searched via the following:

- Google Scholar
- Taylor& Francis Online Journals
- Sage Journals Online
- Eric (EBSCO host)
- EBSCO e-Book Academic Collection
- Springer Link
- Science Direct.

Also checked were the reference lists of the publications found through our reviews. Additionally, we identified two key journals which required hand-searching: *Language Testing* and *Language Assessment Quarterly*. All searches conducted in the scope of this study were limited to papers written in English.

During the process of selecting publications on test impact, articles using the term *washback* in particular, and testing and assessment practices from validity perspectives were excluded mainly because they tend to disregard the social dimensions of consequences, functions, and uses of tests (McNamara 2008; McNamara and Roever 2006; Shohamy 2017).

A total of 60 publications in the literature of language testing and assessment met the inclusion criteria. Each publication included in our review focused on the use, discourse, social dimensions, and power of test/testing/assessment. The selected publications included both theoretical and empirical accounts in the formats of original research articles, books, book reviews, and book chapters published between 1990 and 2017. The length of the texts was generally between 3 and 300 pages.

The content of the publications was analyzed in three steps, which acted as a successive filter to condense the text information. Predefined categories were not used in analyzing the identified publications, but rather an inductive or bottom-up approach in which categories emerged from the data via the following process:

- Selection of information. Each publication was divided into paragraphs as context units. This allowed for the selection of text fragments that explicitly addressed a topic or theme relevant to the power of tests.
- Segmentation of units into propositions/items. Context units were divided into their constituent propositions. Each proposition as a code unit covered a single subject–predicate relationship.
- Grouping code units into themes. Items predicating the same content were combined under the same theme.

Through the analysis, four main themes appeared on “power of tests”: (1) the roles of testers (RT), (2) the meaning of tests in public (MTP), (3) The feelings and meaning test evokes in test takers (FM), and (4) the functions of tests (FT).

8.5 Findings: Key Concepts Generating the Power of Tests

As language tests become a feature of the social lives of individuals, they and their results are used to make judgments about people, knowledge, values, ideas, and languages. Inevitably, such a social role might result in making tests as "the instruments of power" related to political, educational, and social domains rather than just as tools assessing knowledge, skills, or progress (Shohamy 2001a, b). Although language tests are scientifically verified as neutral and objective, they might exercise power and control (Shohamy 2008a, b; Menken 2008b).

The question "Is the test powerful in this testing discourse?" requires an evaluation of the dimensions or the factors of what would constitute "test power." The review of literature on language testing and assessment provides four key concepts related to test power. Table 8.1 shows the dominant themes emphasized in each study. They are concerned with four dimensions that generate test power: the roles of testers, the meaning of tests in public, the feelings and meaning test evokes in test takers, and the functions of tests.

Table 8.1 The Publications analyzed for the scoping review

The Names of the Authors/Year	Themes
Bourdieu 1991; Broadfoot 1996; Douglas 2010; Filer 2000; Filer & Pollard 2000; Foucault 1995; Hamp-Lyons 2007; McNamara and Roever 2006; Menken 2008a; Shohamy 1993, 1997, 1998, 2001a, 2001b, 2005, 2006, 2007c, 2009, 2013; Shohamy, Donitsa-Schmidt and Ferman, 1996	FT,* RT,** FM,*** MTP****
Bachman 2005; Broad 2007; Davidson 2002; Lazaraton 2010; Matsugu 2011; McNamara 2009a; McNamara and Shohamy 2008; Shohamy 2001c, 2007b; Young 2012	FT, RT, FM
Brindley 2008; Hamp-Lyons 2000; Lynch 2001	FT, RT, MTP
Brown and McNamara 2004; Davies 2008, 2012; ILTA 2007; McNamara 2005, 2008; Moder and Halleck 2012; Shohamy 2007a, 2008a, 2008b; Shohamy and Menken 2015; Spolsky 2008a, 2012; Xi and Davis 2016	FT, RT
Cheng 2008; Coniam and Falvey 2007	FT, FM
Hudson 2012; Kunnan 2005a, 2008; McNamara 2001; Menken 2008b, 2017; Ross 2011; Shohamy 1990, 2017; Spolsky 2008b	FT
ILTA 2000	RT

*RT: the roles of testers
**MTP: the meaning of tests in public
***FM: the feelings and meaning test evokes in test takers, and
****FT: the functions of tests

8.5.1 The Roles of Testers

Language testers include "all those who take part in the decisions and actions to make the testing event, or the testing experience, happen… 'A tester' is all those who have made some contribution to the act of testing" (Shohamy 2001b, p. 145). In this respect, of all 60 publications reviewed, 48 include a depiction of how they conceptualize *the roles of the testers* in a testing discourse in relation to both professional and social responsibilities of testers.

Developing high quality language tests for practical purposes has attracted attention since the 1980s (Saville 2012). The usual practice has been that testers set professional standards and determine the quality of tests in terms of their rhetoric and psychometric features. Apparently, setting standards is a technical issue, but in reality it is related to political and ethical stances (Davies 2017). It is because tests are often introduced by powerful organizations attempting to manipulate and control educational systems according to set agendas (Shohamy 2007a, b). These organizations hold the power of making the decisions (what to introduce as tests, what to test, how to score, how to test, how to deliver and how to interpret the results) and have a key role in the development of the power of tests. When granted such power, tests and testers can control directly or indirectly the knowledge and behaviors of test takers and society at large (Bourdieu 1991; Shohamy 2001b). Thus, with the growth of the testing industry in the world, as McNamara and Roever (2006) stress. "the importance of well-rounded training for language testers that goes beyond applied psychometrics…. includes a critical view of testing and social consequences, whether those effects concern the educational sector or society at large" (p. 255). Related to this concern, ILTA (2000) and ILTA (2007), indicate that testers have produced guidelines and codes for test ethics in order to link ethics and professionalism (Davies 2017). Additionally, the roles and responsibilities of testers have also been conceptualized (Xi and Davis 2016).

This scoping review found several studies on the social responsibilities of testers for the development of a democratic testing discourse. They demand that testers investigate the quality of tests in relation to societal, political, economic, ideological, and educational consequences that they bring about for individuals and society at large (ILTA 2000; Shohamy 2001a, b, c, 2013).

The crucial point to be raised here is whether testers are authoritative or collaborative agents in a particular testing context. In cases of authoritative testers, they determine all rules of tests which are often introduced in a top-down manner. They make the important decisions about the administration of tests and the knowledge included in them. In such a testing discourse, stakeholders other than testers have no right to question the test results and methods. Additionally, powerful testers determine the qualities of tests which are compulsory in most cases by their psychometric features. Such power testers hold, as stated by Shohamy (2001b) is "one important feature that grants tests power" (p. 20).

Collaborative testers, on the other hand, choose to share their power in the testing discourse with other stakeholders. They pay attention to the development of democratic testing contexts by sharing authority, responsibilities, knowledge, and values. This scoping review reveals the social roles and responsibilities of testers for the development of shared power relations in a democratic testing discourse as follows:

- stating the test's intended purposes explicitly
- releasing information about tests (the extent of the information about the test, the format, the questions, the content, the context, or its purpose)
- monitoring the uses of tests, their effects, and consequences
- following the ethical and fair principles of their own professions
- providing equal opportunities and treatment to all test takers
- following principles of shared power
- protecting the personal rights of all test takers
- forbidding the misuse of tests
- improving the quality of language testing considering their social political, economic, ideological, and educational effects and consequences
- providing test takers with meaningful feedback that can be constructive in improving their learning.

8.5.2 The Meaning of Tests in Public

This scoping review is suggestive of the crucial roles of the stakeholders other than testers and test takers such as teachers, families, etc. Of the reviewed publications, 23 highlight that the way community members perceive language tests contributes to the generation of the power of tests in a testing discourse.

Attaching high stakes to tests, those who introduce tests know that the far reaching consequences of tests are to influence and then shape individuals and society (Broadfoot 1996). It is because "this is the easiest and quickest way for policy makers to demonstrate action and authority" (Shohamy 2007c, p. 528). They believe that individuals will comply with the demands of tests by changing their behavior due to the power tests exercise (Shohamy 2001b, 2007c). For Bourdieu (1991), not only symbolic power structures but also members of society tend to have a role in legitimizing the power exercised in social systems. This means stakeholders other than testers and test takers also contribute to extending test power in the society. As Shohamy (2007c) explains,

> if English language tests use specific criteria for correctness it is obvious that in high stake situations, these criteria become the very criteria used as part of the teaching and learning English in schools… This decision is often a reaction to public or media demands for action, but in the case of English language teaching, demand appears unlimited. Parents judge success of schools by the proficiency their children attain in the English language. (p. 528)

It appears that the influential nature of tests might stretch beyond expectations of testers due to the complex variables included in testing contexts.

A further problem is that most community members are not academically equipped to achieve a better understanding of the complexities of tests. That is why "the introduction of tests has a strong appeal to the public as it symbolizes social order in areas in which the public normally feels a lack of control, such as education" (Shohamy 2001b, p. 39). When tests are used by test users as evidence for their high stakes decisions, community members attach blind trust to them and rarely raise objection as they have no rights or tools to examine or even question the quality of the tests. As Spolsky (1998) explains, "For much of this century, the general public has been brain-washed to believe in the infallibility, fairness and meaningfulness of the results of tests and examination" (p. 1). In fact, once tests achieve such an acceptance level in the community, the knowledge, identities, and values included in tests are also popularized in societies. In other words, examinations imply the ideal of how an individual in the society should behave (Foucault 1995). When community members expect individuals to meet what is involved in tests, they contribute to the development of test power. Shohamy (2009) states that "there is an unwritten contract between those in power who want to dominate and those who are subjected to the tests in an effort to perpetuate and maintain existing social order" (p. 50). Whatever the intentions of a test are, test users "interpret it as prescriptive, so it is creating a single system in actuality" (Menken 2008b, p. 405). Thus, testers are not the only ones who create the power of tests. How individuals see themselves against a test in a particular testing discourse is highly related to test power. This review of the language testing literature revealed that tests turn into powerful tools when community members perceive tests as signs of the following:

- a serious and meaningful attitude toward education
- policies of schools
- their perceptions about individuals
- single source of knowledge
- gaining social qualifications and identities
- their expectations and evaluations that they form of test takers
- their educational and social order
- discipline
- quality of education
- a single source of knowledge
- their societal values
- their direction, guidance, and social habits
- the main criterion of worth.

8.5.3 Feelings and Meaning a Test Evokes in Test Takers

From the literature reviewed, *the feelings and meaning language tests evoke in test takers* appeared to be another factor within the scope of test power. A total of 32 publications touches upon the significance of listening to the voices of test takers so as to understand the use and power of tests.

Tests take their meaning for test takers via socially and economically valued resources embedded in them. Due to such resources, test takers might attach importance and value to tests. It is because

> test takers are very realistic about the consequences of the tests. Experience has taught them that these consequences go far beyond the test score and may affect a number of crucial future events, for better or for worse. Tests can affect self-esteem, confidence, pride, stigmas and opportunities. (Shohamy 2001b, p. 14)

In such a situation, test takers might feel they have no choice but to comply with tests' rules to gain the benefits related to tests. Therefore, inevitably, tests arouse some positive and negative feelings but mostly negative ones such as stress, fear, and anxiety in test takers (Shohamy 2001b). As put by Shohamy (2017), "it is the powerful uses of tests—their detrimental effects and their uses as disciplinary tools that are responsible for the strong feelings that tests evoke in test takers" (p. 443).

How test takers see themselves vs. tests and testers is highly related to test power (Shohamy 2001a). When tests serve as a "normalizing gaze," test takers feel that they have to confirm to the ideal of how an individual in the society should behave (Foucault 1995). Tests compel test takers to work for the immediate goal of getting good scores. Shohamy (2013) says, "Test takers surrender to the demands of tests, accept the testing discourse as 'the truth,' and comply with it" (p. 225). It is because they are aware that a high score on tests is the only way to get the economic and social values attached to tests, which results in test-related anxiety. Thus, test takers are subjected to test power. Tests rarely face any objection by test takers; rather they believe in their essentiality to ensure an equal and objective stance (Shohamy 2001a, b). Young (2012) agrees that test power is originated from the trust that those who are affected by tests place in them. In other words, what makes tests powerful also depends on the meaning individuals attribute to tests and the feelings evoked in test takers about tests. When individuals are judged on the basis of their performances on tests, tests turn into powerful tools playing central roles in their lives. Therefore, they themselves contribute to generating the power of tests. As Shohamy (2001b) states, "The power of the test, as expressed in the fear and respect that those affected by the test have for it, guarantees an almost automatic response—behaviors will be changed" (p. 35).

Here are some key codes that appear in this scoping review identifying the feelings and meanings tests evoke in test takers when tests exercise power over them:

- fear, frustration, tension, anger, and even humiliation
- lack of control and feeling helpless
- pressure, competition, and anxiety, stress
- central position of test in test takers' lives
- feeling as a victim of test
- need to match their performance to the demands of the tests
- need to change their behaviors in line with the demands of the tests
- need to develop strategies to comply with the demands of the test
- feeling obliged to comply with every decision made through the tests
- dependence, blind trust

- feeling of being in the hands of the testers that have control over them
- feeling they need luck and supernatural forces for doing well
- need to do anything, even unethical behaviors, to maximize their opportunity to succeed on tests
- feeling of being powerless
- feeling of wasted time while studying for tests
- feeling lack of freedom of action
- assigning importance to tests
- symbol of achievement
- sign of gaining identity
- meaning of life
- sign of success or failure in society and life.

As the above given list reveals, test takers seem to live through a great many emotional states, from anger and frustration to seeing the test as the sole meaning of their lives, because "the test taker is understood as a social being whose subjectivity is a function of subject position realized in the test itself" (McNamara and Roever 2006, p. 196). In this way, tests serve the function of a social tool constructing the identities reinforced in them.

8.5.4 The Functions of Tests

This review has found an overwhelming number of studies (59 out of 60) highlighting the functions of the language tests in the generation of their power. They point out that in testing contexts, tests can have educational, political, social, ideological, and economic functions.

As Shohamy states, "tests are often used for a variety of undeclared and covert purposes, other than just 'measuring knowledge'" (2009, p. 51). These purposes vary from making decisions about individuals' lives (Coniam and Falvey 2007) to discriminating, separating, categorizing, and labeling (Hamp-Lyons 2007, p. 487). Tests might "also influence the social systems in which they play a part when results are used to make important decisions" (Saville and Khalifa 2016, p. 78). In Foucault's (1995) analysis, the mechanism used by examinations to structure modern societies socially is presented in the following way:

> Examination combines the techniques of an observing hierarchy and those of a normalizing judgement. It is a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through which one differentiates them and judges them. That is why, in all the mechanisms of discipline, the examination is highly ritualized. In it are combined the ceremony of power and the form of the experiment, the development of force and the establishment of truth. (p. 184)

Driven by Foucault's perspective of "disciplinary power," Shohamy (2001b, 2005) states that tests serve the function of a disciplinary tool by imposing behaviors on those who are tested and affected by their results. Thus, tests might manipulate and control educational systems according to set agendas (Shohamy 2007a, b, 2013).

How tests work as de facto policies so as to achieve the changes in educational and social systems desired by policy makers is also highlighted in the literature (Menken 2008b, 2017; Shohamy 2001b, 2006, 2007a, 2008b). In fact, tests which are scientifically stated as neutral can promote certain social and educational policies, agendas, priorities, and values (McNamara and Shohamy 2008; Shohamy and Menken 2015). Therefore, tests as de facto powerful practices are forms of control and manipulation of policies. As they are used by those in authority, they turn into ideology in practice and implementation.

The underlying functions of tests, whether they are desired and planned or unplanned and undesirable, should be clarified to understand test power. On the basis of this review of language testing and assessment literature, the following functions of tests are presented.

- demonstrating authority
- discipline
- pressure
- changing educational systems/curricula/political systems
- changing the behavior of all those affected
- imposing policy/knowledge/control of mind/certain behaviors
- controlling educational systems
- marketing
- social control/construct a singular, standardized culture
- determining future/individuals' future opportunities
- solving the troubled systems as practical solution
- quantification, normalization, and standardization of people according to a common yardstick
- selection
- entrance/gate keeping
- access to valued social resources, such as wealth, jobs, status.
- observation/surveillance, screening populations
- competition/comparing
- classification/categorization purposes
- maintaining and creating social class and order
- imposing sanctions
- discriminating
- perpetuating ideologies
- making detrimental/high stake/important decisions for individuals such as selection, controlling immigration, accepting jobs, graduating from high schools, entering universities, obtaining high-ranking jobs, or entering elite institutions
- admission
- promotion
- placement
- graduation

- making judgment/assigning what is good/bad/successful
- construction of identities
- providing or taking away opportunities
- managing and controlling the linguistic repertoire of the nation
- determining prestige and the status of languages
- determining language correctness and standards
- establishing prestige and respect/prompting status of the subject
- promoting certain values and diminishing others
- punishment/a tool to threaten/used as a forcing and threatening device.

8.6 Insights Gained: A Framework on the Power of Tests as Social Practice

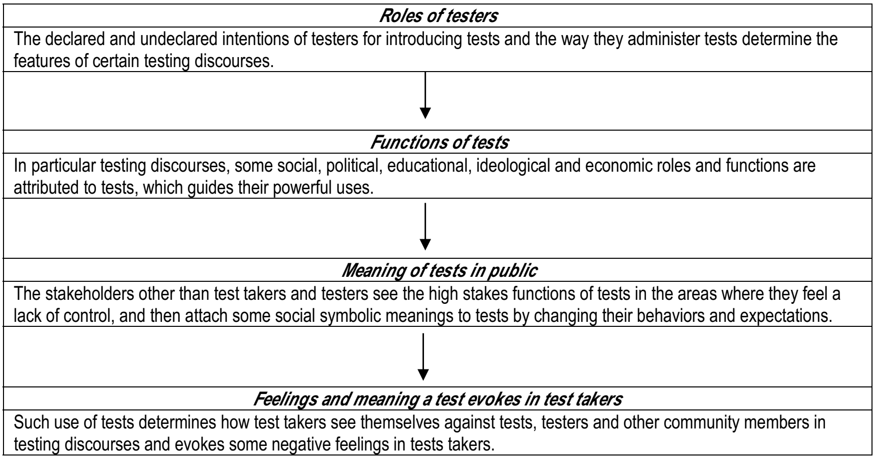
Tests as social practice have increasingly come to the fore in language testing and assessment studies which are primarily driven by the critical examination of testing practices in their social and cultural contexts to detect the identities and discursive power created through their uses (McNamara and Roever 2006; Shohamy 2001b, 2013; Young 2012).

Here, in this study, social practice is defined as the practice of production that people do collaboratively in all domains of life. All social practices are associated with what materials they work on, what means are used for their production and what social relations they produce. To Fairclough (2010),

> all practices involve identification, the construction of social identities- every practice is associated with particular 'position' for people refers to 'position and practices' in terms of which their identities and social relations are specified. However, there are different *performances* in these positions depending on the social memberships and life histories of those who occupy them and different identities attached to different performances. (pp. 172–173)

Regarding tests as social practice implies the uses and roles of testing activities in the constructions of individuals, societies, and discourses (Filer 2000). Because tests are ideologically shaped by power and struggle for power in relations, tests as social practice tend to constitute social order in society. Furthermore, they might contribute to social continuity and social change by leading individuals to generate representations of their identities and performances according to their positions.

Therefore, in a particular testing context, the deployment of tests as social practice might signify the expectations of testers from test takers for some identities and roles in the society. As Young (2012) puts it, "language testing is the construction and reflection of these social expectations through actions that invoke identity, ideology, belief and power" (p. 185). This is the core of the power that tests exercise in testing situations. Figure 8.1 explains how the four factors generated from the literature interact with each other and thereby make tests as exercises of power in testing discourses.



Roles of testers

The declared and undeclared intentions of testers for introducing tests and the way they administer tests determine the features of certain testing discourses.

Functions of tests

In particular testing discourses, some social, political, educational, ideological and economic roles and functions are attributed to tests, which guides their powerful uses.

Meaning of tests in public

The stakeholders other than test takers and testers see the high stakes functions of tests in the areas where they feel a lack of control, and then attach some social symbolic meanings to tests by changing their behaviors and expectations.

Feelings and meaning a test evokes in test takers

Such use of tests determines how test takers see themselves against tests, testers and other community members in testing discourses and evokes some negative feelings in tests takers.

Fig. 8.1 Generation of the power that tests exercise

The power of tests originates in testers' power because in most testing contexts who determines what to test, how to test, how to score, how to administer and determine results, are the testers who hold the power and authority. Through the use of tests, testers exercise non-negotiable and continuous control over individuals and society at large. As van Dijk (2008) points out, "Those who control discourse may indirectly control the minds of people. And since people's actions are controlled by their minds (knowledge, attitudes, ideologies, norms, values), mind control also means indirect action control" (p. 9). When granted such power, tests and testers can control directly or indirectly the knowledge and behaviors of tests takers and society at large because "almost all the constructs that underlie high-stakes language tests are theories of individual cognition that can be measured in one context (the test) and are stable enough to be ported to other non-testing contexts where the language is used" (Young 2012, p. 180). When life-changing functions such as passing a class, attending a university, taking a job, getting degree, etc., are assigned to tests, they inevitably turn to milestones in test takers' lives (Bourdieu 1991).

Such high stakes implications of tests and the roles of testers promote tests' power in testing discourses, but they are not the only dimensions that generate it, as stated above. As Young (2012) explains, power is "co-constructed by all participants—both the powerful and the non-powerful… Non-powerful participants co-construct power by accepting the constraints imposed upon them" (p. 185). What makes tests powerful also depends on the meaning individuals attribute to tests, which often results in changes in their behaviors. How both test takers and test users determine their own subordinate positions in testing discourses also contributes to tests having power.

When the stakeholders other than test takers and testers accept high stakes functions of tests in the areas where they feel a lack of control, they attach some social symbolic meanings to tests. Then, they change their behaviors. For example, teachers teach to the tests and reduce the curriculum to what is tested in tests, as the success of their students is often viewed as a reflection of their teaching in the society. Additionally, parents invest in private tutoring for their children, as success on a test is seen as a reflection of their parenting. They become dependent on tests, developing unchallenged trust in their results and in their power. Therefore, they expect test takers to meet assumed standards of knowledge, skills, and attitudes required in tests and might marginalize those who do not do well in tests.

These three dimensions determine how test takers see themselves in relation to tests, testers, and other community members who guide their subordinate positions in testing discourses. In most cases, as stated in the reviewed literature, test takers who develop some negative feelings toward tests have to adapt their behaviors, values, and knowledge in accordance with the demands of the tests, as they fear detrimental effects on their lives in case they do not well in them.

8.7 Conclusion: Implications for Test Users

This study has argued that because language tests are embedded in their use in education and society, they are social practices playing significant roles in construction and development of knowledge, individuals, and societies. In fact, because of the power "reinforced by dominant social and educational institutions as major criteria of worth, quality and value," language tests exercise high stakes roles in testing contexts (Shohamy 2009, p. 50).

Understanding the power of a test requires looking at the situation from a broader perspective (Shohamy 2008b). The awareness of the powerful nature of tests motivates the introduction of tests as social and policy tools (Menken 2008b; Shohamy 2008b). In fact, although tests are scientifically verified as neutral and objective, they might enjoy power and control (Menken 2008b; Shohamy 2008a, b). Yet, individuals rarely consider tests as the products of the policies because the policies embedded in tests "are typically implicit rather than explicit, though extremely powerful in shaping changes" (Menken 2017, p. 387).

The framework outlined in this chapter is an attempt to break down the disciplinary walls between language testing and other areas of applied linguistics. It suggests questioning and analyzing the operations of the four dimensions in a particular testing discourse. Therefore, the matter prioritized here is mostly based on identifying the power of tests by drawing attention to the roles of testers, the meaning of tests in public, the feelings and meaning test evokes in test takers and the functions of tests. With these dimensions in mind, tests should be considered with their relations to "educational, pedagogical, bureaucratic, psychological, social and political variables that affect people, knowledge, curriculum, teaching, learning, ethicality, social classes, bureaucracy, politics, inclusion and exclusion" (Shohamy 2007c, p. 522).

What is needed is first, to listen to the lived experiences and interpretations of all test users and then have test users view testing practices critically by reflecting on their own experiences so that they could understand their *subordinate position* in the social system. As Shohamy (2001b) explains,

> there is a need for sharing the power of tests by training the public in testing methods, in the testing process and in the rights of test takers. Testing cannot remain a field that belongs only to testers but rather test takers and the public at large need to be part of the discussion. (p. 158)

We argue that a balanced power between tester and test takers should be encouraged by the assumption that testers should not be the only ones who are responsible for all the knowledge of tests. As Davidson (2002) notes, "We must be self-aware and willing to observe what we do, and we must be capable (and I think, trained) to involve and listen to a wide array of interested parties" (p. 107). Only through such an awareness, can test takers and test users view tests critically to question their uses and functions in their lives, understand test power, and become testing literate individuals.

References

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bourdieu, P. (1991). *Language and symbolic power*. Cambridge, UK: Polity Press.

Brindley, G. (2008). Educational reform and language testing. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol. 7, pp. 365–378). New York, NY: Springer.

Broad, B. (2007). The power of tests: A critical perspective on the uses of language tests by Elena Shohamy. *Journal of Writing Assessment, 3*(1), 55–60.

Broadfoot, P. (1996). *Education, assessment and society: A sociological analysis.* Buckingham: Open University Press.

Brown, A., & McNamara, T. (2004). "The devil is in the detail": Researching gender issues in language assessment. *TESOL Quarterly, 38*(3), 524–538.

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment,* (2nd ed., Vol. 7, pp. 349–364). New York, NY: Springer.

Coniam, D., & Falvey, P. (2007). High-stakes testing and assessment. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 457–471). New York, NY: Springer.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–18). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1989). Construct validity after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Champaign, IL: University of Illinois Press.

Davidson, F. (2002). Book review: The power of tests: A critical perspective on the uses of language tests. *Language Testing, 19*(1), 105–107.

Davies, A. (2008). Ethics, professionalism, rights and codes. In E. Shohamy & N.H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol 7, pp. 429–443). New York, NY: Springer.

Davies, A. (2012). Ethical codes and unexpected consequences. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 469–482). New York, NY: Routledge.

Davies, A. (2017). Ethics, professionalism, rights, and codes. In E. Shohamy, L. G. Or, & S. May (Eds.), *Encyclopedia of language and education, Language testing and assessment* (3rd ed., Vol. 10, pp. 397–415). Cham, Switzerland: Springer.

Douglas, D. (2010). Book review: McNamara, T. and Roever, C. *Language testing: The social dimension* (Language Learning and Monograph Series). Malden, MA and Oxford, UK: Blackwell Publishing. *Language Testing*, *27*(2), 283–285.

Fairclough, N. (2010). *Critical discourse analysis: The critical study of language* (2nd ed.). London, UK: Pearson.

Filer, A. (2000). *Assessment: Social practice and social product*. New York, NY: Routledge.

Filer, A., & Pollard, A. (2000). *Social world of pupil assessment: Processes and contexts of primary schooling*. London: Continuum.

Foucault, M. (1995). *Discipline and punish: The birth of the prison* (A. Sheridan, Trans.). New York, NY: Vintage.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System, 28*(4), 579–591.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 487–504). New York, NY: Springer.

Hudson, T. (2012). Standards-based testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 493–508). New York, NY: Routledge.

ILTA. (2000). *International Language Testing Association code of ethics*. http://www.iltaonline.com/images/pdfs/ILTA_Code.pdf. Retrieved March 25, 2016.

ILTA. (2007). *International Language Testing Association guidelines for practice*. https://cdn.ymaws.com/www.iltaonline.com/resource/resmgr/docs/ilta_guidelines.pdf. Retrieved March 25, 2016.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2005a). Language assessment from a wider context. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 779–794). London: Lawrence Erlbaum.

Kunnan, A. J. (2005b). Towards a model of test evaluation: Using the test fairness and test context frameworks. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity. Proceedings of the ALTE Berlin Conference* (pp. 229–251). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2008). Large scale language assessments. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol. 7, pp. 135–156). New York, NY: Springer.

Lazaraton, A. (2010). From cloze to consequences and beyond: An interview with Elana Shohamy. *Language Assessment Quarterly, 7*(3), 255–279.

Luxia, Q. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142–173.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351–372.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum: 87th yearbook of the National Society for the Study of Education, Part I* (pp. 83–121). Chicago, IL: University of Chicago Press.

Matsugu, S. (2011). *Language testing: The social dimension*, by Tim McNamara and Carsten Roever. *Language Assessment Quarterly, 8*(1), 99–102.

Mays, N., Roberts, E., & Popay, J. (2001). Synthesising research evidence. In N. Fulop, P. Allen, A. Clarke, & N. Black (Eds.), *Studying the organization and delivery of health services: Research methods* (pp. 188–220). London: Routledge.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333–349.

McNamara, T. (2005). 21st century shibboleth Language tests, identity and intergroup conflict. *Language Policy, 4*(4), 351–370.

McNamara, T. (2008). The socio-political and power dimensions of tests. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol. 7, pp. 415–428). New York, NY: Springer.

McNamara, T. (2009a). Language tests and social policy. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration: Critical perspectives on language testing regimes in Europe* (pp. 153–163). Amsterdam: John Benjamins.

McNamara, T. (2009b). Principles of testing and assessment. In K. Knapp & B. Seidlhofer (Eds.), *Handbook of foreign language communication and learning* (pp. 607–628). Berlin and New York, NY: Mouton de Gruyter.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Oxford: Blackwell.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics, 18*(1), 89–95.

Menken, K. (2008a). *English learners left behind: Standardized testing as language policy*. Bristol: Multilingual Matters.

Menken, K. (2008b). High-stakes tests as de facto language education policies. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol. 7, pp. 401–414). New York, NY: Springer.

Menken, K. (2017). High-stakes testing as de facto language education policies. In E. Shohamy, L. G. Or, & S. May (Eds.), *Encyclopedia of language and education, Language testing and assessment* (3rd ed., Vol. 7, pp. 385–396). Cham, Switzerland: Springer.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Moder, C. L., & Halleck, G. B. (2012). Designing language tests for specific social uses. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 151–163). New York, NY: Routledge.

Ross, S. J. (2011). The social and political tensions of language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 786–797). New York, NY: Routledge.

Saville, N. (2012). Quality management in test production and administration. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 493–508). New York, NY: Routledge.

Saville, N., & Khalifa, H. (2016). The impact of language assessment. In B. Jayanti & D. Tsagari (Eds.), *Handbook of second language assessment* (pp. 77–94). Berlin: De Gruyter Mouton.

Shohamy, E. (1990). Discourse analysis in language testing. *Annual Review of Applied Linguistics, 11,* 115–131.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning* (NFLC Occasional Papers). Washington, DC: National Foreign Language Center.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*(3), 340–349.

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*(4), 331–345.

Shohamy, E. (2001a). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391.

Shohamy, E. (2001b). *The power of tests: A critical perspective on the uses of language tests.* London: Longman/Pearson Education.

Shohamy, E. (2001c). The social responsibility of the language testers. In R. L. Cooper, E. Shohamy, & J. Walters (Eds.), *New perspectives and issues in educational language policy in honor of B. D. Spolsky* (pp. 113–130). Amsterdam: John Benjamins.

Shohamy, E. (2005). The power of tests over teachers: The power of teachers over tests. In D. J. Tedick (Ed.), *Second language teacher education: International perspectives* (pp. 101–111). New York, NY and London: Routledge.

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. London: Routledge.

Shohamy, E. (2007a). Language tests as language policy tools. *Assessment in Education, 14*(1), 117–130.

Shohamy, E. (2007b). Tests as power tools: Looking back, looking forward. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 141–152). Ottawa, ON: University of Ottawa Press.

Shohamy, E. (2007c). The power of language tests, the power of the English language and the role of ELT. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching: Part one* (pp. 521–531). New York, NY: Springer.

Shohamy, E. (2008a). Introduction to volume 7: Language testing and assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol. 7, pp. 150–159). New York, NY: Springer.

Shohamy, E. (2008b). Language policy and language assessment: The relationship. *Current Issues in Language Planning, 9*(3), 363–373.

Shohamy, E. (2009). Language tests for immigrants: Why language? Why tests? Why citizenship. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration: Critical perspectives on language testing regimes in Europe* (pp. 45–59). Amsterdam: John Benjamins.

Shohamy, E. (2013). The discourse of language testing as a tool for shaping national, global, and transnational identities. *Language and Intercultural Communication, 13*(2), 225–236.

Shohamy, E. (2017). Critical language testing. In E. Shohamy, L. G. Or, & S. May (Eds.), *Encyclopedia of language and education. Language testing and assessment* (3rd ed., Vol. 7, pp. 441–454). Cham, Switzerland: Springer.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298–317.

Shohamy, E., & Menken, K. (2015). Language assessment: Past to present misuses and future possibilities. In W. Wright, S. Boun, & O. Garcia (Eds.), *Handbook of bilingual and multilingual education* (pp. 253–269). Hoboken: Wiley-Blackwell.

Spolsky, B. (1998, June 15). What is the user's perspective in language testing? Paper presented at the Colloquium, *The state of the art in language testing: The user's perspective.* National Foreign Language Center, the Johns Hopkins University, Washington, DC.

Spolsky, B. (2008a). Introduction: Language testing at 25: Maturity and responsibility? *Language Testing, 25*(3), 297–305.

Spolsky, B. (2008b). Language assessment in historical and future perspective. In E. Shohamy & N. H. Hornberger (Eds.) *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol. 7, pp. 445–454). New York, NY: Springer.

Spolsky, B. (2012). Language testing and language management. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 509–519). New York, NY: Routledge.

Taylor, L. (2005). Washback and impact. *ELT Journal, 59*(2), 154–155.

van Dijk, T. A. (2008). *Discourse and power*. New York, NY: Palgrave Macmillan.

Xi, X., & Davis, L. (2016). Quality factors in language assessment. In B. Jayanti & D. Tsagari (Eds.), *Handbook of second language assessment* (pp. 61–76). Berlin: De Gruyter Mouton.

Young, R. F. (2012). Social dimensions of language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 178–193). New York, NY: Routledge.

Chapter 9
The Washback Effect of the Vietnam Six-Levels of Foreign Language Proficiency Framework (KNLNNVN): The Case of the English Proficiency Graduation Benchmark in Vietnam

Phương Hoa Đinh Thị and Handoyo Puji Widodo

Abstract In educational research, tests have been considered one of the dominant determiners of what happens in classrooms that can influence teaching and learning activities. English foreign language tests in particular are no exception. There is a plethora of empirical research on the impact of high-stakes language tests upon language learning and teaching. To extend this scholarship, this chapter reports a case study that investigates the washback of the Vietnam Six-Levels of Foreign Language Proficiency Framework called the KNLNNVN, a compulsory requirement for the National University of Arts Education (NUAE) graduation. The study reported in this chapter adopts a triangulation of numerous methodologies to investigate how the KNLNNVN affects the teaching and learning of English Foreign Language (EFL) for non-English major students at NUAE in Hanoi, Vietnam. The findings of this study include the domains of test validity including teachers' KNLNNVN knowledge and abilities, the influence of the KNLNNVN and its English Test on what and how teachers teach, its influence on teachers' methods of assessment, curriculum, and materials; and its effects on students' awareness of the KNLNNVN and methods of learning. This study has contributed to the knowledge of the nature of CEFR-adapted washback upon the entire teaching and learning of English situated within an EFL context, such as Vietnam.

P. H. Đinh Thị
Hanoi Law University, Hà Nội, Vietnam
e-mail: dinhphuonghoa.ecas@gmail.com

H. P. Widodo (✉)
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: handoyopw@yahoo.com

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_9

9.1 Introduction: Purpose and Testing Context

Recently, the role of English as a global language has encouraged the adoption of standard frameworks for designing language curricula, instruction, and assessment. During the last decades, the Common European Framework of Reference for Foreign Languages (CEFR) has been a global phenomenon where many countries have adopted or adapted this framework to their local educational practices, such as designing language curricula and syllabi, instructional practices, and testing and assessment. As a result, the adoption and adaptation of the CEFR in different educational contexts have been of great interest among policymakers, language testing providers, language curriculum and materials developers, assessment and test designers, language teachers, and language learners worldwide.

In particular, in the field of language assessment, a plethora of empirical research has been conducted to examine the impact of the CEFR on locally tailored language tests in different English as a foreign language and second language contexts, such as in China, Iran, and Malaysia (Afip et al. 2019; Sims and Chen 2019). This phenomenon also has taken place in Vietnam (Pham and Bui 2019) where English has been a compulsory subject at secondary and tertiary settings as well as an elective subject at primary schooling from 1982 to the present. In response to this, the Vietnamese Prime Minister has decided to establish the 2008–2025 National Foreign Languages Project. This project is intended to enact educational innovation and evaluation of foreign language teaching and learning at all levels in the national education system.

Through this project scheme, the Vietnamese Ministry of Education and Training issued the Circular N^0 01/2014/TT-BGDĐT of January 24, 2014, approving the KNLNNVN. This framework comprises six levels that are compatible with the Common European Framework of Reference for Languages and other common international language proficiency levels. These frameworks serve as the basis for designing/writing curricula and teaching plans. As spelled out in the National Foreign Languages Project, higher education institutions that do not specialize in foreign languages require undergraduate students to participate in the new language-training program. In this training program, undergraduate students need to have a language proficiency of KNLNNVN level 3 upon the completion of their undergraduate degree. Following this framework, the English Proficiency Tests from Level 2 to Level 5 (EPT.2 and EPT.3-5) are set up and administered. In this respect, EPT.2 is aligned with A2 of CEFR, and EPT.3-5 is aligned with B1, C1, and C2 of CEFR. Thus, the KNLNNVN has become a very high-stakes test among non-English majors.

9.2 Testing Problem Encountered

Through the lens of higher education administrators or policymakers, setting a graduation benchmark is one of the best ways to monitor students' English learning outcomes or to enhance the quality of English language education (Tsai and Tsou 2009). In response to this global trend, the KNLNNVN is designed based on the CEFR, situated in the Vietnamese educational context. This framework comprises six levels and contains *can-do* descriptors that are compatible with CEFR and other common international language proficiency levels. Therefore, the KNLNNVN is used as a reference when developing syllabi, creating tests/exams, marking/grading exams, assessing language learning needs, designing courses, developing learning materials, and calibrating language policies/assessment.

The KNLNNVN operationalizes foreign language proficiency at three broad bands with six main levels: Levels 1 and 2, Levels 3 and 4, Levels 5 and 6. The scale starts at Level 1 and finishes at Level 6. This scaling is aligned with the CEFR from A1 to C2 as seen in Table 9.1.

The KNLNNVN helps to clearly define certain requirements for language skills, such as listening, speaking, reading, and writing. Hence, at the English Level 2 (A2) of the KNLNNVN, learners are supposed to be able to perform four main language activities such as listening, speaking (spoken interaction), reading, and writing (written production) within the public, personal, educational, and occupational domains in which they can work on some types of texts and questions.

In response to the national reform to enact the KNLNNVN, the Rector of NUAE decided to choose the English proficiency of KNLNNVN Level 2 (A2). This is part of the university policy for undergraduate graduation from 2015 to 2020. Consequently, teachers and students have begun to learn about KNLNNVN and EPT. Hence,

Table 9.1 The six-levels of the KNLNNVN

KNLNNVN (Level)		General descriptions
A—Basic user	Level 1 (A1—Breakthrough)	Can communicate in basic English with help from the listener
	Level 2 (A2—Way-stage)	Can communicate in English within a limited range of contexts
B—Independent user	Level 3 (B1—Threshold)	Can communicate essential points
	Level 4 (B2—Vantage)	Can use English effectively, with some fluency, in a range of contexts
C—Proficient user	Level 5 (C1—Effective Operational Proficiency)	Can use English fluently and flexibly in a wide range of contexts
	Level 6 (C2—Mastery or Highly proficient)	Can use English, very fluently, precisely, and sensitively, in most contexts

many Vietnamese educational institutions and KNLNNVN preparation courses have come to exist in the Vietnamese market in which NUAE's teachers of English are KNLNNVN preparation course trainers. The KNLNNVN and EPT appear to have so strong an impact upon NUAE teachers of English that all EFL teaching policies are likely to change, that is, lead to changes of the teaching and learning EFL activities.

Consequently, this case study aims to investigate the changes from the KNLNNVN to institutional policies of EFL teaching and learning as well as examine key dimensions of teaching and learning activities at NUAE. To collect empirical data regarding those issues, the data were garnered from questionnaires, individual interviews, observations, and focus groups and were triangulated in order to ensure rigor and trustworthiness.

9.3 Review of Literature

9.3.1 Defining Washback and Language Proficiency Framework

In applied linguistics, the term *washback*, or backwash, is defined as the effect or impact of tests (e.g., standard tests) on curriculum/syllabus design, language teaching, and language testing (Alderson and Hamp-Lyons 1996; Alderson and Wall 1993; Bailey 1999; Đinh 2019; Hughes 1989; Messick 1996; Pearson 1988; Watanabe 1996). Previous studies show that washback effects could be positive if tests are properly designed and appropriately used and/or negative if tests exert negative influence on students' learning (Shohamy 2014; Tsagari and Cheng 2017). The washback of tests can also have direct or indirect effects on student learning (Alderson and Hamp-Lyons 1996; Forbes 1973; Wall and Alderson 1993; Watanabe 1996).

In the educational evaluation, washback is considered as the impact of tests or examinations that can drive learning and teaching. This leads to measurement-driven or test-driven instruction (Popham 1987, cited in Cheng and Curtis 2004, p. 4). Fitz-Gibbon (1996) defined impact as any effect of the service [or of an event or initiative] on an individual or a group. This definition indicates that the impact can be positive or negative and may be intended or accidental. Following this definition, measuring impact deals with identifying and evaluating change (Streatfield and Markless 2009).

Furthermore, Messick (1989) expanded the concept of consequential validity, changing the previous notions about score interpretation and test use. The concept of washback in test validity research is primarily associated with Messick's concept of consequential validity. Therefore, washback is operationalized as an "instance of the consequential aspect of construct validity and a focal point of validity research" (Messick 1996, p. 242), which covers components of test use, the impact of testing on test takers and educators, the interpretation of results by decision-makers, and any possible misuses, abuses, and unintentional effects of tests. The influences of tests on teachers, students, institutions, and society are accordingly deemed to be one type of validity evidence. Other researchers have also emphasized the meaning of justifying test use and exploring its consequences (Cronbach 1988; Shohamy 2000). With this in mind, washback also plays a key role in the process of educational innovation and evaluation in language teaching and learning (Đinh 2017; Shohamy 1992).

In addition to the washback of testing, it is important to operationalize a language proficiency framework. Before defining this framework, what competence means needs to be defined. Council of Europe (2001) (hereafter CoE) defines the term, *general competences* as those not specific to language, but which are called upon for actions of all kinds, including language activities. In addition, the term *competence* is referred to as the accumulation of knowledge, skills, attitudes, and experiences that allow a person to perform particular tasks (Widodo 2015). Therefore, when this definition is applied to language performance, *the communicative language competence* of a learner is manifested through "the performance of the various language activities, including reception, production, interaction or mediation in the public, the personal, the educational and the occupational domains" (CoE 2001, p. 14).

In the CEFR, the "framework" is used to define levels of proficiency that allow the progress of language learners to be measured at each stage of learning and on a life-long basis (CoE 2001, p. 9). Furthermore, It serves as the basis for developing language learning and teaching activities, the principles of language testing and assessment more effectively. The framework is widely used by language learners, teachers, examiners, textbook writers, teacher trainers, educational administrators, and those interested in language curriculum, learning, teaching, and assessment (CoE 2001).

9.3.2 Previous Washback Studies

In language testing and assessment, washback, either a positive or negative effect, has been widely researched. In particular, the CEFR impact on assessment has been well documented to date (Coste 2007; Little 2007, cited in Jones and Saville 2009, p. 53) because of its role in consequential validity among others. For example, the washback model of Alderson and Wall (1993) is considered a classic and landmark study. Alderson and Wall (1993) used an observation method to investigate the washback of English language teaching and learning in the Sri Lankan context. Alderson and Wall (1993, pp. 120–121) developed the fifteen hypotheses (WHs) that included different possible aspects of washback effects on what to teach/learn; how to teach/learn; the rate and sequence of teaching/learning; the degree and depth of teaching/learning; and the attitudes to content, methods of teaching/learning. Alderson and Hamp-Lyons's (1996) model (p. 296) used interviews and one-week classroom observations so as to review and correct WHs of Alderson and Wall (1993) that "tests will have different amounts and types of washback on some teachers and some learners than other teachers and learners" (p. 296). These two studies could be the point of departure for more investigations into the washback effects of tests on educational (curriculum, instruction, and assessment) practices.

Recently, tertiary-level EFL graduation requirements with reference to CEFR have been adopted in many Asian universities (e.g., Japan, Korea, Malaysia, Taiwan, Vietnam) in order to hone university students' EFL proficiency and to meet university graduates' global market competitiveness (Pan and Newfields 2012; Sims and Chen 2019). Therefore, Asian universities set English proficiency graduation requirements in which CEFR has become a norm. Indeed, this policy views a high-stake or standardized test as a major determinant of course design and curriculum/classroom practices. In response to this global phenomenon, studies into the washback effects of CEFR-oriented EFL proficiency graduation requirements on university EFL curriculum, instruction, and assessment have burgeoned. Most of these studies have focused upon how university students view the adoption of standardized English language proficiency (ELP) tests as a tool for assessing their English competence for graduation (Tsai and Tsou 2009).

To begin with, in the Taiwanese EFL context, Pan and Newfields (2012) examined the impact of English proficiency graduation requirements on 17 tertiary educational institutions in Taiwan in response to Taiwan's Ministry of Education (TME) policy on English proficiency graduation requirements or English thresholds for graduates to obtain a level of English proficiency modified according to the CEFR B1 or A2 levels. Data were collected through a questionnaire survey and structured interviews. The survey and interview data showed that the English graduation requirements had an impact on increased motivation for English study, more time allocated to English study, more variation in the methods adopted to study English, and more test-related practice.

In the same vein, Wu and Lee (2017) explored Taiwanese university students' views of the English proficiency graduation policy in three universities where students had to take the General English Proficiency Test (GEPT) prior to graduation. The findings indicated that most of the university students with mixed English proficiencies had a positive attitude toward the English graduation benchmark policy. In particular, the intermediate group showed more positive attitudes toward the graduation requirement policy than the high-intermediate group. The results revealed that the university students' attitudes toward the English graduation requirement positively influenced their learning motivation although no significant relationship between the attitudes toward the policy and test performance existed. This empirical evidence contributes to a better understanding of university students as major stakeholders who know the context of test use.

In the Vietnamese university context, Pham and Bui (2019) investigated students' voices/views on the nationwide enactment of the English graduation benchmark policy in Vietnamese universities based in Northern, Central, and Southern Vietnam regions. They also looked into how university students with elementary, intermediate, and upper-intermediate English proficiencies perceived the policy. A total of 902 students were recruited as participants. Data were collected through a questionnaire survey and analyzed by using inferential statistical tools. ANOVA and Hochberg's GT2 post hoc tests yielded significant differences in students' voices regarding the "Benefits" and "Anxiety" factors. MANOVA and the Bonferroni analyzes also showed differences in students' voices on the "Benefits," "Anxiety," and

"Test-oriented learning" factors. These results suggest that the Ministry of Education and Training (MOET), policymakers, higher education institutions, and university teachers need to direct the policy for innovative curriculum practices.

Despite the ever-increasing use of English language proficiency as a graduation requirement in Taiwanese and Vietnamese universities (Pham and Bui 2019; Tsai and Tsou 2009; Wu and Lee 2017), university policymakers', university teachers', and students' voices on this one-size-fits-all policy remain under-researched. In particular, a research gap in the language assessment literature is that university students' voices about the benchmark policy as a catalyst for needs analysis have been little heard. In the Vietnamese EFL context, scant empirical evidence (Pham and Bui 2019) seeking Vietnamese students' opinions on the policy has been reported. To extend empirical scholarship into the washback effect of the CEFR-based English language proficiency requirement for graduation, the current study investigated the washback effect of KNLNNVN on language learning and teaching situated in non-English major departments in one of the Vietnamese universities. The contribution of this study may provide insight into educational innovations driven by the English language benchmark policy in order to aim for improved future English education at a tertiary level.

9.4 Methodology

This study was conducted between January 2014 and November 2018. It aimed to explore the changes in language learning and teaching when KNLNNVN was introduced in 2014. Under this scheme, the first cohort of NUAE's students had to sit the EPT. 2 graduation examination in 2017. For this study, data were collected through (1) educational artifact documentation, (2) focus group interviews, (3) structured questionnaires, and (4) classroom observations. It is important to highlight a couple of things. First, all these data were mutually supporting. Second, regarding the questionnaire, qualitative input and piloting procedures were carried out in order to ensure content validity (Low 1988, cited in Cheng 2004, p. 151) and thus its consequential validity (Messick 1996).

9.4.1 Sample

Following the receipt of ethical approval, students, English teachers, and policymakers were recruited. Thus, the sample of the study included 679 students (18–22 years old) from different areas of Northern Vietnam, the Rector of NUAE, Head of Training Department, and 13 English teachers of NUAE (Director, two Vice Directors of Foreign Language Center, and nine academic staff). Because the first author worked as a reseacher, she was not a participant of this study. All these faculty members were recruited for this study because they participated in the national reform project, so they could provide relevant and accurate information regarding the impact of the CEFR on student learning. In doing so, the interviews were conducted in three phases: the first interview with the Rector; the second interview with leaders of the Training Department and Foreign Language Central; and the third interview with nine teachers of English.

NUAE is one of the Vietnamese universities that offer undergraduate programs that do not specialize in foreign languages. At this university, all undergraduate students are supposed to obtain an English language proficiency of KNLNNVN Level 3 upon graduation. However, the Rector of NUAE decided to apply for the English proficiency of KNLNNVN Level 2 (A2) upon graduation because (1) the students received limited training time; (2) they were placed in large mixed-ability classes (from 55 to over 65 students); (3) classrooms had no microphone facilities; and (4) the students had low English proficiency although they had studied English for 10 years. Based on the placement test, students' English proficiency was at the beginner level (A0); therefore, EPT.2 (A2) of KNLNNVN has been a compulsory requirement for NUAE graduation since 2016; and EPT.3 (B1) will come into effect by 2021.

9.4.2 Instruments

9.4.2.1 Educational Artifact Documentation

The first author collected all institutional policy documents, called educational artifacts, which contained curricular guidelines, assessment materials, syllabi, and supplementary materials according to KNLNNVN and EPT 2. These documents served as the qualitative data in that they provided thick description which is particularly useful for phenomenological research in education. These artifacts were used for document analysis because they could provide "information about what has been encouraged or discouraged; about what has happened or will happen …" (Hinchey 2008, p. 77). Because all the documents are written in Vietnamese, all the verbal texts were translated and coded into several themes that related to the research questions.

9.4.2.2 Structured Questionaire

After piloting questionnaire items, the questionnaire was undertaken from 25 December 2017 to 12 January 2018. Simple random sampling was employed in this study. To compare the responses given by each group, a Teacher Questionnaire (TQ) and Student Questionnaire (SQ) which consisted of four parts. Both instruments utilized the same items in a modified and adapted form based on Cheng (2004). Due to limited space, both the TQ and SQ are described in Table 9.2.

Table 9.2 TQ and SQ

Numerical order	Concepts	Variables	Scales
A: Personal information			
Part 1	English proficiency, ages	4 (TQ)/2(SQ)	Nominal Scale
B. EFL teaching and learning activities:			
Part 2	Contents and communicative method of teaching EFL (including listening, speaking, reading, writing skills)	372 (including 4 skills)	5-point Likert scale of frequency
While-lesson activities	Topics	13 × 4 skills	
	Texts	18 × 4 skills	
	Question types	8 × 4 skills	
	Activities	7 × 4 skills	
After school (Homework)	Topics	13 × 4 skills	
	Texts	18 × 4 skills	
	Question types	8 × 4 skills	
	Activities	7 × 4 skills	
Post-lesson activities	Correct and Comment	1 × 4 skills	
Part 3	Materials	13	Nominal Scale
Part 4	Assessment (including listening, speaking, reading, writing tests)	10	Nominal Scale

9.4.2.3 Observations

In the Vietnamese context, one English teacher is responsible for teaching one English class. Between 2017 and 2018, there were 17 classes of English A2 in Semester 2 at NUAE. Every English teacher might teach one class of English A2 with 55 class periods on average.

After obtaining permission from all the participants, 10 classes (English level A2) of ten teachers were observed. Multiple observations took place from January to March 2018. Classes were scheduled one day per week with substantial uninterrupted work periods. Each of the class periods lasted approximately 200 min (four periods) per day. There were 55 class periods of English Level A2 from December 25, 2017 to March 23, 2018. Every class observation spanned 50 min. There were two observation rounds: Round 1 took place before the mid-term examination, and Round 2 was scheduled before the term examination to explore the differences of influences of KNLNNVN between two rounds as seen in the following observation timelines (also see Table 9.3).

Table 9.3 Observation timelines

Duration: Spring semester, 2018		
Round 1	**The length of classroom observation period**	**Time**
10 English lessons	50 min for each observation of one English lesson	from January 5 to March 23, 2018
Round 2	**The length of classroom observation periods**	**Time**
30 English lessons	150 min for each observation of three English lessons	from March 26 to March 30, 2018

The observation scheme was designed and adapted from Cheng (1999) in order to complement the questionnaire data (Part A of COLT). These qualitative data were used to probe into how KNNNNVN and EPT.2 influence teachers and students.

9.4.2.4 Interviews

After multiple observations, the focus group interviews were conducted in order to enrich the data collected from the classroom observations. Each of the interviews took 20–45 min. The open-ended questions were used to get more nuanced responses from the participants, such as the attitudes of the teachers and students, what learning and assessment tasks teachers used and taught, and how students responded to such tasks (Creswell 2008, as cited in Boyce 2010, p. 43). Following individual interviews, focus group interviews were carried out to explore more lived experiences of participants in the national CEFR-driven English Benchmark Requirement for Graduation Project. These different types of interviewing produced a myriad of specific information that might be comparable across the groups of participants (Cohen et al. 2000, as cited in Boyce 2010, p. 44). All the interviews were audio-taped and then selectively transcribed and finally translated into English.

9.4.3 Analysis Procedures

The data analyzes for each phase of this study are briefly outlined here. The frequency distributions were calculated for all the document analysis and questionnaire items/observation/interview data. The analysis involved calculation of the amount of time/times. This was applied to the observation data: Parts 1, 3, and 4 of TQ and SQ by using Excel and Statistical Product and Services Solutions software. All institutional policies on curriculum, the official course documents, methods of assessment, and supplementary materials, were assessed to determine whether teaching content and/or methods of assessment changed due to the KNLNNVN and EPT.2.

The survey was distributed to 12 teachers and 679 non-English major students at NUAE. Of the 691 surveys distributed, 691 valid questionnaires were returned. The survey explored the phenomenon of washback or backwash, and the influences of KNLNNVN and EPT.2 on teacher instruction and students' learning. The observation data were analyzed by using Part A of COLT to determine whether washback or influences of KNLNNVN and EPT.2 exist and to what extent they operate in classroom activities. Following the survey and observations, participants who agreed to be interviewed were recruited. All the interview data were analyzed to be triangulated. A method of triangulation with a complementary multiple-method design was used in this study to minimize errors arising from the data collection and analyzes. In this report, all the interview data, questionnaire responses, and document analysis were scrutinized to determine the influences of washback from the KNLNNVN and EPT.2 on all the curriculum practices at the university.

9.5 Findings

9.5.1 Findings from Document Analysis

The document analysis involved institutional policies on the curriculum, the official course documents, methods of assessment, and supplementary materials used by teachers. Relevant details of the analyzes are presented in Table 9.4.

9.5.2 Curriculum and Methods of Assessment

(1) Teaching contents and methods of assessment changed. Table 9.4 illustrates the changes in teaching contents and methods of assessment
(2) In 2017, some more authentic materials were included as the official documents (see Appendices 2 and 3).
(3) Teachers of English were encouraged to use texts taken from journals, books, and news for listening/speaking/reading and writing skills. Learning tasks designed by teachers included short-answer questions, gap-filling/identifications, and sentences/paragraphs identical to EPT.2 of KNLNNVN or practice tests at Levels A1 and A2.

The analysis of the official course documents indicated that the official course documents set before 2013 for semesters 1 and 2 were not aligned with EPT.2 of KNLNNVN or practice tests at Levels A1 and A2. This shows the impact of the EPT.2 of KNLNNVN on English learning and teaching before 2013. Since 2014, teachers of English have been encouraged to use a variety of authentic materials in addition to the official course documents. This was also reported by university leaders and teachers in the interviews.

Table 9.4 The changes in teaching contents and methods of assessment

Year	Teaching periods of semester 1	Teaching periods of semester 2	Teaching contents of semester 1	Teaching contents of semester 2	Formative assessment	Summative assessment (achievement test)	Learning outcomes of University graduation
2013	80	55	From Unit 1 to 14 of Lifeline textbook (Elementary)	From Unit 1 to 6 of Lifeline textbook (Pre-intermediate)	Questions and Answers or Writing Test (Grammar or Reading exercise)	Writing Test (Grammar and Reading exercise)	
2017	80	55	Four skills and grammar/vocabulary of KNLNNVN level 1	Four skills and grammar/vocabulary of KNLNNVN level 2	Speaking Test/Reading Test/Listening Test or Writing Test	Writing Test (Objective test and Writing test)	EPT.2 of KNLNNVN

9.5.2.1 Supplementary Materials Used by Teachers

The results of the analysis of the supplementary materials used and enacted by English teachers and students indicated that they used various authentic materials (see Appendix 3), such as commercial publications, journals, books, and news for listening/speaking/reading and writing skills. They covered most Cambridge ESOL materials (CESOL) that were available in Vietnam. Teachers and students did not use other kinds of materials, as reported in the results of the interviews and observations. The effects of CESOL-tests were evident in the official course documents, but these materials were used after 2013. The English teachers tended to use the materials from the CESOL resources to prepare students for mid-term and final term examinations and EPT.2 examination. The analysis revealed that other kinds of materials have no influence on teachers and students. This indicates that the EPT.2 and CESOL examinations have an impact on the choice of materials that the English teachers made and therefore on students' learning.

The university leaders claimed that the formative assessment and term English examinations were similar to EPT.2 and CESOL-tests (apart from the sub-writing of term examinations). However, because of time limitations and mixed-ability large sized classes, one of the four sub-tests (listening/speaking/reading/writing tests) was applied to both formative assessment and semester examinations at NUAE. The analysis showed that a part of the semester examination focused on testing the mastery of grammar and vocabulary and that the type of English test was unchanged. There were changes in the nature of the examination, and the changes in questions seemed the same as EPT.2 and CESOL item types and content (see Appendix 3). Thus, the term semester examinations were based on the EPT.2 and CESOL examinations in the four sub-tests (listening/speaking/reading/writing tests) as far as the item types and contents were concerned.

However, there were some adjustments in terms of the length of time for different sub-tests and the levels of difficulty of each test at different semester levels. This resulted in the number of questions for each sub-test being different. The scores of the examination still followed the traditional Vietnamese scale of a 10-point scale in which Point 1 was the lowest, and Point 10 was the highest. The differences in the score were indicated in the semester examinations/achievement tests and the EPT.2 and CESOL-tests as proficiency tests. To conclude, the semester examinations were EPT.2 and CESOL-based tests.

9.5.3 Results from the Questionnaire

The responses of 12 English teachers and 679 English students showed that the contents of their teaching and learning activities focused on four skills that were designed according to topics/texts/question types and activities of KNLNNVN. However, there were a few differences between teacher and student responses to teaching and learning listening/speaking/reading and writing activities in the classroom as seen in Table 9.5.

After examining the results for the TQ and SQ, the first author decided to observe what happened in the English classroom and why the English teachers and students did not give the same answers to one part of the questionnaire. The small differences in Table 9.5 might be attributed to the washback of KNLNNVN and EPT.2 on teaching and learning activities.

9.5.4 Findings from Observations

Because the KNLNNVN and EPT.2 exerted influence on EFL teaching at NUAE, the first author observed 10 English teachers to see whether washback existed in their classroom practices.

9.5.4.1 Round 1

Out of 12 English teachers, 10 agreed to be observed, and two teachers along with their students agreed to have their in-class interactions video-taped. Ten teachers were female with ten years of teaching experience. Ten teachers and their students used materials from the CESOL type. Textbooks were the third edition (*Pre-intermediate*/A2-B1) (Oxenden et al. 2012). The supplementary materials were Cambridge Key English Tests 1 and 2 (*Cambridge Key English Test 2* 2003) and English Grammar in Use of Murphy (2011). Ten teachers focused on four skills, grammar and vocabulary during Round 1 (listening: 9.2%, speaking: 15.6%, reading: 13.6%, writing: 10.2%, grammar: 17.8%, vocabulary: 26%, and pronunciation: 4%). The findings in the field notes showed that the majority of students (75%) worked in pairs or groups and made oral presentations, and then ten teachers corrected their errors sometimes (3.8%). Some students read the assigned textbook and drew some information from it. The rest (around 25%) played games on their mobile phones; they seemed reluctant to take part in teaching and learning activities and thus, they did not contribute to the discussion and other learning activities. Because of time limitations, not all students had a chance to speak English (see Appendix 1). Ten teachers used authentic materials (53.67%). The others read materials of their own choice which they designed themselves (see Appendix 2).

Table 9.5 Teacher and student responses to teaching and learning activities in the classroom (%)

Part 2	Contents and communicative methods of teaching EFL listening, speaking, reading, writing skill	Listening/M(Average)		Speaking/M(Average)		Reading/M(Average)		Writing/M(Average)	
		T	S	T	S	T	S	T	S
While-lesson activities	Topics	89.91	50	90.83	50	85.61	50	84.03	50
	Texts	76.04	53.76	73.95	50.74	72.91	52.61	70.38	55.37
	Question types and Activities	64.29	65.52	63.33	61.85	66.66	64.73	64.28	61.55
After school (Homework)	Topics	51.39	55.70	53.03	53.92	53.78	50	52.08	54.16
	Texts	50	50	50	50	50	50	50	50.66
	Question types and Activities	50	51.01	47.22	51.61	50	50.63	50	50.58
Post-lesson activities	Correct and Comment	100	50	100	51.55	100	50.37	100	60.24

Note S = Student; T = Teacher

9.5.4.2 Round 2

The first author observed ten classes of English A2 with 30 English lessons in Round 2. All the teachers observed were female, with ten years of teaching experience. They used materials from the CESOL type; they did not use textbooks. The supplementary materials were Key English Test 1, 2 (*Cambridge Key English Test 2* 2003) and *English Grammar in Use* by Murphy (2011). Ten teachers focused on practicing reading and writing (46.67%) and a written test (24%) during Round 2. Students worked in pairs or groups and made oral presentations, and then ten teachers corrected their errors sometimes (7.47% and 1.13%). Because of time limitations, not all students had a chance to speak English (see Appendix 1).

In short, the results of observations corresponded to the responses of teachers and university leaders. The ten teachers used a variety of materials from the CESOL type. These materials were in line with the EPT.2 of KNLNNVN's approach. The methodology used by these teachers was a communicative approach. It was hard to define whether the EFL teaching methodology was influenced by the EPT.2 of KNLNNVN's approach or by the methodology of the materials used. Moreover, 25% of the students did not focus on teaching and learning activities; therefore, they did not understand what they learned; their respones were different from those of their teachers. This suggests that the EPT.2 of KNLNNVN exerted influence upon EFL teaching and learning.

9.5.5 Findings from Interviews

Informal conversational interviews were conducted with nine out of 12 English teachers after four classroom observations and focus group discussions. Semi-structured interviews were also conducted with the Rector of NUAE, leaders of the Training Department and Foreign Language Center (see Appendix 4). The following are the interview findings.

100% of the English teachers had already obtained M.A. degrees at universities in Vietnam, Australia, and the USA. They had teaching experience of more than seven years; they could understand the changes in the national and institutional policies on EFL teaching and learning between 2013 and 2014. All of them often collected CESOL, EPT.2 and CESOL-type materials, which could be used in class. They also asserted that there were many practice tests for EPT.2 and CESOL examinations. They reported that they had been using the materials and practice tests because the CESOL-test materials were included in the course documents; therefore, they did not design learning tasks for students. They also expressed that they wanted their students to be familiar with numerous text types and content of the EPT.2 and CESOL-tests. This suggests that there was a relationship between teaching and learning, which in turn was related to the washback of KNLNNVN and EPT.

Interview data showed that there had been many more materials on the market that were designed to prepare for EPT.2 and Cambridge ESOL examinations. It could also be said that teachers reacted differently to the needs of the test, and self-designed learning tasks were also a problem for inexperienced teachers. The selection of the supplementary materials was an indicator of the washback impact of the KNLNNVN and EPT on the use of teaching and learning materials. Some of the teachers did not think that they taught to the tests. They claimed that they honed students' English ability. Thus, the teachers acknowledged that the EPT.2 and CESOL-tests were prevalent. In addition, 70% of the teachers said that they changed their teaching methods to meet the changes in formative assessment and semester exams.

As the Rector of NUAE reported, the number of students admitted to NUAE was increasing to meet the demands of society, and society demanded a high quality of training outcome, particularly a high level of student English language proficiency. For this reason, the assessment of EFL learning outcomes at NUAE must be improved to meet the necessities of society. The Rector asserted that he wanted to maintain the institutional policies on English language teaching according to KNLNNVN in the coming years because of its benefits.

The leaders of the Training Department and Foreign Language Center asserted that the English semester exams were geared for EPT.2 and CESOL-tests and that they were of the EPT.2-type, except for the writing sub-test and the score scheme. Furthermore, the English teachers were acquainted with EPT.2 and CESOL-tests, and they understood that EPT.2 and CESOL-tests influenced the semester EFL exams. They believed that their tests were standardized because their tests were designed based on EPT.2 and CESOL-tests. Therefore, the semester EFL exams significantly influenced the learning and teaching of EFL at NUAE. This empirical evidence showed the washback impact of the EPT.2 of KNLNNVN.

9.6 Insight(s) Gained

The responses of the students, teachers and leaders revealed that the EPT.2 and KNLNNVN had a washback effect on what the teachers taught (teaching content) and on semester exams. Most of the teachers agreed that formative assessment and semester exams corresponded to one of the EPT.2 sub-tests. Nonetheless, a few teachers acknowledged that the EPT.2 and KNLNNVN had little washback effect on what they taught. Accordingly, KNLNNVN and EPT.2 had different washback effects on some teachers and learners than on other teachers and learners. Similar empirical evidence was also reported by Alderson and Hamp-Lyons (1996).

Drawing on this empirical evidence, the KNLNNVN and EPT.2 have been considered as one of the dominant determiners of what occurred in language classrooms that influenced EFL teaching and learning activities at NUAE. The nature of those influences was direct and indirect as well as either positive or negative. For example, the positive influences of tests brought about innovations in the official language curriculum, the official course documents, instructional methods, assessment methods, and supplementary materials. However, some inexperienced teachers did not design the learning tasks for students but relied instead on the available published materials. Moreover, teachers focused on practicing grammar, vocabulary, reading, and writing skills because of time constraints that were related to negative washback effects.

To sum up, the insights from the findings show that both KNLNNVN and EPT.2 have both positive and negative impacts upon the institutional policies, curriculum, the assessment of EFL learning outcomes, and the teaching and learning of EFL for non-English majors at NUAE.

9.7 Conclusion: Practical Implications for Test Users

The findings suggest that the Vietnam Ministry of Education and Training should issue a set of pre-constructed English tests that are based on EPT or CESOL-tests. Afterwards, all schools would draw from this set to design a new version that is adapted to suit their own EFL teaching and learning context. In addition, university teachers should be trained in how to design and administer educational assessments and tests that help them to include test items and tasks which suit their students' needs for EFL tests. There should also be solid collaboration among Vietnamese policy-makers, language educators, and test writers, test users, English teachers in preparing educational assessments and tests that can have the beneficial washback effects of KNLNNVN and EPT.2 upon the design and enactment of language curriculum, instruction, and assessment as a whole. The findings reported in this chapter have contributed to the knowledge of the nature of washback and provide a better understanding of ways to identify different levels of washback effects using different empirical data. With this mind, the washback effects of tests could be positive if such tests are tailored to students' needs for their further studies or future employment because the students' needs could be a catalyst for increasing students' motivation to learn. As Tsai and Tsou (2009) emphasize, stakeholders' views/perspectives (e.g., learners and university teachers) should be viewed as the basis for needs analysis in order to know whether the adoption of CEFR-oriented ELP tests as a tool for assessing university students' English competence for graduation can work well and have a positive washback effect on learning and teaching as a whole.

Appendix 1

Participant Organization: Percentages of Observation Time by VISIT 1 (%)

Participation Organization	Whole class (Interaction)			Student			Teacher			The medium of instruction		Total (%)
	T > S/C	**S > S/C**	Choral work	individual	pair	group	Pre-lesson activities	While-lesson activities	Post-lesson activities	English	Vietnamese	
T1	46	54	6	22	12	14	10	86	4	50	100	100
T2	40	60	0	10	20	30	40	58	2	40	100	
T3	50	50	0	30	0	20	40	58	2	40	100	
T4	44	56	10	0	14	32	20	76	4	20	100	
T5	32	68	8	46	14	0	20	76	4	20	100	
T6	48	52	0	14	0	38	10	88	2	20	100	
T7	52	48	0	38	0	10	20	78	2	20	100	
T8	48	52	0	42	4	6	16	80	4	40	100	
T9	46	54	10	16	12	10	10	86	4	50	83	
T10	38	62	0	22	10	30	30	88	10	60	100	
M (Average)	44.40	55.6	3.4	24.00	8.60	19.00	21.60	77.40	3.80	36.00	98	

Contents: Mean Percentages of Observation Time by VISIT 1 (%)

Teacher	Topics	Texts	Question types	Listening	Speaking	Reading	Writing	Grammar	Vocabulary	Pronunciation	Homework	Total (%)
T1	Food and drink	Pictures	Word completion	0	0	0	0	0	50		4	100
				0						20		
		Conversation	Multiple-choice	26						0		
T2	Free time and entertainment	Pictures	Word completion	0	0	0	0	0	34	0	2	
		Dialogues	Gap-fill		0			24				
			Open questions		40			0	0			
T3	Health and weather	Journal articles	Completing forms	0	0	0	0	70	0	0	4	
		Pictures	Guided writing				26	0				
T4	Transportation	Pictures	Gap-fill	0	0	0	0	20	20	0	4	
		Dialogues	Word completion	20	0			0	0			
			Role-play	0	36			0	0			
T5	Free time and entertainment	Pictures	Matching	0	0	0	0	40	0	0	2	
			Completing forms		0		0					
			Open questions		40		0	0				
			Guided writing		0		18	0				
T6	Shopping	Broadcast	Open questions		0	0	0	0	78	0	2	
		Pictures	Matching	0								
			Word completion	0								

(continued)

(continued)

Contents: Mean Percentages of Observation Time by VISIT 1 (%)

Teacher	Topics	Texts	Question types	Listening	Speaking	Reading	Writing	Grammar	Vocabulary	Pronunciation	Homework	Total (%)
		Dialogues	Open questions	20					0			
T7	Work and Jobs	Pictures	Matching	0	0	70	0	0	0	0	2	
			Guided writing				28					
T8	Work and Jobs	Pictures	Matching	0	0	66	0	0	0	0	4	
			Guided writing				30					
T9	Food and drink	Pictures	Multiple-choice	26	0	0	0	0	50	20	4	
T10	Free time and entertainment	Dialogue	Gap-fill	0	0	0	0	24	0	0	2	
			Open questions		40			0	0			
		Pictures	Word completion		0			0	34			
M (Average)				9.2	15.6	13.6	10.2	17.8	26.6	4.0	3.0	

Appendix 2

10 Teacher’s Use of Materials Used in the EFL Class: Percentages of Observation Time by VISIT 1 (%)

Materials	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Total (%)	Mean (Average)
Global (Elementary)	10	10	0	0	0	0	0	0	10	10	100	4.00
English File (Pre-intermediate)	10	0	10	10	10	10	10	10	10	0		8.00
New English File (Pre-intermediate)	0	0	10	0	0	0	0	0	0	0		1.00
New Headway 3rd Ed (Pre-intermediate)	0	0	0	10	0	0	0	0	0	0		1.00
Solutions (Pre-intermediate)	0	0	0	0	0	10	0	10	0	0		2.00
Lifeline (Pre-intermediate)	10	0	0	0	0	0	0	0	10	0		2.00
Videos	10	0	0	0	0	0	0	0	10	0		2.00
CD	10	0	0	10	0	10	0	0	10	0		4.00
Pictures	10	10	10	10	10	10	10	10	10	10		10.00
Journal articles	0	0	10	0	0	0	0	0	0	0		1.00
Laptop	10		10	10	10	10	0	0	10	0		6.67
Projector	10	0	10	10	10	10	0	0	10	0		6.00
Blackboard	0	10	10	10	10	10	0	0	0	10		6.00
Mean (Average)	6.15	2.50	5.38	5.38	3.85	5.38	1.54	4.23	6.15	2.31		53.67

Appendix 3

Participant Organization: Percentages of Observation Time by VISIT 2 (%)

Participation Organization	Whole class (Interaction)			Student			Teacher			The medium of instruction		Total (%)
	T > S/C	**S > S/C**	Choral work	individual	pair	group	Pre-lesson activities	While-lesson activities	Post-lesson activities	English	Vietnamese	
T1	14.67	12.00	0.00	73.33	12.00	0.00	10.00	88.67	1.33	40	100	100
T2	21.33	5.33	0.00	73.33	5.33	0.00	13.33	85.33	1.33	40	100	
T3	20.00	6.67	0.00	73.33	6.67	0.00	10.00	88.67	1.33	40	100	
T4	21.33	5.33	0.00	73.33	5.33	0.00	13.33	85.33	1.33	40	100	
T5	14.67	12.00	0.00	73.33	12.00	0.00	14.67	80.00	1.33	40	100	
T6	28.00	5.33	0.00	66.67	5.33	0.00	28.00	72.00	1.33	40	100	
T7	28.00	5.33	0.00	66.67	5.33	0.00	18.00	80.67	1.33	40	100	
T8	27.33	6.00	0.00	66.67	6.00	0.00	23.33	76.00	0.67	40	100	
T9	22.00	11.33	0.00	66.67	11.33	0.00	13.33	86.67	0.00	50	100	
T10	21.33	5.33	0.00	73.33	5.33	0.00	13.33	81.33	1.33	40	100	
M (Average)	21.87	7.47	0.00	70.67	7.47	0.00	15.73	82.47	1.13	41.00	100	

Contents: Percentages of Observation Time by VISIT 2 (%)												
Teacher	Topics	Texts	Question types	Listening	Speaking	Reading + Writing	Grammar	Vocabulary	Pronunciation	Written Test (Multiple choices + essay)	Homework	Total (%)
T1	Entertainment and media; The natural world; House and home; Places and building; Sport; Hobbies and leisure.	Notices Storyline Sentences Newspaper Magazine Articles A short letter or email. Notes, adverts	Matching; Multiple choice; Multiple choice Close; Word completion; Sentences completion; Open close; Information transfer; Guided writing.	0	0	46.67	0	0	0	26.67	1.33	100
T2				0	0	46.67	0	0	0	26.67	1.33	
T3				0	0	46.67	0	0	0	26.67	1.33	
T4				0	0	46.67	0	0	0	26.67	1.33	
T5				0	0	46.67	0	0	0	26.67	1.33	
T6				0	0	46.67	0	0	0	20.00	1.33	
T7				0	0	46.67	0	0	0	20.00	1.33	
T8				0	0	46.67	0	0	0	20.00	0.67	
T9				0	0	46.67	0	0	0	20.00	0.00	
T10				0	0	46.67	0	0	0	26.67	1.33	
M (Average)				0.00	0.00	46.67	0.00	0.00	0.00	24.00	1.13	100

Materials: Percentages of Observation Time by VISIT 2 (%)

Materials	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Total (%)	Mean (Average)
KET[a] 1	10	0	0	0	0	0	0	0	10	0	100	2.00
KET[a] 2	0	0	10	0	0	0	0	10	0	0		2.00
KET[a] 3	0	0	0	10	0	10	0	0	0	10		3.00
KET[a] 5	0	10	0	0	0	0	0	0	0	0		1.00
KET[a] 7	0	0	0	0	10	0	10	0	0	0		2.00
New English File—Test Booklet (Pre-intermediate)	0	10	10	0	0	0	0	0	0	10		3.00
Written Test	10	10	10	0	10	10	10	10	10	10		9.00
English Grammar in Use	0	0	0	10	0	10	10	0	0	0		3.00
laptop	10	10	10	10	10	0	0	0	0	10		6.00
projector	10	10	10	10	10	0	0	0	0	10		6.00
Blackboard	10	10	10	10	10	10	10	10	10	10		10.00
Mean (Average)	4.55	5.45	5.45	4.55	4.55	3.64	3.64	2.73	2.73	5.45		4.27

[a] The Key English Test

Appendix 4

Q1: How do you feel KNLNNVN influences your teaching methodology and materials?

Q2: How do you feel KNLNNVN influences your institutional policies on EFL teaching and learning?

Q3: How do you feel KNLNNVN influences the semester examinations of English at NUAE?

References

Afip, L. A., Hamid, M. O., & Renshaw, P. (2019). Common European framework of reference for languages (CEFR): Insights into global policy borrowing in Malaysian higher education. *Globalisation, Societies and Education, 17*(3), 378–393. https://doi.org/10.1080/14767724.2019.1578195.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280–297. https://doi.org/10.1177/026553229601300304.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129.

Bailey, K. M. (1999). *Washback in language testing*. Princeton, NJ: Educational Testing Service.

Boyce, A. (2010). *The effectiveness of increasing language learning strategy awareness for students studying English as a second language.* https://pdfs.semanticscholar.org/5365/d86a8ff2668fb0263f04d03180c22de470e1.pdf. Accessed 2 August 2019.

Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and Teacher Education, 15*(3), 253–271. https://doi.org/10.1016/S0742-051X(98)00046-8.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 146–170). Mahwah, NJ: Lawrence Erlbaum.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 3–17). Mahwah, NJ: Lawrence Erlbaum.

Coste, D. (2007). *Contextualising uses of the Common European Framework of Reference for Languages*. Paper presented at Council of Europe Policy Forum on use of the CEFR, Strasbourg, 2007. http://www.coe.int/T/DG4/Linguistic/Source/SourceForum07/D-Coste_Contextualise_EN.doc. Accessed 25 September 2008.

Council of Europe. (2001). *Common European framework of reference for languages*. Cambridge: Cambridge University Press.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cambridge Key English Test 2. (2003). Cambridge: Cambridge University Press.

Đinh, T. P. H. (2019). Washback of The Vietnam Six-levels of Foreign Language Proficiency Framework on teaching English as foreign language for non-English major students at National University of Arts Education. In R. K. Prabhakara (Ed.), *Proceeding of the 5th International Conference Language, Society, and Culture in Asian Context 2018* (pp. 1550–1578). Jakarta, Indonesia: MNC Publishing.

Đinh, T. P. H. (2017). Washback in language teaching and learning worldwide and Vietnam: The results of literature review. *The Vietnam Journal of Art Education, 23,* 51–57.

Fitz-Gibbon, C. T. (1996). *Monitoring education: Indicators, quality and effectiveness.* London: Cassell.

Forbes, D. (1973). Selling English short. *English Language Teaching Journal, 27,* 132–137.

Hinchey, P. H. (2008). *Action research in education.* USA: Peter Lang.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Jones, N., & Saville, N. (2009). European language policy: Assesment, learning and the CEFR. *Annual Review of Applied Linguistics, 29,* 51–63. https://doi.org/10.1017/S0267190509090059.

Low, G. D. (1988). The semantics of questionnaire rating scales. *Evaluation and Research in Education, 22,* 69–79.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: MacMillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241–256.

Murphy, R. (2011). *English grammar in use.* Oxford: Oxford University Press.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2012). *English file (Pre-intermediate).* Oxford: Oxford University Press.

Pan, Y., & Newfields, T. (2012). Tertiary EFL proficiency graduation requirements in Taiwan: A study of washback on learning. *Electronic Journal of Foreign Language Teaching, 9*(1), 108–122.

Pearson, I. (1988). Tests as levers for change. In D. Chamberlain & R. J. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (pp. 98–107). London: Modern English.

Pham, T. N., & Bui, L. T. P. (2019). An exploration of students' voices on the English graduation benchmark policy across Northern, Central and Southern Vietnam. *Language Testing in Asia, 9*(15), 1–20. https://doi.org/10.1186/s40468-019-0091-x.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal, 76*(4), 513–521.

Shohamy, E. (2000). Using language tests for upgrading knowledge. *Hong Kong Journal of Applied Linguistics, 5*(1), 1–18.

Shohamy, E. (2014). *The power of tests: A critical perspective on the uses of language tests.* New York: Routledge.

Sims, J. M., & Chen, S. (2019). Court ruling on the English benchmark requirement for graduation in Taiwan. *The Journal of Asia TEFL, 16*(1), 345–348. https://doi.org/10.18823/asiatefl.2019.16.1.23.345.

Streatfield, D., & Markless, S. (2009). What is impact assessment and why is it important? *Performance Measurement and Metrics, 10*(2), 134–141.

Tsagari, D., & Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment: Encyclopedia of language and education* (3rd ed., pp. 359–372). Cham: Springer.

Tsai, Y., & Tsou, C. (2009). A standardised English Language Proficiency test as the graduation benchmark: Student perspectives on its application in higher education. *Assessment in Education: Principles, Policy & Practice, 16*(3), 319–330. https://doi.org/10.1080/09695940903319711.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10,* 41–69.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13,* 318–333.

Widodo, H. P. (2015). *The development of Vocational English materials from a social semiotic perspective: Participatory action research.* Unpublished Ph.D. thesis, The University of Adelaide, Australia.

Wu, J., & Lee, M. C.-L. (2017). The relationships between test performance and students' perceptions of learning motivation, test value, and test anxiety in the context of the English benchmark requirement for graduation in Taiwan's universities. *Language Testing in Asia, 7*(9), 1–21. https://doi.org/10.1186/s40468-017-0041-4.

Chapter 10
Avoiding Scoring Malpractice: Supporting Reliable Scoring of Constructed-Response Items in High-Stakes Exams

Kristina Leitner and Benjamin Kremmel

Abstract Scoring reliability of constructed-response items is a key concern in high-stakes testing. Constructed-response items, often used for their authenticity, potentially allow for a multitude of acceptable answers that were neither intended nor anticipated, and can therefore be problematic for reliable scoring. This chapter examines the use of a specially developed marker support system for the Austrian EFL school-leaving exam, which uses such items but without centralized marking and therefore potentially suffers from inconsistent scoring that could affect 40,000 students annually. The study investigates the impact of three different scoring guide conditions on test taker results in four constructed-response tasks for listening at CEFR B2 level. The first scoring condition (A) is exact scoring based on the scoring guide developed by the item writing team before the task had been field-tested. The second scoring condition (B) is based on an extended scoring guide that was improved in a centrally run scoring session after piloting the items. The third scoring condition (C) is based on the highly comprehensive scoring guide that was enhanced during the scoring of the national live exam through a marker support system in the form of an online helpdesk and a telephone hotline. The statistical analyzes show an overall improvement in the reliability of the test from scoring condition A to scoring condition C. Consequently, the findings of the study suggest that the practice of improving and refining the scoring guides through the implemented marker support system increase the comparability, reliability, and fairness in test taker scores.

K. Leitner (✉)
Austrian Federal Ministry of Education, Science and Research, Vienna, Austria
e-mail: kristina.leitner@bmbwf.gv.at

B. Kremmel
University of Innsbruck, Innsbruck, Austria
e-mail: benjamin.kremmel@uibk.ac.at

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_10

10.1 Introduction: Purpose and Testing Context

The reliability of scoring procedures is essential to any kind of test in order to guarantee that test takers receive fair results. It is the scores on a test from which inferences are drawn, and based on which important decisions are made. Consistency of assessment records is therefore crucial, particularly in high-stakes exams. However, in exam contexts where no centralized marking is employed, the consistency of scoring procedures may be compromised and the fairness of the test thus jeopardized. While many tests rely on objectively scorable selected-response test formats for exactly this reason, this chapter discusses the issue of scoring reliability in the context of the high-stakes secondary school-leaving exam in Austria that also features items for which a subjective scoring decision needs to be made.

This chapter will focus on such constructed-response items for listening tasks, commonly referred to as short-answer questions (Buck 2001), open-ended items (Harding et al. 2011; Eberharter and Frötscher 2013), or limited production response (Bachman and Palmer 2010). The main characteristic of this item type is that test takers have to come up with answers themselves, within a restricted number of words. Open-ended items can be designed to test a number of different listening skills or “language operations” (Brindley 1998, p. 172), such as understanding main ideas, listening for specific information, and inferring the speaker’s meaning and are therefore among “the more commonly used formats” (Brindley 1998, p. 177). Despite their usefulness, however, constructed-response items potentially allow (1) a broader scope of acceptable answers and therefore assessors need to “apply their own interpretations of the construct in judging responses that fall outside the information provided in a marking guide” (Harding et al. 2011, p. 108) and (2) the marking of open-ended items tends to be “more resource-hungry, in terms of availability and management of suitably trained personnel and the time needed for marking” (Taylor and Geranpayeh 2011, p. 97). These issues concerning the marking of constructed-response items may have an impact on marker consistency, as well as on the resulting test scores that are awarded to test takers.

These issues are particularly pertinent in high-stakes exams. The Austrian school-leaving examination, or Matura, which is the context for the issue and research presented in this chapter, employs, among numerous other formats, such a limited production response format for the receptive skills. The exam itself has undergone a major reform since 2007 and has now become obligatory for all secondary schools. The new Austrian national standardized school-leaving examination (*Standardisierte Reife-und Diplomprüfung*, hereafter SRDP) is a high-stakes proficiency test that allows test takers to enter university and is, like the Austrian national curriculum itself (BMUKK 2004), based on the Common European Framework of Reference. The SRDP is a centrally developed, CEFR-linked exam that is delivered to all test takers across Austria on a common date (Spöttl et al. 2018; Weiler and Frötscher 2018). However, although the SRDP is centrally developed and administered, for a number of (mainly political) reasons, the marking is still conducted by individual class teachers (Spöttl et al. 2016), creating potential issues with reliability and thus a major challenge for the test developers.

10.2 Testing Problem Encountered

Constructed-response items, often used for their authenticity, potentially allow for a multitude of acceptable answers that were neither intended nor anticipated, and can therefore be problematic for reliable scoring. As the Austrian EFL school-leaving exam uses such items but without centralized marking, the exam therefore potentially suffers from inconsistent scoring that could unfairly affect 40,000 students annually. The test development team addressed this issue by (a) providing extended scoring guidelines based on trial responses of test takers to the class teachers, who are themselves responsible for grading the exam papers of their own students, and (b) setting up a comprehensive marker support system in the form of an online helpdesk and a phone hotline that extends the answer keys further during the marking period of the live examination to support examiners when they are faced with unexpected responses. This chapter examines the usefulness and impact of these two measures on scores and their reliability and thus fairness.

All tasks that are considered for use in the national exam are developed following internationally accepted standards and have to pass a number of quality control procedures in a rigorous test development cycle (Spöttl et al. 2018). At three stages in this cycle, scoring guides are developed and enhanced for these open-ended items.

Pre-trial, item writers develop an initial scoring guide, adapting it depending on peer and expert item moderator feedback. This version of the scoring guide usually only comprises the intended model answers from the sound file that the item writers are targeting.

Post-trial, central correction is the second stage, at which the scoring guide is developed and improved further. Trialing open-ended items is key to ensure that these items are not ambiguous and to detect alternative answers that were not foreseen by the task developers (Buck 2001). Based on unforeseen responses from trial data (N > 100) that go beyond the model scoring guide, item writer teams, item moderators, and native speakers enhance the original scoring guide in centrally run sessions as recommended by Alderson (2000). Assessors receive detailed recommendations for making decisions on these responses to ensure consistent decision-making across multiple teams. In accordance with Alderson (2000), markers are, broadly speaking, instructed “not to penalise responses which show an understanding of the text and task, but are expressed in ‘incorrect’ language” (p. 199). The recommendations include instructions on accepting spelling variations that appear to resemble the correct response phonetically, and on accepting misspellings that resemble the target answer phonetically but spell another word only if the meaning is unambiguous in the context of the item. If both meanings, the correct answer and the misspelled one, lead to ambiguity in the given context and it is not clear whether the candidate has actually understood the correct answer, they are instructed to mark the response as incorrect. Assessors are not allowed to penalize grammatical errors as long as the communicative function between a candidate’s answer and the marker is not impeded, that is, the answer is sufficiently correct and the morphological inadequacy does not change the meaning. Each marking team is instructed to list all acceptable answers that deviate from the correct one. They are also asked to keep a record of answers not accepted (except for utterly meaningless answers). Each team identifies answers that fall outside the original scoring guide, discusses these within their group, and decides on a principled case-by-case basis whether a test taker's response is acceptable or not. The overall decisions taken are carefully recorded in the scoring guide, which tends to expand considerably. Discussions are an important means in the decision-making process to ensure more reliable marking.

The third stage of scoring guide expansion takes place during the live administration at the meetings of the helpdesk and hotline marker support system. As mentioned above, although Austria has introduced a centralized exam with standardized, externally developed tasks, the marking of student papers is still in the hands of the class teachers. This particular situation poses a "threat to the reliability of the marking and even the resulting test scores" especially because "experience has shown that responses not anticipated occurred frequently [despite the fact that the] marking guides are based on actual test taker responses from the field tests ($N < 100$) and include justifications written by the item writers" (Eberharter and Frötscher 2013, p. 236). Therefore, a marker support system in the form of an online helpdesk and a telephone hotline has been set up. This innovative system was acknowledged with the Innovation in Assessment Award 2013 by the British Council. All teachers involved in marking can access this support system and query an expert team on whether an unexpected answer should be accepted or not. The expert team of test developers and trained native speakers collects all enquiries, systematically answers them, and checks for consistency before all decisions are entered in a database. This aims to ensure consistent decision-making and enhanced reliability. See Eberharter and Frötscher (2013) for more details on how the support structures have been implemented.

While the central correction sessions and the marker support system, and thus the stepwise enhancement of scoring guides, were set up with the intention to avoid scoring malpractice and address potential inconsistency and unfairness in marking, it is yet to be demonstrated empirically to what extent such enhanced scoring guides do improve the reliability of scores. The study presented in this chapter therefore seeks to investigate to what extent different marking procedures have an impact on test taker results by tracing scoring guides for short-answer items from their inception at the task development stage to live administration, and to analyze whether the quality control procedures and support structures that have been created have actually helped to improve reliability of test scores of English listening tasks at the stipulated level B2 of the Common European Framework of Reference (CEFR) (Council of Europe 2001). If these measures taken are actually ensuring fairer and more reliable test scores, then this would be convincing evidence that the time and money invested in the support structures is well-justified. We thus attempted to compare three scoring conditions, (A) the pre-trial stage, (B) the post-trial or central correction stage, and (C) the post-live administration or post-helpdesk and hotline stage, and investigated whether exact marking versus marking with extended scoring guides results in a significant difference in test taker results of English constructed-response tasks. The study tries to answer the following research questions:

RQ1: Which differences can be detected at the score and item level when four English listening constructed-response tasks at CEFR B2 level of the national standardized Austrian school-leaving examination are marked with (A) a rigid scoring guide from the pre-trial stage, (B) an enhanced post-trial central correction scoring guide, and (C) a further extended post-live administration scoring guide?

RQ2: Which differences can be detected in test reliability when the same four English constructed-response tasks are marked with (A) a rigid scoring guide, (B) an enhanced post-central correction scoring guide, and (C) a further extended post-live administration scoring guide?

RQ3: Do the three different scoring conditions measure the same construct?

10.3 Review of Literature

This section attempts to discuss critically the use of constructed-response items and approaches to their marking. Constructed-response format items such as short-answer questions or note-form items require a test taker to complete a gap or to respond to a question with a stipulated number of words: in Austria, a maximum of four words. According to Field (2013), this item format has a number of advantages. The primary one is that it reflects more authentic listening behavior as note-taking or listening to answer a specific question is "closer to real-life" (p. 167). Also, it does not provide the test takers with various options from which they need to select the appropriate answer; therefore, such items have less "effect on the cognitive processes involved in listening" (Elliott and Wilson 2013, p. 167). Furthermore, it can be assumed that a candidate's answer is based on what the test taker has actually understood (Alderson 2000); hence the guessing factor or the deduction of an answer by way of eliminating other options can largely be avoided. Constructed item types are also deployable in testing a variety of different constructs, which makes them a versatile instrument to "measure specific areas of language knowledge, as well as comprehension in receptive language use tasks" (Bachman and Palmer 2010, p. 335).

Despite the positive aspects of this response type, short-answer items also entail certain risks. Not only do they require the test taker to read and comprehend an item correctly, but they also introduce an element of writing, which may lead to construct-irrelevant variance, which is often quoted as the main disadvantage of this item type (e.g., Buck 2001). Brown and Yule (1983a, as cited in Lynch 2009) identify four reasons why test takers might not give a correct answer even though they were able to understand the relevant part of the sound file.

- The test takers have misunderstood the question, which is a problem of reading.
- The test takers have made a slip in their answer, which is a problem of writing.
- The test takers may not have noticed a specific detail required for a correct answer to an item, which is a problem of attention.
- The test takers may have forgotten what they heard and understood, which is a problem of memory (p. 123).

According to Brown and Yule (1983), the effects of memory load can be reduced if test takers are allowed to answer questions during the listening. Lack of attention can be avoided if the questions reflect a realistic target as in "what you would expect a competent native listener to have understood or noticed" (Lynch 2009, p. 123). One way of reducing the effects of writing and reading is to minimize the amount of the second language that needs to be produced or processed (Brown and Yule 1983). If questions are kept short and simple and elicit short and unambiguous answers, other abilities, such as reading the items and writing responses, will not impact greatly on the construct being tested and construct-irrelevance can be kept to a minimum (Weir 2005). Weir argues that "if all candidates have equal ability in these other skills then scores should not be affected" (p. 137). All of these concerns are taken into account in the context of the present study.

However, Buck (2001) lists two more potential problems when assessing listening through short-answer items, especially when understanding on a deeper level is required, which he defines as listening that goes beyond the "superficial understanding of clearly stated information." He questions "what constitutes a reasonable interpretation of the text" and "what constitutes a sufficient response to the question" (p. 140). Bachman and Palmer (2010) elaborate on this when they discuss two implications for scoring. One is the increased difficulty of ascertaining that the marking criteria correspond to the construct definition; the other is the need to develop detailed scoring guides, which will list any responses that are considered acceptable. If such correction guides are not supplied, it may become necessary for scorers to apply their personal judgment while marking, which consequently could constitute a potential source of inconsistency. Therefore, some researchers (e.g., Alderson 2000; Buck 2001; Weir 2005) strongly suggest exhaustively piloting the items to comprise scoring guides that are as comprehensive as possible and that "should ideally be trialed and refined through an iterative development process to include a variety of acceptable responses" (Harding et al. 2011). The study presented in this chapter is an attempt to document the impact of such an iterative scoring guide development process.

In making decisions for such scoring guides, some issues are prone to cause discussions among developers. Elliott and Wilson (2013) identified spelling as the most problematic area because there are several ways of dealing with misspelled responses. The researchers list three main policies which may be adopted regarding spelling:

- Accepting all plausible phonetic misspellings of a word.
- Accepting a limited, prescribed range of misspellings of a word and no others.
- Accepting only the correct spelling of a word (p. 168).

Elliott and Wilson (2013) state that "spelling does not form part of a narrowly defined construct of listening" (p. 168) and hence different scores due to misspelling could introduce construct-irrelevant variance. Candidates with limited literacy skills will be disadvantaged by a strict spelling policy. A further argument supporting the acceptance of spelling variations is the cognitive load that is demanded from candidates as they focus on listening and not on writing during the limited time span they have to answer an item during a listening test (Elliott and Wilson 2013, p. 169). Therefore, more liberal scoring guides regarding spelling seem fairer and more valid in terms of the construct being tested because test takers are not penalized for their writing skills as long as their answers correspond sufficiently to the key answer.

However, if the approach of a liberal scoring guide is adopted, markers are faced with a number of issues regarding the adequacy of a response and of accepting incorrectly spelt answers. Subjective marker decision-making might pose a threat to the reliability of the marking. Limiting the number of words a candidate is allowed to write reduces the range of possible correct answers, and hence contributes to consistency. Furthermore, the more test takers have to write, the greater the effect of writing on the item and the threat of introducing construct-irrelevance.

Hackett et al. (2006) compiled a useful list of potential issues regarding spelling. First, they mention the scope of interpretation as to what resembles the correct answer phonetically. Different markers may have different understandings of which variations are still acceptable. Secondly, a misspelled answer may be phonetically acceptable, but the actual letters create a different word; again, it is up to individual markers to decide whether a candidate has answered an item correctly or not. A third issue refers to the question whether a test taker has unmistakably understood the answer to an item or has simply tried to reproduce a number of sounds they have heard.

Clear guidelines should therefore be developed so that markers are aware of these issues and can react accordingly. Before a marking session is started, markers need to develop a common understanding of what can constitute an acceptable answer and discuss possible variations of the key answer. If a target answer is likely to be misspelled and the misspelling becomes another word, markers need to know which policy they have to adopt. One possibility may be to accept the answer, although the misspelling has led to another word, if the meaning of the answer is unambiguous in the context of the item. However, if the answer that has been caused by misspelling leads to ambiguity, it should be rejected.

Moreover, some of the issues discussed by Hackett et al. could also be counteracted already at the test development stage. Measures to avoid the stated problems could include selecting items as responses that are less likely to cause spelling problems or be confused phonetically; or choosing words that are below the targeted language level. It can be assumed that this way the answer is a more high-frequency word or chunk that the test taker should be familiar with at the target level and has therefore mastered writing it, if not yet correctly, at least acceptably.

The Cambridge ESOL Main Suite Listening papers follow a related approach. Elliott and Wilson (2013) state that for the Cambridge listening exams “nouns tend to be used more often as keys” because it is assumed that “nouns are easier for candidates to identify than other word forms”; in addition, specific, concrete information, which is the focus of the tasks, is “most likely to be communicated via nouns and noun phrases” (p. 179). Scoring guides can be kept reasonably restricted in this way and, since candidates are generally not asked to write words above the targeted level, incorrect spelling may be less of a problem. However, such an approach may be at the expense of not exploiting the full range of the listening construct that a test seeks to cover. Especially when testing advanced levels, selecting lower level words or phrases may be found to be too limiting. For example, it might not reflect the targeted main idea or may not adequately capture the intention of the speaker. Ultimately, whatever approach is adopted, Taylor and Geranpayeh (2011) stress that test takers need to know the "criteria against which their performance will be judged so they can dedicate resources (e.g., monitoring/checking time) to this if necessary" (p. 97).

These drawbacks to open-ended questions need to be considered carefully when the marking is approached because each policy may involve a "trade-off between validity in terms of the listening construct on the one hand and fairness and reliability of marking on the other" (Elliott and Wilson 2013, p. 168).

To date, little research has been conducted on the effects of different marking guides on test takers' scores and applying marking guides in the assessment of listening, particularly with regard to short-answer items. Harding et al. (2011) have made an important contribution to the field by analyzing decision-making processes on the part of the assessors while marking open-ended items for a specific purpose English listening test. In their study flexible scoring guides are analyzed that "arguably allow for a scoring procedure that captures more accurately a candidate's ability to listen" because assessors are asked to weigh carefully whether "an answer indicates an *appropriate* [italics added] response to the question" (p. 112). Their study is in part a replication and continuation of an earlier one carried out by Harding and Ryan (2009), in which the researchers have identified three broad categories of decisions that need to be made during the assessment process:

1. Decisions regarding spelling
2. Decisions regarding the correctness of an overelaborate response
3. Decisions regarding the adequacy of response
   This third category is further divided into:

   a) semantic distinction
      "Making a decision about whether an alternate word or phrase in a response demonstrated understanding of what the speaker had said, or whether it showed understanding of a different concept" (Harding and Ryan 2009, p. 107)
   b) sufficiency of answer
      "Making a decision about whether enough information was included in an alternate answer to sufficiently match the idea represented by the answer in the marking guide" (Harding and Ryan 2009, p. 107).

The findings of these two studies have influenced the development and improvement of quality control procedures in the task development cycle of the Austrian school-leaving examination, which is the focus of a progress report by Eberharter and Frötscher (2013). The researchers discuss the challenges that test developers are faced with when working with open-ended test items and propose measures to improve the reliability of marking. They have extended the scoring guidelines in place, based on Harding and Ryan's (2009) work, and added an item analysis grid that is completed by the assessors during the actual marking. The grid has been designed to identify problematic items after the piloting phase and thus provides qualitative information in addition to the quantitative data on item performance. Moreover, Eberharter and Frötscher (2013) delineate ways of supporting external markers (e.g., teachers) during the correction phase after the live administration.

Hackett et al. (2006) investigated the impact of the revision of an FCE 4 productive task mark scheme and found that working with a liberal scoring guide resulted in minor increases in item facility and minor increases in item discrimination when compared to the results of marking with a more stringent guide on orthography.

While Harding et al. (2011) focus on thought processes and assessor decision-making by means of a qualitative approach, namely stimulated recalls, and Eberharter and Frötscher (2013) consider ways of improving existing structures and standardizing the decision-making during the actual marking process, the present study seeks to examine whether adopting different scoring guides will impact on the psychometric properties of constructed-response tasks. In this respect it is similar to the study conducted by Hackett et al. (2006), but it will not focus only on degrees of leniency toward spelling, but rather compare scoring guides that differ in their degrees of comprehensiveness of acceptable answers, as they emerged at three different stages of a real high-stakes testing context.

10.4 Methodology

This section describes the methodological procedure that has been used to collect and analyze the data and provides information about the population sample. The research instruments are outlined and the three types of scoring guides that were applied are discussed; in addition, the actual process of marking will be explained in more detail. After this, the methods of statistical analysis will be presented briefly to facilitate the understanding of the subsequent results.

10.4.1 Participants

To address the three research questions, a quantitative research approach was chosen. The participants were pupils in their penultimate year of upper secondary education in Austrian secondary grammar schools. Usually, pupils take their final exam at the age of 18 when they should have reached a stipulated CEFR B2 level in English according to the Austrian national curriculum (BMUKK 2004). The trial took place during the penultimate week before the summer break. Although this random population sample did not have the full four years of upper secondary schooling, it was very close to the official trial population. The test booklet was administered under standardized trial administration conditions at six different schools in four different provinces of Austria, resulting in data from a random sample of 142 pupils. The test takers completed the listening test in approximately 40 min. The sound files of each task were played twice, as is standard in the live exam.

Table 10.1 Overview of the listening comprehension test

Task	1	2	3	4	Total
Sound file length	04:10	03:45	03:08	02:15	35:20
No of items	10	9	7	6	32
Mean FV	60.1%	66.3%	52.8%	53.4%	58.1%

10.4.2 Research Instrument: Listening Test

The four short-answer listening tasks that were selected for the purpose of this study were part of previously administered CEFR-linked listening test booklets of the Austrian standardized school-leaving examination between 2010 and 2013 at CEFR level B2. All tasks that are developed and considered for the SRDP are based on authentic recordings. Table 10.1 gives an overview of their length and number of items, as well as the tasks' mean facility values obtained under the most sophisticated and enhanced scoring condition (C). The Cronbach alpha value for the entire test booklet was .901.

10.4.3 Scoring Procedures

Since the same answer sheets needed to be marked with three different scoring guides, the 142 answer sheets that were filled in and returned to the researcher were photocopied. This way the three sets of answer sheets could be marked separately on paper on three different occasions, one per scoring guide. Responses were recorded as zero (incorrect answer), one (correct answer), and nine (missing answer) and entered into SPSS. In turn, each set of responses was marked using each of the three scoring conditions, or scoring guides, A, B, and C.

10.4.3.1 Scoring Guide A: Pre-trial

The first round of marking followed the principles of exact marking; hence, only the original responses listed in scoring guide A were accepted. Any deviation from the key either in terms of incorrect spelling or grammar was marked as incorrect. Capitalization was also taken into account.

10.4.3.2 Scoring Guide B: Post-central Correction

In addition to the model answers listed in scoring guide A, scoring guide B contained all decisions that had been taken during central correction of these already administered tasks. The procedure at central correction is to "accept spelling variations and misspellings, to accept abbreviations […], to disregard grammatical errors that do not affect meaning, and to consider synonyms or alternative phrasings of answers" (Harding et al. 2011, p. 112) as long as an answer clearly reflected the key answer.

10.4.3.3 Scoring Guide C: Post-live Administration

The final round of marking was the most time-intensive marking session, as the post-live administration scoring guides are comprehensive lists where not only all misspellings deemed acceptable are listed, but also any unacceptable response variations. These extensive scoring guides can be confusing and markers need to pay particular attention to keep mistakes to a minimum because extensive marking instructions tend to become complicated and difficult to handle and therefore prone to error (Hackett et al. 2006).

10.4.4 Methods of Data Analyzes

Classical test theory was employed to calculate and compare facility values and reliability for all four tasks in the three different scoring conditions. The difference between test taker results (total scores per test taker per scoring condition) was calculated in order to answer the research questions, namely, to what extent the test scores and the reliability of the tasks differ depending on the scoring condition employed. The skewness and kurtosis ratio was calculated to check for normality of distribution, and the results were corroborated by running the Shapiro–Wilk test of normality. Since the assumption of parametric data was not met, it was necessary to run the non-parametric Friedman test to determine whether there was a significant difference in the scores participants had received depending on the scoring guide employed. Wilcoxon signed-rank tests with a Bonferroni adjustment were run on the different combinations of the three groups of different scoring conditions to examine where the group differences actually occur. To answer RQ3, correlations on test scores per scoring condition were used to examine the relationships between the different pairs of scoring guides as strong correlations might indicate corresponding constructs.

10.5 Findings

This section presents the results of the statistical analyzes, which form the basis of the subsequent discussion. After presenting the descriptive statistics of the individual scoring conditions to facilitate the interpretation of the findings, results of the non-parametric analyzes, as well as the correlation coefficients, will be outlined.

10.5.1 Facility Values

Descriptive statistics provide a first impression of the distribution of scores obtained by the test takers. Table 10.2 illustrates that the facility values (mean scores) of the overall results of exact marking (scoring condition A), ($N = 142$) range from 0.7% (item q15) to 95.8% (item q12); for scoring condition B they range from 12%

Table 10.2 Item facility value per scoring condition

	A (%)	B (%)	C (%)
Task1_q1	59.15	62.68	62.68
Task1_q2	56.34	63.38	64.08
Task1_q3	30.99	42.25	45.77
Task1_q4	19.72	58.45	61.27
Task1_q5	53.52	62.68	62.68
Task1_q6	39.44	61.27	64.08
Task1_q7	18.31	30.28	57.04
Task1_q8	47.89	54.23	54.23
Task1_q9	4.93	11.97	52.82
Task1_q10	31.69	70.42	76.06
Task2_q11	48.59	64.79	68.31
Task2_q12	95.77	95.77	97.18
Task2_q13	59.15	66.90	72.54
Task2_q14	29.58	81.69	85.21
Task2_q15	0.70	30.99	66.90
Task2_q16	11.27	58.45	66.90
Task2_q17	16.90	40.14	42.25
Task2_q18	14.08	17.61	26.76
Task2_q19	22.54	68.31	70.42
Task3_q20	61.97	64.08	65.49
Task3_q21	37.32	66.20	68.31
Task3_q22	1.41	17.61	23.24
Task3_q23	2.82	35.92	35.92
Task3_q24	14.08	70.42	73.24
Task3_q25	38.73	46.48	46.48
Task3_q26	28.17	54.93	57.04
Task4_q27	47.18	62.68	62.68
Task4_q28	22.54	35.21	38.03
Task4_q29	61.97	75.35	75.35
Task4_q30	29.58	45.77	48.59
Task4_q31	6.34	23.24	23.24
Task4_q32	64.08	72.54	72.54

(item q9) to 95.8% (item q12), and for C from 23.2% (items q22 and q31) to 97.2% (item q12). All items have an equal or a higher facility value when marked with the post-live administration scoring guide than with the post-central correction scoring guide. Task 3 had the lowest overall mean facility value throughout all three scoring conditions; hence it was consistently the most difficult task out of the four for the sample population. The mean facility value for Task 4 changed the least from scoring condition A (38.6%) to scoring condition C (53.4%), and it only changed minimally from scoring condition B to C with a difference of 0.9%, whereas the mean facility value of Task 1 increased the most from scoring condition B (51.8%) to scoring condition C (60.1%).

Highlighted in Table 10.2 are potentially problematic items. Green (2013) argues that facility values between 20 and 80% can provide useful information about a test taker's proficiency provided the items discriminate and contribute to the test's internal consistency (p. 26). Applying these values, 11 items show facility values below 20% for scoring condition A and one item reaches a facility value of 95.8% (item q12). Only three items have facility values below 20% for scoring condition B (items q9, q18, q22), but two items reach facility values above 80% (items q12 and q14), and with scoring condition C no item has a facility value below the threshold level while the same two items (items q12 and q14) have facility values above.

The total mean score improved quite considerably from scoring condition A to C. When exact marking was applied the mean was 10.77 ($SD = 5.24$), for scoring condition B the mean score was 17.13 ($SD = 7.07$) and for scoring condition C the mean score was 18.87 ($SD = 7.36$) out of 32 items.

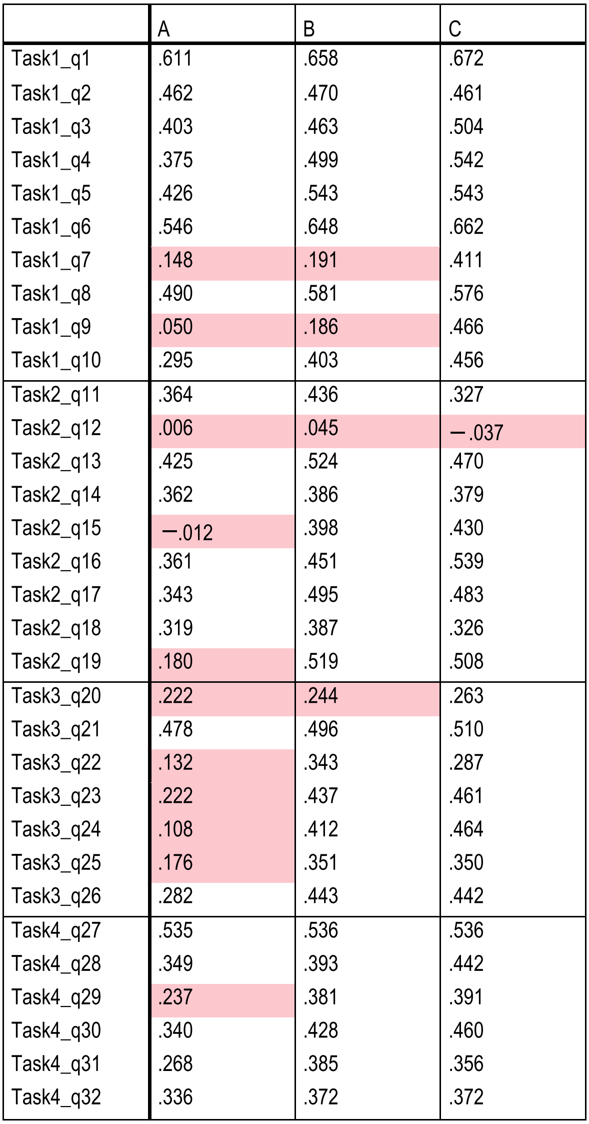
10.5.2 Discrimination and Reliability

Table 10.3 shows the impact of the three scoring conditions on the item discrimination, as indicated by the corrected item-total correlation (CITC). A CITC of .25 or above may be seen as suggesting acceptable discriminatory power (Henning 1987). All values below this level are highlighted in red. As can be gleaned from Table 10.3, 11 items fall below this limit for scoring guide A and four for scoring guide B. Only one item shows unsatisfactory discrimination for scoring guide C.

All three scoring conditions had satisfactory reliability values when analyzed with classical test theory. The Cronbach alpha for scoring condition A was .82, for scoring condition B .89, and for scoring condition B .90. When exact marking was applied, six items show problematic values for the Cronbach's Alpha If Item Deleted. For scoring condition B items q7 and q12 show slightly problematic reliability values, while for scoring condition C item q12 contributes negatively to the test's internal consistency.

Skewness and kurtosis ratios of the total scores per scoring condition indicated a departure from symmetry because not all of the resulting values were within ±2 (Green 2013); hence a normal distribution could not be assumed. Furthermore, the Shapiro–Wilk test of normality was significant at the .05 significance level for all

Table 10.3 Item discrimination per scoring condition (CITC)



	A	B	C
Task1_q1	.611	.658	.672
Task1_q2	.462	.470	.461
Task1_q3	.403	.463	.504
Task1_q4	.375	.499	.542
Task1_q5	.426	.543	.543
Task1_q6	.546	.648	.662
Task1_q7	.148	.191	.411
Task1_q8	.490	.581	.576
Task1_q9	.050	.186	.466
Task1_q10	.295	.403	.456
Task2_q11	.364	.436	.327
Task2_q12	.006	.045	−.037
Task2_q13	.425	.524	.470
Task2_q14	.362	.386	.379
Task2_q15	−.012	.398	.430
Task2_q16	.361	.451	.539
Task2_q17	.343	.495	.483
Task2_q18	.319	.387	.326
Task2_q19	.180	.519	.508
Task3_q20	.222	.244	.263
Task3_q21	.478	.496	.510
Task3_q22	.132	.343	.287
Task3_q23	.222	.437	.461
Task3_q24	.108	.412	.464
Task3_q25	.176	.351	.350
Task3_q26	.282	.443	.442
Task4_q27	.535	.536	.536
Task4_q28	.349	.393	.442
Task4_q29	.237	.381	.391
Task4_q30	.340	.428	.460
Task4_q31	.268	.385	.356
Task4_q32	.336	.372	.372

three scoring conditions, suggesting violation of the assumption of normality, which is why the Friedman test was employed to provide information on whether there were overall significant differences in the scores participants had received. The Friedman test indicated that there was a statistically significant difference in the results of test takers' scores depending on which scoring guide was used, $\chi^2(2) = 265.89, p = .000$. The median values showed an increase of test scores from scoring condition A ($Md =$

11) to scoring condition B (*Md* = 18), and a further increase with scoring condition C (*Md* = 20). Post hoc analyzes with Wilcoxon signed-rank tests were conducted with a Bonferroni adjustment applied, resulting in a significance level set at $p < .017$. There were statistically significant differences between scoring conditions A and B ($z = -10.32$, $p = .000$, $r = .612$), between scoring conditions A and C ($z = 10.35$, $p = .000$, $r = .614$), and also between scoring conditions B and C ($z = 9.17$, $p = .000$, $r = .544$), each condition having a large effect size.

Spearman's rho correlations were calculated for the three sets of scoring conditions to examine the strength of the relationship between the scoring guides. Results for 142 sets of scores showed strong, positive correlations for scoring condition A and scoring condition B ($r = .929$. $p < .001$. 2-tailed), scoring condition A and scoring condition C ($r = .923$. $p < .001$. 2-tailed), and scoring condition B and scoring condition C ($r = .979$. $p < .001$. 2-tailed). The strongest correlation can be found between scoring conditions B and C, with a shared variance of 95.8%.

10.6 Insights Gained

This section will attempt to answer the research questions about the impact of marking with different scoring guides.

RQ1: Which differences can be detected at the score and item level when four English listening constructed-response tasks at CEFR B2 level of the national standardized Austrian school-leaving examination are marked with (A) a rigid scoring guide from the pre-trial stage (B) an enhanced post-trial central correction scoring guide, and (C) a further extended post-live administration scoring guide?

Comparing the results of item statistics calculated for the three different scoring guides, one learns that the total mean score improves considerably from scoring condition A to scoring condition C. The mean score for exact marking (A) is 11. For the post-central correction scoring guide (B) it is 17, and the mean score increases further to 19 when candidates' responses are marked with the post-Matura scoring guide (C). In addition, the facility values per item improve for all items from scoring condition A to scoring condition C, with the biggest increase for item q15, which shows an improvement of 66.2%. The same holds true for CITC values. Discrimination values generally improve from scoring condition A to scoring condition C, except for two items (q2 and q11), which show slightly weaker values when the post-live administration scoring guide (C) was applied. In these two cases scoring condition C seems to have accepted answers from weaker candidates, which might not reflect their general performance on the test as a whole. Overall, the results on item and score level indicate that on average test takers receive higher scores and hence better grades on their listening performance the more comprehensive the scoring guides are.

A further finding is that for some items (q1, q5, q8, q23, q25, q27, q29, and q32) the facility values did not change from scoring condition B to scoring condition C. Thus, for a number of items the central correction scoring guide was sufficiently comprehensive to mark test takers' responses. This seems to confirm the significance of central correction meetings. The better and the more comprehensive the central correction scoring guide for a given task is, the easier and more reliable marking is for external markers (i.e., teachers), and the more reliable are the decisions taken during the helpdesk and hotline stage. Therefore, central correction is an essential stage in the test development cycle.

RQ2: Which differences can be detected in test reliability when the same four English constructed-response tasks are marked with (A) a rigid scoring guide, (B) an enhanced post-central correction scoring guide, and (C) a further extended post-live administration scoring guide?

Overall test reliability improves from a Cronbach alpha value of .82 for exact marking (A), over .89 for the central correction scoring condition (B) to .90 for the post-Matura scoring condition (C). Although a Cronbach alpha coefficient of .82 for exact marking is acceptable, in a high-stakes testing situation one aims for the highest possible internal reliability coefficient and therefore the improved value for the post-helpdesk and hotline scoring condition (C) confirms that the effort of maintaining and refining these support structures is well worthwhile.

RQ3: Do the three different scoring conditions measure the same construct?

The correlation analyzes show strong, positive relationships between the three sets of scoring guides, with the strongest correlation of .98 between scoring condition B and C, i.e., a shared variance of 95.8%. This seems to confirm that the marking decisions taken are consistent and in line with the construct that is being tested.

Correlation coefficients are above .9 for all three pairs of variables which seem to suggest that the construct being measured, namely, listening for main ideas and supporting details at CEFR B2 level, is basically the same for all three scoring conditions, and that the scoring guide used has little impact on the underlying construct being tested. However, the lowest correlation coefficient out of the three is achieved between scoring condition A and scoring condition C with a value of .92, which means that the internal reliability of the test can be improved further when comprehensive marking is carried out. In fact, test quality improves each time the scoring guide is expanded.

If all psychometric properties are taken into account, we can see that test takers obtain higher scores and thus better results when their responses are marked with the post-live administration scoring guide (C) and that this scoring condition also best measures the underlying construct of listening comprehension. Since the Austrian standardized school-leaving examination is a high-stakes test, the procedures of central correction and the support structures provided through helpdesk and hotline seem to be worth the time, money and effort spent in order to guarantee the best possible and fairest results for test takers. The findings of the present paper strongly support current practice and recommend improving and refining the implemented support structures further.

10.7 Conclusion: Implications for Test Users

This chapter has attempted to investigate the impact of three different scoring conditions on the test taker results on four B2 listening constructed-response tasks for English as foreign language. In particular, it has looked at the effects of exact marking in comparison with more flexible scoring guides that have been enhanced and improved during two stages of the test development cycle.

In doing so, this study has followed the development of scoring guides for constructed-response listening items from their inception at the pre-trial stage, to the extended version of the post-trial stage, when the scoring guides are improved through a standardization process at central correction, and finally to the post-live administration stage, when the scoring guides are further enhanced while offering marking support through an online helpdesk system and a telephone hotline service for teachers. The initial scoring guide served as the exact scoring guide baseline.

The scores obtained from 142 Austrian test takers at a presumed B2 level of the CEFR were statistically analyzed and the results showed considerable improvement in terms of item facility, discrimination and reliability values, from the exact scoring condition (A) to the considerably extended post-live administration scoring guide (C). These findings seem to support the results of an earlier study by Hackett et al. (2006), although their study only showed minor increases in item reliability when using a comprehensive scoring guide compared to a stringent scoring guide that only allowed a restricted range of misspellings. The present study confirms that the results from marking with comprehensive scoring guides noticeably improve not only test reliability, but also test takers' scores, while still equally or better reflecting the underlying construct. Hence, the measures taken to further develop the scoring guides and to support the actual marking process after the Matura exam have a positive impact on the quality of the exam.

However, there is still a need for more in-depth research on item types and the variability between different items depending on the scoring conditions, and the reasons for this. In doing so, clearer guidelines for item writing could be developed to further increase the quality of note-form tasks that target the construct of listening for main ideas and supporting details and to decrease the risk of construct-irrelevant variance (Khalifa and Weir 2009) in such tasks. Another area of research could be the development of a database in which all decisions regarding an answer are recorded, which could then serve as the basis for future decisions if the same or a similar response is targeted again in a different task.

Taylor (2013) states that "scoring validity accounts for the extent to which […] scores on constructed-response tasks are arrived at through the application of appropriate criteria, exhibit agreement, are as free as possible from measurement error, stable over time, appropriate in terms of their content sampling and engender confidence as reliable decision-making indicators" (p. 30). The results of the present chapter have, hopefully, contributed not only to the limited body of research on the impact of different scoring procedures on marking open-ended listening items, but also delineate potential ways of improving the reliability of marking in a high-stakes testing situation.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

BMUKK [Bundesministerium für Unterricht, Kunst und Kultur]. (2004). *Oberstufenlehrplan für die Erste und Zweite Lebende Fremdsprache für Allgemein Bildende Höhere Schulen.* http://www.bmukk.gv.at/medienpool/11854/lebendefremdsprache_ost_neu0.pdf. Accessed 13 October 2013.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics, 18,* 171–191.

Brown, G., & Yule, G. (1983). *Teaching the spoken language.* Cambridge: Cambridge University Press.

Buck, G. (2001). *Assessing listening.* Cambridge, UK: Cambridge University Press.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf. Accessed 2 November 2013.

Eberharter, K., & Frötscher, D. (2013). Quality control in marking open-ended listening and reading test items. In D. Tsagari, S. Papadima-Sophocleous, & S. Ioannou-Georgiou (Eds.), *International experiences in language testing and assessment: Selected papers in memory of Pavlos Pavlou* (pp. 229–242). Frankfurt: Peter Lang.

Elliott, W., & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. B. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 152–241). Cambridge, UK: Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. B. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge, UK: Cambridge University Press.

Green, R. (2013). *Statistical analyses for language testing*. Basingstoke: Palgrave Macmillan.

Hackett, E., Geranpayeh, A., & Somers, A. (2006). *Listening skills group spelling project: Investigating the impact of the revision of an FCE 4 productive task mark scheme based on the recommendations of four external consultants* (Cambridge ESOL Internal Report).

Harding, L., Pill, J., & Ryan, K. (2011). Assessor decision making while marking a note-taking listening test: The case of the OET. *Language Assessment Quarterly, 8*(2), 108–126.

Harding, L., & Ryan, K. (2009). Decision making in marking open-ended listening test items: The case of the OET. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7,* 99–114.

Henning, G. (1987). *A guide to language testing: Development, evaluation and research.* Cambridge, MA: Newbury House.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge, UK: Cambridge University Press.

Lynch, T. (2009). *Teaching second language listening*. Oxford, UK: Oxford University Press.

Spöttl, C., Eberharter, K., Holzknecht, F., Kremmel, B., & Zehentner, M. (2018). Delivering reform in a high stakes context: From content-based assessment to communicative and competence-based assessment. In G. Sigott (Ed.), *Language testing in Austria: Taking stock* (pp. 219–240). Berlin: Peter Lang.

Spöttl, C., Kremmel, B., Holzknecht, F., & Alderson, J. C. (2016). Evaluating the achievements and challenges in reforming a national language exam: The reform team's perspective. *Papers in Language Testing and Assessment, 5*(1), 1–22.

Taylor, L. (2013). Introduction. In A. Geranpayeh & L. B. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 1–35). Cambridge, UK: Cambridge University Press.

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes, 10*(2), 89–101.

Weiler, T., & Frötscher, D. (2018). Ensuring sustainability and managing quality in producing the standardized matriculation examination in the foreign languages. In G. Sigott (Ed.), *Language testing in Austria: Taking stock* (pp. 241–260). Berlin: Peter Lang.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Basingstoke: Palgrave Macmillan.

Chapter 11
Score Changes with Repetition of Paper Version(s) of the TOEFL in an Arab Gulf State: A Natural Experiment

Betty Lanteigne and Hana Sulieman

Abstract Although replaced by the internet-based TOEFL (iBT), the paper-based TOEFL (PBT) was still used in some countries until July 2017. This natural experiment investigates a case where, November 2006–May 2009, a testing center in the Middle East did not follow TOEFL time interval protocol at that time, with 1,575 test takers allowed to attempt paper TOEFL versions as frequently as desired. This study analyzes score increase/decrease with frequency of/time intervals between attempts at taking paper TOEFL version(s). These test takers attempted the TOEFL multiple times, with 48% taking it at least twice, some up to 22 times. Results indicate a negative relationship between number of attempts and score increases/decreases between consecutive tests greater than Standard Error of Measurement (>*SEM*), and a positive relationship between score change and average time interval. Score decreases showed greater variability, with smaller average time interval than increases. Percentages of individuals' score decrease averaged in repeater group positively correlated with number of attempts. Consecutive test scores decreased (>*SEM*) at least once for 900 test takers. Score decreases showed greater variability and significantly smaller average time intervals than increases. There was very strong positive correlation between increase in number of attempts and consecutive score decreases. Repetitions within one week showed greater score variations. These findings raise questions about construct-irrelevant effects of test repetition and time interval on score change. One explanation is that score increases >*SEM* occurred gradually over more time, in smaller amounts, in keeping with increases in ability through learning, but decreases >*SEM* occurred in shorter time intervals and larger amounts, possibly reflecting more immediate negative affect-related factors.

B. Lanteigne (✉)
LCC International University, Klaipeda, Lithuania
e-mail: blanteigne@lcc.lt

H. Sulieman
American University of Sharjah, Sharjah, UAE
e-mail: hsulieman@aus.edu

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_11

11.1 Introduction: Purpose and Testing Context

The TOEFL (Test of English as a Foreign Language) is frequently used as a high-stakes test required for admission of non-native-English speakers to universities in America and some English-medium universities in other countries. For many students, achieving the required TOEFL score means that, instead of having to study in an intensive English program (IEP) to improve their English skills, they can enroll in university courses and begin their university studies.

Currently, the Educational Testing Service (ETS) has two forms of the TOEFL: internet-based (iBT) and paper-delivered where internet-based testing is not possible. Formerly, the paper-based TOEFL (PBT) was used, but it began to be phased out in 2012 in most areas of the world (Educational Testing Service 2012), although it continued to be offered in regions where infrastructure did not support consistent computer/internet access, a function replaced by the paper-delivered TOEFL in July, 2017 (Educational Testing Service 2019a, b, d). However, Institutional Testing Program (ITP) TOEFL paper versions are used by universities, language institutes, and other organizations for institutional, program, and classroom purposes in locally administered institutional applications. As such, these paper versions of the TOEFL (the former PBT, ITP, and Paper-delivered) required and still require standardized administration for reliable and valid results.

This study analyzes a phenomenon in which the ETS-established time interval policy between test administrations at that time (once a calendar month) was not followed at a testing center in an Arab Gulf State, November 2006–May 2009. During that time period, paper versions of the TOEFL were being used in determining admission to the university associated with that particular testing center.

11.2 Testing Problem Encountered

The focus in this research is on patterns of score change greater than the Standard Error of Measurement (>*SEM*) related to repetition of the TOEFL in *naturally occurring* data. In this context, naturally occurring indicates that test takers themselves chose to take the TOEFL for their (unknown) personal reasons, as opposed to a controlled experiment eliciting specified numbers of attempts at prescribed time intervals. In the fall of 2009, English teachers in the IEP associated with the testing center reported to friends who were students of the first author, that many of their students said they took the TOEFL on multiple occasions close together (as often as three times a week) yet their TOEFL scores decreased drastically with repeated attempts. Note: Only after this protocol-violating practice of test repetition was stopped by the testing center did the researchers begin to investigate this practice.

The testing center which administered these tests has acknowledged that during this time period it did not consistently adhere to ETS policy in effect at that time. In May 2011, the first author contacted the testing center in question and found there was a new director who could not comment on prior practices, but who stated that ETS time interval guidelines were being followed under her jurisdiction at that time. The test offered was called the PBT, although the scores were reported to test takers on the testing center website (instead of reports from ETS), which would seem to indicate that the paper versions of the TOEFL offered were ITP.

Also, in this case of non-standard paper TOEFL version administration, the scores were publicly available (posted for test takers by unique identification numbers on the testing center website). Since the scores of individual test takers could thus be tracked, this gave opportunity for analysis of effects of repetition of the TOEFL in shorter time intervals than ETS policy permitted at that time, making possible a natural experiment in the sense that the researchers' only involvement was analysis of the data after the fact.

The descriptive component of this analysis was similar to Wilson's (1987), looking at how often the students were taking the TOEFL at this center. Thus the first research question addressed number of attempts:

Research question 1: What is the incidence of 1-time, 2-time, …, *n*-time test taking?

The second issue was whether or not it was true that some students were experiencing large score decreases. Thus the second research question addressed score change:

Research question 2: What are the patterns of change in TOEFL performance for repeating test takers by number of times tested?

The third issue was whether or not there was a connection between score change greater than the *SEM* (especially score decrease) and number of attempts and/or time interval. *SEM* is a statistic calculating how much variation in scores is due to factors other than what is intended to be measured, and thus looking at score change greater than *SEM* is more likely to produce relevant and meaningful insight. Hence consecutive score change >*SEM* was addressed in the third and fourth research questions:

Research question 3: What is the relationship between consecutive score changes greater than the *SEM*, and number of attempts?

Research question 4: What is the relationship between consecutive score changes greater than the *SEM*, and time interval between consecutive attempts?

11.3 Review of Literature

This testing center's lack of compliance with ETS time interval protocol may have compromised the accuracy of information obtained about the test takers' English ability. Issues pertinent to this situation include construct-irrelevant variance, responsibility for test administration reliability, and test repetition.

A potential problem with this practice of allowing students to take the TOEFL on multiple occasions close together is possible introduction of construct-irrelevant variance (Messick 1989) distorting accurate measurement of test taker English proficiency through introduction of elements having nothing to do with the construct in question, which for the PBT included English reading, writing, listening, and structure.

Factors which could result in inaccurately low test performance include “inattention, anxiety, low motivation, fatigue, adverse testing conditions, and insufficient test-wiseness” (Messick 1984, p. 227). Test anxiety and fatigue could cause test scores to be lower than what true scores would warrant (Elliot and McGregor 1999; MacIntyre 1995; Zeidner 1998). Specifically concerning language testing, Henning (1987) points to fatigue, sickness, and emotional disturbance as possible elements negatively affecting scores.

While test anxiety and fatigue may result in inaccurately low scores, factors such as practice effect, item disclosure, and test-wiseness may inaccurately increase test scores. In pre-test/post-test evaluation, too short a time interval may introduce a practice effect from test takers repeating a test in a short period of time. Steinborn et al. (2009) investigated the score-increasing effect of practice, and Cohen and Wollack (2006) found that item overexposure can occur when the same test is given within a short time interval, allowing test takers to remember specific items. On the other hand, Barkaoui (2018) found that score changes for test takers repeating the Pearson Test of English Academic—Writing were not significantly affected by number of attempts or time interval between test attempts. Regarding the PBT, Hale et al. (1980) found that item disclosure increased PBT scores. Henning (1987) mentions the practice effect as a possible factor in language testing, and Wilson’s (1987) score change analysis suggests that, apart from language learning, test-wiseness may result in increased PBT scores.

Consistency in administrative conditions must be achieved so test results can be compared between individuals or institutions (Cohen and Wollack 2006; Shepard 2003). Linn (1998) advocates that responsibility for consequences of test score use be shared among test sponsors, developers, users, and reviewers. This view of language testing ethics indicates test users such as testing center administrators must comply with test administration protocol established by test developers.

One aspect of standardized administration is time interval between test repetitions. While numerous studies have involved test-retest validation research (validation using comparison of test taker scores on the same test given twice), some test repetition studies have investigated factors associated with test takers repeating a test. (See Alderman 1981; Barkaoui 2018; Cliffordson 2004; Kingston and Turner 1984; and Swinton et al. 1983.) Empirical research into negative affective effects of students attempting the same test on multiple occasions is limited in language testing contexts, but Meyer (2010) cautions that time intervals between test administrations should be spaced so as to avoid fatigue from test repetition. Also, Koretz (2008) mentions the possibility that test takers repeating the same test on multiple occasions may be fatigued or “simply fed up with testing” (p. 149).

Concerning time between test-retest administrations in determining reliability, Haertel (2006, p. 70) says the "interval chosen should be long enough to allow for the influence of transient fluctuations in test performance, but not so long as to allow for significant influence from learning or maturation." Repetition of the TOEFL, with different time intervals, is investigated in four studies. Henning (1987) looked at reliability estimates of the TOEFL, using test-retest coefficients with an eight-day interval of time, assuming that for the students in his study no significant learning took place in that time period. Zhang (2008) analyzed test-retest reliability of the iBT, using scores of repeaters taking the iBT in consecutive calendar months, assuming that "learning would be unlikely to occur within a month in the usual course of language acquisition" (p. 1). Powers and Lall (2013) report a study by Lall looking at TOEFL score change with time intervals ranging from less than 30 days to more than 120 days, with some test takers showing decreases in test scores and particularly large score decreases with time intervals of more than 120 days. They conclude that it is possible some large score decreases may be due to more than random fluctuations. Powers and Lall also indicate that TOEFL repeaters are self-selected, which raises the question of regression to the mean (Alderman 1981), a general tendency for extreme characteristics to become more moderate over time, thus closer to the average. Wilson (1987) investigated patterns of PBT score change and test taker characteristics, including analysis of score changes with respect to time intervals. He observed that a longer time interval between attempts may provide opportunity for language learning and the "number of times tested may be an indirect measure of effort, motivation, financial resources, practice effects, and so on" (p. 38).

Although investigating a different test taker population than did Wilson, this study also focuses on repetition of paper version(s) of the TOEFL, as well as time interval between attempts.

11.4 Methodology

Part of a larger project including quantitative and qualitative analysis, this study analyzes publicly available TOEFL scores of test takers taking paper versions of the TOEFL at a testing center from November 2006, through May 2009.

11.4.1 Data Collection

To investigate whether or not there was a decrease in scores with repeated attempts and if there were significant correlations between size of decrease and time interval or between size of decrease and number of repetitions, TOEFL scores (November 2006–May 2009) were downloaded in the fall of 2009 from the testing center's website where they were posted by test identification numbers. After the scores were verified with the testing center, scores of 3,328 test takers were subsequently analyzed in a time series, identifying 1,575 who had taken the TOEFL on more than one occasion.

Note: It is recognized that some test takers may have taken the TOEFL elsewhere or outside of the November 2006–May 2009 time period. Also, it is unknown which version(s) of the TOEFL were taken by which test takers. The tests, as mentioned above, were very likely ITP instead of PBT, but when contacted, the testing center did not say (and may not have known) which versions were used for which students.

11.4.2 The Test Takers

No demographic information was available about the test takers. However, based on the demographics of the student population of this university and the associated IEP in 2006–2009, it is likely that most of these self-selected test takers were male and female non-native-English speakers from Africa, Asia, the Middle East, Europe, and/or South America, with some native English speakers also included. These test takers were self-selected test repeaters who could have been students just graduating from secondary school, IEP students, or working adults applying for undergraduate or graduate study. It is not possible that all of the test takers were IEP students, though, since the number of test takers was greater than the IEP enrollment at that time. Considering the university's population of undergraduate and graduate students, their ages could have ranged from 17 to 45 or older. Registration costs for the TOEFL were paid by the individual test takers. Students applying for undergraduate studies at that time had to achieve a PBT score of 530 or greater, and those applying for graduate studies had to achieve a score of 550 or greater.

11.4.3 Data Analysis

Score change was defined by the absolute difference (positive or negative change) between two consecutive test scores, thus including score increase and score decrease. For research questions three and four, only score changes greater than the Level 1 PBT *SEM* of ±13 points (Educational Testing Service 2018) were analyzed, to exclude score fluctuations within the range of measurement error. [Because the exact paper TOEFL version(s) used by this testing center are unknown, the Level 1 PBT *SEM* of ±13 points is used in analysis of this data, meaning that only score changes greater than 13 points were analyzed.] For each test taker three averages of score change were calculated: all recorded score changes, only positive score changes (score increase), and only negative score changes (score decrease). The corresponding averages for time interval (measured by number of days between two consecutive tests) were also calculated. In addition, percentages of score decreases and score increases were computed.

To examine the relationships between mean score change and each of the two independent variables (time interval and number of test attempts), scatterplot charts and Pearson's correlation coefficient were used, with scatterplots displaying score changes for each time interval and each test attempt group. Pearson's indicates the strength and direction of relationship between these variables. While interpretation of strength of correlation varies according to the purpose of analysis (Hopkins 1998), in this research, following Dancey and Reidy (2004, p. 171), Pearson's correlation coefficient was deemed to be strong (±0.70 – ±1.0), moderate (±0.40 – ±0.69), or weak (±0.1 – ±0.39). Direction of relationship could be positive or negative. For example, a positive relationship between number of attempts and score increase would indicate that as test takers took the test more times, their score increases would be greater, while a negative relationship would indicate greater number of attempts coinciding with smaller score increases.

Difference in time interval/score change means (averages) for decreases and increases was examined using confidence interval, a statistic giving a range of scores likely to include the true average of a set of scores.

Regression involves statistical analysis of the effect of independent variable(s) on a dependent variable. One type of regression, hierarchical regression, allows for exclusion of confounding effect by an independent variable, and in this study hierarchical regression was utilized to examine the effect of number of attempts (independent variable) on score change (dependent variable) when the effect of time interval (independent variable) was controlled for. The nominal 5% significance level was used to declare significance, indicating that there is only a 5% chance of mistakenly claiming that the findings are statistically significant. [For more information about these statistics, see Bachman (2004), Bailey (1998), and Dancey and Reidy (2004).]

11.5 Findings

The results of the statistical analyses and what these findings indicate are presented here in order of the research questions.

11.5.1 Research Question 1. What Is the Incidence of 1-Time, 2-Time, …, n-*Time Test Taking?*

Of the 3,328 students taking the TOEFL, 1,575 (47.33%) took it on multiple occasions, up to 22 attempts. In Table 11.1 summary statistics of the 3,328 sets of TOEFL scores [mean, standard deviation (*SD*), coefficient of variation (*CV*), median, minimum, and maximum] are reported for each number-of-attempts group. The *SD* indicates how spread out the scores are from the mean (average), and the *CV* is a statistic showing how different data sets compare in terms of their variation, even

Table 11.1 Summary statistics of TOEFL scores by number of attempts

# Attempts	# Students	% Students	Mean	*SD*	*CV* (%)	Median	Min, Max
1	1,753	52.64	558.63	44.06	7.9	553	483, 677
2	445	13.36	492.71	47.50	9.6	503	310, 590
3	305	9.16	490.4	40.17	8.2	500	353, 577
4	204	6.10	482.28	41.71	8.6	490	347, 567
5	154	4.62	484.05	38.86	8.0	490	313, 570
6	120	3.60	485.07	36.20	7.5	490	367, 567
7	83	2.50	485.01	34.62	7.1	490	373, 570
8	83	2.50	485.81	34.77	7.1	490	323, 573
9	47	1.41	480.93	36.14	7.5	487	343, 570
10	34	1.02	482.23	32.86	6.8	487	357, 557
11	25	0.75	487.58	30.36	6.2	490	397, 560
12	18	0.54	478.13	34.40	7.2	483	353, 547
13	15	0.45	484.68	34.21	7.0	490	350, 557
14	11	0.33	483.69	30.47	6.3	487	380, 553
15	9	0.27	478.08	32.30	6.7	483	380, 540
16	7	0.21	481.84	32.43	6.7	483	373, 540
17	1	0.03	482.06	29.24	5.9	490	420, 517
18	4	0.12	492.85	26.60	5.4	498	423, 533
19	3	0.09	482.54	34.04	7.1	487	393, 542
20	3	0.09	452.34	38.0	8.4	450	357, 533
21	1	0.03	453.46	24.15	5.4	457	403, 495
22	3	0.09	469.06	30.54	6.5	471	393, 517
Total	3,328	100					

when they have different means. The median is the score that is in the middle of a range of scores.

Overall, mean TOEFL scores and maximum scores are lower for test takers with larger numbers of test attempts than for test takers with smaller numbers of attempts. The *CV* (measuring how dispersed the scores are relative to the mean) also generally lessens with greater numbers of attempts. The one-attempt group had the highest mean, median, and maximum of all of the groups. It is also noted that the repeater groups varied in size, with 445 test takers in the two-attempts group, and only one test taker in each of the 17- and 21-attempts groups.

11.5.2 Research Question 2. What Are the Patterns of Change in TOEFL Performance for Repeating Examinees by Number of Times Tested?

Table 11.2 shows summary statistics of score change calculated by number of test attempts. Score change is the absolute value (positive or negative change) of the difference between a current score and its preceding score. The table exhibits an overall pattern in which mean score change and maximum score change decrease as the number of attempts increases. Minimum score differences of 0 were not included in mean SC calculation, although minimum score differences of 0 are included in the table.

Table 11.2 Patterns of TOEFL score change (SC[a]) by number of test attempts

# Attempts	# Students	SC	Mean SC	*SD* SC	*CV* SC (%)	Median SC	Min, Max SC
2	445	436	30.03	21.72	72	26.00	0, 153
3	305	595	22.20	16.67	75	20.00	0, 113
4	204	596	21.04	16.95	81	17.00	0, 83
5	154	601	20.45	17.72	87	17.00	0, 130
6	120	590	20.54	16.88	82	17.00	0, 100
7	83	489	19.34	15.76	81	16.00	0, 93
8	83	576	20.40	15.83	78	17.00	0, 117
9	47	373	22.21	18.76	84	20.00	0, 127
10	34	303	19.59	15.02	77	17.00	0, 97
11	25	246	17.34	13.95	81	14.00	0, 70
12	18	198	20.86	14.36	69	18.00	0, 73
13	15	178	19.29	17.25	89	14.00	0, 93
14	11	141	17.50	14.49	83	13.00	0, 57
15	9	123	18.61	16.24	87	14.00	0, 93
16	7	103	17.24	14.00	82	14.00	0, 80
17	1	16	17.69	10.85	61	17.00	0, 43
18	4	68	16.79	11.61	69	17.00	0, 50
19	3	54	19.83	16.34	82	14.00	0, 60
20	3	55	21.71	14.61	67	20.00	0, 60
21	1	20	19.60	10.18	52	20.00	4, 40
22	3	63	14.29	13.35	94	13.00	0, 70

[a] SC: Score Change = |current score-previous score| (Score differences of 0 not included in SC)

11.5.3 *Research Question 3. What Is the Relationship Between Consecutive Score Changes Greater Than the* SEM*, and Number of Attempts?*

Concerning the relationship of score change with number of attempts, Fig. 11.1a shows individual values of score changes greater than *SEM* vs number of test attempts. There were larger score changes for test takers with smaller numbers of attempts compared to those with larger numbers of attempts. To further reveal the relationship between score change greater than *SEM* and number of attempts, mean score change greater than *SEM* per test taker is calculated and plotted against number of attempts in Fig. 11.1b. It is clearly seen that mean score change lessens as number of attempts increases, in a curvilinear pattern. Figures are placed side by side to show contrasts.

Figures 11.2 and 11.3 depict the relationship of mean score changes with number of test attempts for mean score decreases and mean score increases greater than

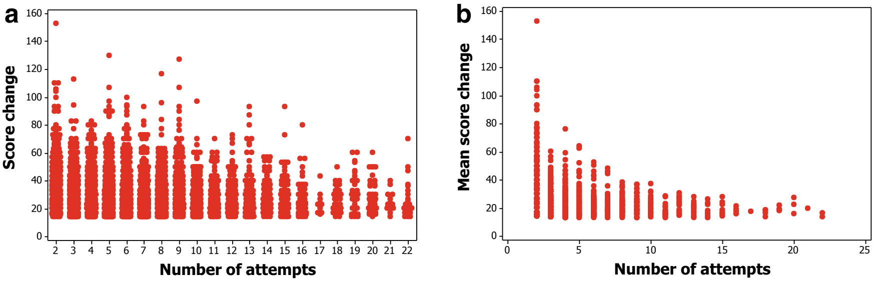


Fig. 11.1 **a** Individual score change greater than *SEM* vs number of attempts. **b** Mean score change greater than *SEM* vs number of attempts

Fig. 11.2 Mean score decreases greater than *SEM* vs number of attempts

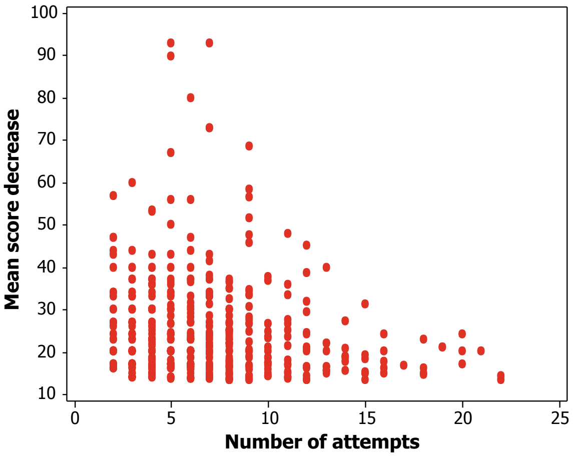
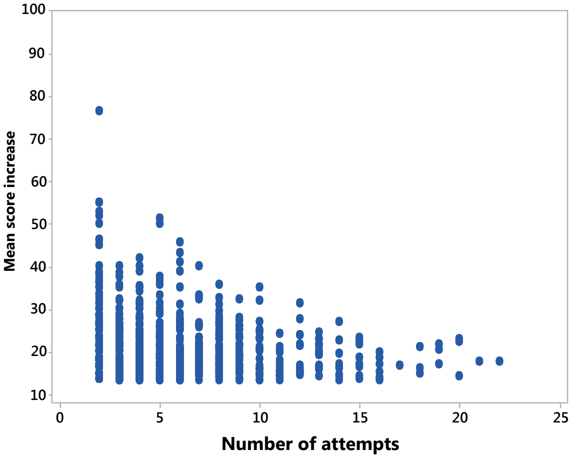


Fig. 11.3 Mean score increases greater than *SEM* vs number of attempts

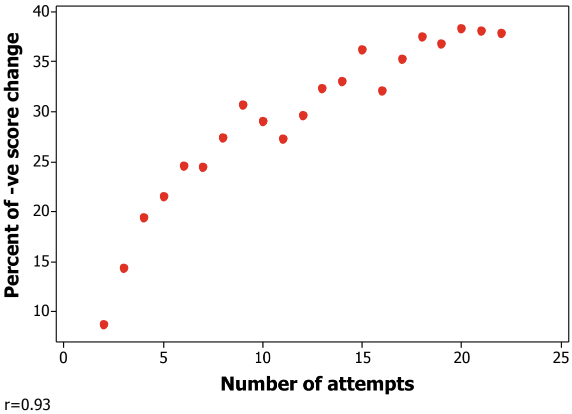


SEM, respectively. The overall behavior of the scatter plots resembles the curvilinear pattern seen in Fig. 11.1b. However, the mean decreases (Fig. 11.2) show greater fluctuation in a larger range of values and relatively weaker rate of decrease as number of attempts increases, than the corresponding mean score increases (Fig. 11.3). In particular, there are large decreases (between 60 and 100 points) between five and ten attempts.

Of the 1,575 test repeaters, 900 (57%) had at least one score decrease between consecutive attempts, and 504 of the 900 test takers with decreases (56%) had mean score change greater than *SEM*. It should be noted that the few large score changes (greater than 80 points) were decreases by test takers who had fewer than ten test attempts.

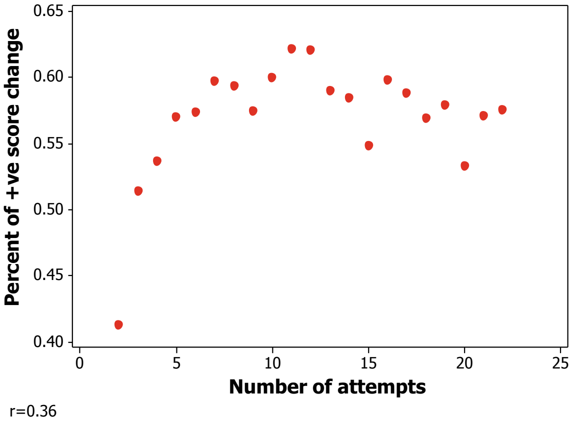
Figure 11.4 (following) shows percentage of score decreases per test taker averaged by number of test attempts, plotted against number of attempts. The increasing

Fig. 11.4 Percentage of decreases per test taker averaged by repeater group vs number of attempts



r=0.93

Fig. 11.5 Percentage of increases/test taker averaged by repeater group vs number of attempts



linear pattern suggests that as number of attempts increases, percentage of decreases (negative score change) increases. The strong Pearson's correlation coefficient (*r* = 0.93) is significant.

Figure 11.5 displays the pattern of the percentage of positive score change (score increases) versus number of test attempts. It is evident that the score increases relationship is not positively linear, in contrast to that of the score decreases. In fact, as number of test attempts goes beyond 12 attempts, percentage of score increases declines. The range for percentages of increases/positive score change (42–63%) is greater than that of decreases/negative score change (5–38%).

11.5.4 Research Question 4. What Is the Relationship Between Consecutive Score Changes Greater Than the SEM, and Time Interval Between Consecutive Attempts?

The relationship of all individual score changes greater than *SEM* with corresponding time interval (in days) between consecutive tests shows an overall trend that is positive and weakly linear with a correlation coefficient of 0.22. This finding indicates that as time interval between consecutive tests increases, there is a tendency for score change to increase. Looking at decreased vs increased score changes greater than *SEM* in relationship to time, the overall positive relationship of score change is still evident, but with a much weaker correlation coefficient for the decreased score changes (*r* = 0.10) than the increased score changes (*r* = 0.23). (See Table 11.3.)

Also, time interval range is greater in the increases, with more test takers attempting the test in intervals greater than 200 days.

Further summary statistics for score decreases and increases greater than *SEM* are reported in Table 11.4. The average time interval for mean score decreases is less than

Table 11.3 Time interval (days) vs consecutive individual SC[a]

All individual score changes vs time interval	$r = 0.22$
Individual score decreases vs time interval	$r = 0.10$
Individual score increases vs time interval	$r = 0.23$

[a] Score differences of 0 not included in SC

Table 11.4 Time interval for increases vs decreases greater than *SEM*

	Increases	Decreases
Time interval (days)	49.91	40.60
Average	47.70	44.25
SD	96%	109%
CV	(48.12, 51.70)	(37.76, 43.44)
95% *CI*		
Score change	21.05	24.52
Average	6.90	11.63
SD	32%	47%
CV	(20.61, 21.50)	(23.49, 25.56)
95% *CI*		

that for increases, with a larger variation ($CV = 109\%$) relative to the mean. Since the two *CI*s do not overlap and the interval for increases is larger than the interval for decreases, it is evident that the score decreases have significantly smaller average time intervals than do the increases. Also, in terms of score change, the decreases show a higher level of variability and larger mean value than do the increases. *CI*s for the mean value of decreases and increases do not overlap, indicating that the mean score decreases are significantly larger than the mean score increases.

In closer examination of the effect of time interval between consecutive tests on score changes, mean score change was categorized into three categories: within one week (1–7 days), within one month (1–30 days), and over a month (31+ days). (Within one week is a subset of within one month.) Table 11.5 presents the three groups of time interval for mean score change greater than *SEM*. The *CV* values indicate that the within-one-week data have the largest amount of variation relative to the mean, while the within-a-month score changes show the least amount of variation.

Table 11.5 Within a week, within a month and over a month mean SC *greater than SEM*

Time interval	No. of students	Amount of SC	Mean	*SD*	*CV* (%)	95% confidence interval
1–7 days	29	31[a]	32.25	19.73	62	(25.30, 39.20)
1–30 days	982	1787	28.67	13.89	48	(28.03, 29.32)
31+ days	1014	1680	32.84	16.79	51	(32.04, 33.65)

SC = |current score-previous score|

[a] Two students attempted the test three times in one week

Table 11.6 Effect of number of attempts on score change, controlled for time

Variable	*R-Sq* change	*F*-change	Coefficient	St. error	*t*-value	*p*-value
Step 1	0.43	156.07				0.00
Time interval			0.21	0.0168	12.49	0.00
Step 2	0.09	23.50				0.00
Time interval Number of attempts			0.19 −0.81	0.0163 0.1670	11.64 −4.85	0.00 0.00

To examine the effect of number of attempts on score change (positive and negative) when time interval was controlled for, hierarchical regression analysis was carried out with time interval entered first. (See Table 11.6.)

When the number of attempts was added to the model in step 2, the explained variability of score change measured by R-squared increased by 9%. The effect of the independent variable (number of attempts) on score change (dependent variable) is characterized by a negative slope (−0.81) and is significant (*p*-value = 0), indicating a strongly negative effect of number of attempts on score change. The regression coefficient value of time interval (0.19), indicates a weakly positive effect on score change. These results are also in accordance with the correlation patterns seen in Figs. 11.1a and 11.b, and Table 11.4: Mean score change greater than *SEM* (increases and decreases) lessened as number of attempts increased. Concerning decreased vs increased score changes greater than *SEM* in relationship to time interval, the overall positive relationship of score change was still evident, but much weaker for the decreased score changes than the increased score changes.

11.6 Insights Gained

The findings of this study indicate that score decreases did indeed occur with frequent repetition of the TOEFL (most notably between five and ten attempts). Test takers attempted the TOEFL up to 22 times with approximately 48% taking it at least twice. Scores did decrease (greater than *SEM*) at least once between consecutive tests for 900 (56%) of the test takers. These findings (seen in Table 11.1) overall initially seem to be generally in keeping with regression to the mean in that the repeater groups with larger number of attempts tend to have lower means, medians, and maximum scores. Alternatively, it is plausible that those attempting the TOEFL fewer times had higher levels of English language proficiency to begin with and thus did not need to repeat the test as often in order to achieve their desired score.

The disparity in repeater group size (see Table 11.1) indicates that for the repeater groups with greater numbers of test takers, the variation is primarily between test takers, while for the groups with smaller numbers of test takers the variation is primarily within test takers. For example, the 444 score changes for the 445 test takers attempting the test twice would reflect differences between them, whereas the

20 score changes for the one individual attempting the test 21 times would reflect variation within that one test taker's attempts.

Several patterns emerged in analysis of the test score data. First, the most striking results with these test takers concerned decreases. Score decreases showed a greater level of variability and had significantly smaller average time intervals than the increases. Also, as number of attempts increased, percentage of score decreases between consecutive tests increased, in a very strong positive correlation, indicating that the test takers repeating the TOEFL more often were more likely to experience score decreases.

Second, score change between consecutive tests negatively correlated with number of test attempts, indicating that as number of attempts increased, the extent of score change lessened. This relationship held true for both score increases and decreases, indicating that test takers repeating the TOEFL more frequently were likely to see smaller changes in consecutive tests than those taking it less frequently. At first glance these findings raise the question of regression to the mean, with means, medians, and maximum scores lower, and also score changes lesser in absolute value, with greater numbers of attempts. However, these lessening score changes are averages for repeater groups, not for individual test takers. Also, the volatility of the *CV* values of decreases, particularly between five and ten attempts, indicates variation from the mean instead of regression to the mean, as does the variability of score change with attempts within one week.

Third, score change between consecutive tests positively correlated with average time interval; i.e., as time interval between tests increased, score change increased. However, the relationship holds true more strongly for increases than for decreases. Longer time intervals may allow for greater learning to take place (likely associated with score increase) or for test takers to forget what they learned if they were not studying, e.g., over a summer vacation (with likely score decrease). Similarly, Wilson (1987) observed this pattern of greater score change with score increases in longer intervals, while Powers and Lall (2013) found particularly large score decreases with time intervals of more than 120 days. Possible reasons for greater score decreases with greater time intervals include gaps in learning with associated decrease in English proficiency, fatigue from test repetition, and factors such as "inattention, anxiety, low motivation, fatigue, adverse testing conditions, and insufficient test-wiseness" (Messick 1984, p. 227), sickness, and emotional disturbance (Henning 1987). Also, one possible explanation is that score increases greater than *SEM* occurred more gradually over long periods of time and in smaller amounts, which is in keeping with increases in language ability through learning, but score decreases greater than *SEM* occurred in shorter time intervals and in larger amounts, which could possibly reflect a more immediate effect of negative affect-related construct-irrelevant factors.

Fourth, frequent repetition in short time intervals (within one week) was associated with greater variation in TOEFL scores (seen in *CV* values). Both Henning (1993) and Zhang (2008) indicate that a one-week time interval is appropriate for evaluation of test-retest of TOEFL reliability. However, these findings indicate that these within-one-week TOEFL repeaters demonstrated score fluctuation in a time frame frequently used in determining test-retest reliability for the specific purpose of avoiding score change due to possible language learning. This difference can be at least partially accounted for by the uniqueness of these self-selected test repeaters, compared to the worldwide population of TOEFL takers.

11.7 Conclusion: Implications for Test Users

These findings emphasize that standardized administration of the TOEFL in terms of time interval is crucial to maintain reliability and validity. While it is true that test familiarization can decrease test anxiety (Messick 1996; Cizek and Burg 2006) and it may be helpful for test takers to utilize practice tests, it is seen in the findings of this research that extensive test repetition had negative effects on some of these test takers' performance as reflected in test scores.

Limitations of this research center around the uniqueness of this instance of a testing center allowing test takers to repeat the TOEFL as often as they liked (a practice which should not have occurred) and the anonymity of the test score data making test taker characteristics unknown. Also, in naturally occurring data, such as in this research, it is not possible to control variables such as number of participants per time interval or per number of repetitions.

This natural experiment raises issues for test users such as testing center administrators to consider. Messick (1989) says that, while test developers are not responsible for effects such as test taker fatigue and remembering test items through test practicing, developers are responsible to determine and communicate to test users appropriate time intervals between test administrations to limit such effects, and tests users are responsible for administering the test following test administration guidelines. Specifically, in these administrations of paper versions of the TOEFL, the ETS time interval protocol should have been followed. Allowing students to take the TOEFL multiple times within a month, or even within a week, resulted in score fluctuations unlikely to have been caused by change in actual language ability.

These findings also raise questions about interpretation of test results by test users such as university admissions or placement officials. Wilson (1987) cautions that PBT scores could increase with repetition because of improvement in English language proficiency but also because of test-wiseness and test familiarity. In light of the score variability (including decreases) occurring with short time intervals documented in this research, test users should evaluate paper TOEFL scores of TOEFL repeaters cautiously. Because of score fluctuation including decreases when the paper TOEFL was repeated frequently (especially five–ten times) and/or in short time intervals (less than once a month and particularly within one week), test users should consider the profile of the test taker's TOEFL attempts to obtain a more accurate measure of language proficiency. A gradual increase in TOEFL score is in keeping with an increase in proficiency. However, extreme and/or numerous fluctuations in test scores with frequent attempts, particularly in short time intervals, may indicate test scores are affected by factors such as fatigue from test repetition, negative affect-related factors, and/or test-wiseness and test familiarity.

The greatest value, though, of this natural experiment is in raising questions rather than providing generalized conclusions. Thus the following questions are posed:

- In light of the positive relationship between number of attempts and percentage of score decreases, contrary to effects of practice, familiarity, and test-wiseness, and similar to fatigue from test repetition mentioned by Koretz (2008) and Meyer (2010), should number of attempts be part of test administration policy? Can extensive test repetition affect test taker scores? What is "extensive" test repetition? How many test repetitions in what time intervals would yield negative effects?
- In light of the volatility of scores within seven days seen in this research, even though the ITP time interval policy is currently determined by institutions (Educational Testing Service 2019c), should time interval for ITP be one calendar month if used for high-stakes purposes such as university admission?
- What would motivate test takers to repeat the TOEFL so often? So close together in time? Would reasons for repeating the TOEFL make a difference in the effect of number of attempts or time interval?
- Why did some of these test takers repeat the TOEFL so extensively? Do other test takers do so (as seen in Wilson's study)?
- What is the role of factors such as parental pressure, peer pressure, wanting to practice the test, hoping to get the required score, or hoping to get the same questions?
- Are test takers who frequently repeat the TOEFL affected by societal views that the test score is what is important instead of the skill the score should reflect?

References

Alderman, D. L. (1981). *Student self-selection and test repetition* (College Board Report 81-5). New York: College Entrance Examination Board.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bailey, K. (1998). *Learning about language assessment: Dilemmas, decisions, and directions.* Pacific Grove, CA: Heinle & Heinle.

Barkaoui, K. (2018). Examining sources of variability in repeaters' L2 writing scores: The case of the PTE Academic writing section. *Language Testing*, 1–23. https://doi.org/10.1177/0265532217750692.

Cizek, G. J., & Burg, S. S. (2006). *Addressing test anxiety in a high-stakes environment: Strategies for classrooms and schools.* Thousand Oaks, CA: Corwin Press.

Cliffordson, C. (2004). Effects of practice and intellectual growth on performance on the Swedish Scholastic Aptitude Test (SweSAT). *European Journal of Psychological Assessment, 20*(3), 192–204. https://doi.org/10.1027/1015-5759.20.3.192.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). Westport, CT: American Council on Education and Praeger.

Dancey, C. P., & Reidy, J. (2004). *Statistics without maths for psychology: Using SPSS for Windows*^{TM} (3rd ed.). London: Pearson.

Educational Testing Service. (2012). *TOEFL*. http://www.ets.org/toefl/important_update/pbt_ending. Accessed 8 February 2012.

Educational Testing Service. (2018). *Select your TOEFL® testing location*. http://www.ets.org/toefl. Accessed 5 January 2019.

Educational Testing Service. (2019a). *About the TOEFL ITP® assessment series*. http://www.ets.org/toefl. Accessed 5 January 2019.

Educational Testing Service. (2019b). *About the TOEFL® PBT test*. http://www.ets.org/toefl. Accessed 5 January 2019.

Educational Testing Service. (2019c). *Administration and scoring for the TOEFL ITP® test*. https://www.ets.org/toefl_itp/administration_scoring. Accessed 5 January 2019.

Educational Testing Service. (2019d). *The TOEFL® test*. https://www.ets.org/s/toefl_itp/pdf/toefl_itp_score.pdf. Accessed 5 January 2019.

Elliot, A. J., & McGregor, H. A. (1999). Test anxiety and the hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 76*(4), 628–644.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Hale, G. A., Angelis, P. J., & Thibodeau, L. A. (1980). *Effects of item disclosure on TOEFL performance* (TOEFL Research Report 8). Princeton, NJ: Educational Testing Service.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Boston: Heinle & Heinle Publishers.

Henning, G. (1993). *Test-retest analyses of the TOEFL® test* (TOEFL Research Report 45). Princeton, NJ: Educational Testing Service.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn and Bacon.

Kingston, N., & Turner, N. (1984). *Analysis of score change patterns of examinees repeating the Graduate Record Examinations General Test* (GRE Board Professional Report No. 83-5P). Princeton, NJ: Educational Testing Service.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practices, 17*(2), 28–30.

MacIntyre, P. D. (1995). How does anxiety affect second language learning? A reply to Sparks and Ganschow. *The Modern Language Journal, 79*(1), 90–99.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215–237.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241–256.

Meyer, P. (2010). *Understanding measurement: Reliability* [e-book]. New York: Oxford University Press USA. https://doi.org/10.1093/acprof:oso/9780195380361.001.0001.

Powers, D. E., & Lall, V. (2013, October). *Supporting an expiration policy for reporting TOEFL® scores* (Research Memorandum). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (2003). Standardized tests and high-stakes assessment. In J. W. Guthrie (Ed.), *Encyclopedia of education* (Vol. 7, 2nd ed., pp. 2533–2537). New York: Macmillan Reference USA.

Steinborn, M. B., Flehmig, H. C., Westhoff, K., & Langner, R. (2009). Differential effects of prolonged work on performance measures in self-paced speeded tests. *Advances in Cognitive Psychology, 5,* 105–113. https://doi.org/10.2478/v10053-008-0070-8.

Swinton, S. S., Wild, C. L., & Wallmark, M. M. (1983). *Investigation of practice effects on item types in the Graduate Record Examinations Aptitude Test: Practice effect*. http://www.ets.org/Media/Research/pdf/RR-82-56-Swinton.pdf. Accessed 5 January 2019.

Wilson, K. M. (1987). *Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language*™ (TOEFL Research Report No. 22). Princeton, NJ: Educational Testing Service.

Zeidner, M. (1998). *Test anxiety* [e-book]. New York: Plenum Press. Available via http://site.ebrary.com.ezproxy.aus.edu/lib/aus/docDetail.action?docID=10048275.

Zhang, Y. (2008). *Repeater analysis for TOEFL® iBT* (Research Memorandum). Princeton, NJ: Educational Testing Service.

Part II
Learning from Tests of World Languages

The main theme of this part is about how we can learn from thinking about testing world languages. This part starts with three experience-based papers (Chapters 12–14). The one data-based paper in this part is Chapter 15. The chapters in Part II are as follows:

- The first (Chapter 12), "Whose English(es) Are We Assessing and by Whom?" by Cheng, Im, and Jabeen, examines the English language testing construct from the perspectives of international contexts. They suggest a number of interesting communication features that both test designers and users should ponder.
- The next (Chapter 13), "Challenges in Developing Standardized Tests for Arabic Reading Comprehension for Secondary Education in the Netherlands," by de Graaf, confronts challenges that occur in testing Arabic language reading comprehension in a CEFR-type framework. She demonstrates using the Dutch language tests specifications and test development approaches for meeting such challenges.
- The final experience-based paper (Chapter 14), "The Conflict and Consequences of Two Assessment Measures in Israel: Global PISA vs. The National *MEITZAV*," by Inbar-Lourie and Shohamy, examines the negative washback effect described in the literature of two standardized tests used in Israel and the massive external testing policies surrounding them. They then argue for using a critical assessment literate view to improving such large-scale national/international tests and how they are used.
- (Chapter 15), "How to Challenge Prejudice in Assessing the Productive Skills of Speakers of Closely Related Languages (the Case of Slovenia)," by Ferbežar and Likar Stanovnik, explains problems related to bias in rating the Slovenian language essays of speakers of other South Slavic languages. They suggest incorporating foreignness and prejudice issues when thinking about language testing fairness.

All four chapters are related to the central theme of this part and to each other in that they address tests in languages other than British, Australasian, and North American (BANA) varieties. Chapter 12 deals with other global Englishes (including World Englishes, English as an international language, and English as a lingua franca); Chapter 13 describes using a Dutch test as a model for an Arabic test;

Chapter 14 deals with two standardized tests in Israel administered in Hebrew and Arabic. And, Chapter 15 addresses problems in Slovenian productive language tests when administered to speakers of closely related South Slavic languages. Three of the Chapters (12–14) are related to each other in that they are experience-based, while Chapter 15 is data-based.

Chapter 12
Whose English(es) Are We Assessing and by Whom?

Liying Cheng, Gwan-Hyeok Im, and Rubaiyat Jabeen

Abstract In this chapter, we review the conceptualizations of English from Standard English to English as a lingua franca. Using this lense, we evaluate the current test design and testing practices in terms of construct representation in international environments (i.e., Im & Cheng 2019) and a university setting (i.e., Jabeen 2016). We problematize the challenges that language testing currently faces through these two empirical studies and, provide solutions and suggestions to cope with the challenges. By doing so, this chapter provides test developers and test users with the aspects of communication that are important in designing language testing and assessing English language proficiency of non-native speakers.

12.1 Introduction: Purpose and Testing Context

English is now increasingly used as a contact language among speakers of different first languages, i.e., English as a lingua franca (ELF) (Jenkins 2009). As a result, the number of multilingual users of English has surpassed that of native speakers of English (McKay 2002), and such use of English has led to the increasing diverse and emergent nature of ELF communication. However, the current international English language assessments such as the Test of English as a Foreign Language (TOEFL), the Test of English for International Communication (TOEIC), and the International English Language Testing System (IELTS) have shown little inclination, so far, to take ELF communication into account in their test design, administration, and scoring. Instead, these assessments continue to assess candidates' English language proficiency with reference to the norms of native speakers of English (Jenkins and Leung

L. Cheng (✉) · G.-H. Im · R. Jabeen
Queen's University, Kingston, ON, Canada
e-mail: liying.cheng@queensu.ca

G.-H. Im
e-mail: gwan.im@queensu.ca

R. Jabeen
e-mail: 17rj3@queensu.ca

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_12

2014). This may result in insufficient reflection of the use of English in international communication, and consequently lack of validity in score interpretations and uses of such English assessments.

English has been the dominant language for international communication over the past few decades (Charles 2007; Ehrenreich 2010). Within this global context, English has been conceptualized in broadly four phases: Standard English (SE), World Englishes (WE), English as an international language (EIL), and English as a lingua franca (ELF).

Standard English (SE) is defined as the variety of English that educated English speakers use for writing and speaking (Trudgill and Hannah 1995) in British-Australian-North American (BANA) communities (Pakir 2009). English is conceived as a stabilized and standardized language leased out on a global scale and controlled by native English speakers (NESs) (Widdowson 2003). Hence, English produced by non-native English speakers (NNESs) has been described as learners' English or interlanguage (Kim 2012), and their English has been defined as English as a second language (ESL) or English as a foreign language (EFL) (Kim 2012). The goal of NNESs is to communicate with NESs and to assimilate into NESs communities (Sato 2014).

However, an alternative view on English was recognized by researchers as *World Englishes* (WE). Pluricentricity of English is emphasized, accepting that language spreads and adapts itself to new environments (Mollin 2006). Specifically, postcolonial Englishes such as Nigerian, Indian, and Singaporean English are recognized as English varieties in WE. WE is defined as "non-native models of English [that] are linguistically identifiable, geographically definable" (Kachru 1992, pp. 357–358), and therefore studies on the WE linguistically focused on types of English varieties. The use of English was categorized by Kachru (1988) in three different groups of English users (Kachru 1992): Inner Circle (English as a native language, e.g., British, Australian, and North American countries), Outer Circle (English as an official language or second language, e.g., Nigeria, India, and Singapore), and Expanding Circle (English as a foreign language, e.g., China, Japan, Korea, and Brazil).

This model is considered useful and influential due to the depiction of the spread of English. However, several limitations of the WE have been acknowledged (Jenkins 2009). Kachru's (1988) model recognizes the English used by the Expanding Circle groups as English as a foreign language (EFL) users. This implies that EFL speakers may need to acquire the native speaker-based communicative competence. Jenkins (2002) argues conceptually that the Expanding Circle users are not foreign speakers of English, but rather international speakers, and therefore the target community should not be the Inner Circle of native English speakers. Furthermore, the WE may lead to the potential confusions as the term WE can be interpreted in the following ways: (1) as an umbrella label to include any kinds of English varieties (Bolton 2004) and (2) as a term to refer to varieties of English used and codified in the Outer Circle group countries (Bolton 2004), as well as an approach to the study of English worldwide, specifically within applied linguistics (Jenkins 2006). As the WE mainly focus on English varieties, especially in the Outer Circle, Canagarajah (2013) pointed out that the WE failed to consider how English users in each circle communicate with each other. Therefore, the term WE may have a limitation to conceptualize the use of English in international communication.

Going beyond WE, English was labeled as an international language (EIL) (Canagarajah 2013). EIL treats English varieties in the same circle as having equal status, e.g., East Asian or South Asian English in the Expanding Circle and Inuit, Quebec, and Athabascan English within Canadian English in the Inner Circle (McArthur 1987). Canagarajah (2013) pointed out that EIL simply added English varieties to WE and has the same issues in WE in terms of the contact between English users in the three circles.

Most recently, *English as a lingua franca* (ELF) (Jenkins 2006; Seidlhofer 2011) has been adopted to address the limitations in WE to go beyond local to global contexts of the use of English, although both WE and ELF commonly recognize the existence of English varieties and the use of English beyond native English speakers (Saraceni 2008). ELF is defined as the use of English in international communication. The term has been defined by a number of researchers in quite different ways in terms of inclusion of native English speakers in ELF users. Some (e.g., Firth 1996; House 2003) limited ELF to the use of English only by non-native English speakers while others (e.g., Jenkins 2006; Seidlhofer 2011) include all English users in ELF. Specifically, Seidlhofer (2011) defined ELF as "any use of English among speakers of different first languages for whom English is the communicative medium of choice, and often the only option" (p. 7).

In fact, ELF has been researched and conceptualized in two traditions according to Canagarajah (2013): (a) codifying linguistic features of English varieties in international communication (hereafter, tradition A) (Jenkins 2006, Seidlhofer 2011) and (b) focus of pragmatics in the use of English (hereafter, tradition B) (Firth 1996; House 2003). Scholars in tradition A attempted to codify regular features of ELF used by non-native English speakers in terms of phonology (Jenkins 2000), phraseology (Prodromou 2008), and lexicogrammar (Cogo and Dewey 2012) and built the spoken ELF corpora: the Vienna-Oxford International Corpus of English (Seidlhofer 2011) and the Helsinki-based corpus of English as a Lingua Franca in Academic Settings (Mauranen and Ranta 2008). However, this practice of codifying ELF has been criticized in terms of creating another English norm used by non-native English speakers (Swan 2012). On the other hand, ELF research in tradition B (e.g., Bjørge 2010; Firth 1996; House 2003) focused on pragmatics (i.e., speakers' management to adjust language use considering contextual factors, using linguistic resources to achieve the communicative goals). Firth (1996) named this tradition B of ELF as *Lingua franca English* (LFE) with more focus on the function (use) of English for communication. Gramkow Andersen (1993) pointed out that there are no common linguistic features in ELF, but rather variabilities in the speakers' English proficiency, and accordingly the speakers seem "to negotiate and govern their own variety of lingua franca use in terms of proficiency level, use of code-mixing, degree of pidginization, etc." (Gramkow Andersen 1993, p. 108).

12.2 Testing Problem Encountered

As mentioned above, the English language is the most widely acquired second language globally today. For more than two decades, it has been an unchallenged lingua franca, and the most common means of global communication (Crystal 1997; Kachru 1992; Medgyes 1994; Prodromou 1992). According to a TESOL International Association Annual Report (2014), there are 1.5 billion English learners worldwide. This number is rapidly growing in the contexts of both English as a foreign language (EFL) and English as a second language (ESL). Despite the fact, international tests continue to assess candidates' English language proficiency with reference to the norms of native speakers of English (Jenkins and Leung 2014). This may result in insufficient reflection of the use of English(es) in international communication, and consequently lack of validity in score interpretation and use of such English assessments.

This section below problematizes the challenges that language testing currently faces through two recent studies (Im and Cheng 2019; Jabeen 2016). Drawing on contemporary validation frameworks, Im and Cheng (2019) evaluated how well the Test of English for International Communication (TOEIC) reflects ELF communication in international business workplaces using an analysis of the testing organization's official documents. They found TOEIC's construct to be under-representative of this fast growing ELF context. Jabeen (2016) investigated the assessment practices and perceptions of native English-speaking teachers (NESTs) and non-native English-speaking teachers (NNESTs) when they rated a university English as a second language (ESL) student's oral presentation. The findings showed a disparity between the actual practices and articulated perceptions within each group of raters. This study put forward new nuanced understanding of language assessments within the complex ELF communication in higher education.

12.2.1 The Case of TOEIC's Construct Representation

As multinational companies emphasize employees' English communication skills and require prospective employees to submit English test scores, the TOEIC, which is composed of listening, reading, speaking, and writing components, is used by English language learning programs and government agencies in over 150 countries, as well as by roughly 14,000 companies around the world (Educational Testing Service [ETS] 2016a, para. 4).

The TOEIC measures English skills of non-native speakers working in international environments (ETS 2016b) and has been used for high-stakes decisions such as hiring, promotion, and overseas assignments in international business workplaces. With this use of TOEIC scores for high-stakes decisions, ETS has put much effort into modifying and revising TOEIC items over the past 30 years. Despite the ETS's effort, concerns have been raised about the accuracy of TOEIC scores (Jenkins and Leung

2014), due to the test's potentially limited reflection of the actual use of English in international environments (Leung and Lewkowicz 2012). This important issue must be addressed to ensure employers can make valid judgments about an employee's English language proficiency, thereby avoiding gratuitous negative consequences on the individuals, the company, and the society.

As a preliminary study through document analysis, Im and Cheng (2019) examined construct representation of the TOEIC using Kane's (2013) interpretation and use argument (IUA). Kane's (2013) IUA provides a systematic and practical guideline for test validation as it frames the scoring, generalization, extrapolation, and decision rules inferences to be investigated during test validation. Among these inferences, the extrapolation inference was the main focus in Im and Cheng (2019) because this inference pertains to the test's construct representation of English language proficiency in international business communication environments. To evaluate the extent to which the TOEIC reflects the required English language proficiency in the environments, Im and Cheng (2019) reviewed the test formats of each component of the TOEIC through analyzing ETS's official documents publicly available online.

The authors found that the TOEIC Listening includes the variety of English in British-Australian-North American (BANA) and New Zealand communities and further some colloquial forms and fragments of full sentences (ETS 2016b). In the TOEIC Reading, text messages and online chat dialogues for passages in the test are included to reflect the actual use of English widely used in international business contexts. In the TOEIC Speaking, one distinct feature in the criteria for evaluating test takers' responses from other English tests such as the IELTS includes the intelligibility, while other criteria such as accuracy and cohesion are similar to those in the IELTS. Those test takers' responses to TOEIC Speaking and Writing questions are evaluated by raters who are residents in the North American countries and have teaching experiences for non-native English speakers, on the platform of the Online Network Scoring system, which allows raters to go through a calibration test every time they rate.

Through reviewing the TOEIC format in each component, a couple of issues in the TOEIC were noticed in relation to the extrapolation inference when it comes to English as a lingua franca. For example, the authors noted that the TOEIC Speaking test may not fully capture the important aspects of communication in international business, which are interactive communication skills required for communications among English users of different first languages. Interactive skills include meaning negotiation, accommodation, and repair for successful communication (Louhiala-Salminen et al. 2005). However, as it is administered through a computer, the TOEIC Speaking test may not fully measure these skills. In addition, although the intelligibility in the criteria of the TOEIC Speaking is distinct, the definition of intelligibility has not been empirically supported (Kang et al. 2018). As raters for evaluating test takers' responses to TOEIC Speaking questions are recruited only from the North American countries, it remains unknown what portion of raters of different first languages is accounted for in the rater pool and how fairly the existing raters from the North American countries could evaluate test takers' responses in terms of how well the test takers could achieve successful communication in ELF contexts, which requires an ability to understand different varieties of English in terms of "different accents, different syntactic forms and different discourse styles" (Harding and McNamara 2017, p. 577). Based on this definition of ability in ELF contexts, another issue to be concerned with is that the TOEIC Listening only includes accents from English-speaking communities. Although the TOEIC is an international English test as its name represents, only inclusion of so-called Standard English varieties may lead to construct underrepresentation of the actual use of English in international environments.

12.2.2 The Case of Native and Non-Native English-Speaking Teachers' Assessment of University ESL Student's Oral Presentation

Due to the increasing use of English as a lingua franca (ELF), intelligibility of speech can be more important than having native English-like pronunciation in meaningful communication. However, in addition to standardized English language proficiency testing, many postsecondary classroom assessments also continue to assess international students' English language proficiency with reference to norms of native speakers of English. Jabeen (2016) demonstrates how English language teachers practice traditional norms of assessment even when they perceive the intelligibility of speech as more important than Standard English grammar and pronunciation.

Today, not only the number of international English language learners but also the number of international English language teachers is increasing rapidly worldwide. Largely, English language teachers can be divided into two groups: native English-speaking teachers (NESTs) and non-native English-speaking teachers (NNESTs). The major distinguishing factor between the NESTs and the NNESTs is their first language (L1) (Liu 1999; McKay 1992; Paikeday 1985). In Jabeen (2016), the teachers who self-identified English as their L1 were regarded as NESTs and the teachers who self-identified a different language as their L1 were regarded as NNESTs.

In light of the advent of the concepts of WE, EIL, and ELF as mentioned at the beginning of the chapter, there has been a shift from traditional pedagogies to the adoption of Communicative Language Teaching (CLT) in the field of English language teaching and learning. CLT is an approach to language teaching that is based on the theory that the primary function of language is meaningful communication (Brandl 2008). Therefore, to investigate the NESTs' and NNESTs' assessment practices and perceptions regarding the assessment of English language speaking skills of ESL students, Jabeen (2016) conducted a study at a university in the United States of America.

The study was conducted on a group of 19 NEST and 12 NNEST raters. The data were collected using an online survey consisting of two sections. Section one was designed to explore the actual assessment practices of the NEST and NNEST raters. It comprised of a video-recording of an ESL student's oral presentation in a university classroom, an analytic rating scale, and a segment for comprehensive written feedback. Section two contained a questionnaire designed to elicit the NEST and NNEST raters' perceptions regarding the assessment criteria on the rating scale in section one. The study did not reveal any statistically significant difference between the rating practices of the NEST and NESTS rater-participants. However, a difference did emerge between their actual assessment practices and their perceptions regarding assessment criteria. While rating and providing feedback to the ESL student's oral presentation, the study found that both the NEST and NNEST raters emphasized having native-like grammar and pronunciation in speech as most important. Interestingly, however, when it came to the raters' perceptions, the study found that both groups of raters perceived intelligibility and comprehensibility of speech as more important than grammatical accuracy or native-like pronunciation in the ESL student's oral presentation.

This disparity between the raters' actual assessment practice, which adhered to traditional norms of correcting grammar and pronunciation, and their perception in which intelligibility of speech is more important, clearly implies that the English language teachers need to revisit and reevaluate their assessment practices in light of their perceptions.

12.3 Solution/Resolution of the Problem and Insight Gained

Overall, both studies pointed out the critical yet complex needs for test developers and test users to pay closer attention to the changing landscape of English on a global scale. The complex definition of key terms of English, i.e., Standard English (SE), World Englishes (WE), English as an international language (EIL), and more recently, English as a lingua franca (ELF) conceptualized in two traditions is only the beginning of our increasing understanding of this changing landscape. In this sense, language testing researchers are only starting to conduct empirical studies to examine what the changing landscape means to the testing of English on a global scale (Cheng et al. 2018).

One major concern identified from Im and Cheng (2019) was TOEIC's potentially limited reflection of the use of English in international business environments. To address this kind of concern in other contexts, some empirical studies (e.g., Major et al. 2002) have examined whether the inclusion of non-native English speakers' accents in the listening section in the Test of English as a Foreign Language (TOEFL) affected listening comprehension of native and non-native English speakers. Harding (2012) also investigated whether test takers gained advantages from shared first language speakers in the listening sub-test of the University Test of English as a Second Language (UTESL). Both studies showed that non-native English speakers' varieties discriminated against a certain group of test takers, which threatens the validity of score meaning. On the other hand, some studies (e.g., Taylor 2002; Xi and Mollaun 2011) on inclusion of non-native English speakers as raters in high-stakes testing showed that non-native English speaker raters performed as well as native speaker raters. McNamara (2012, 2014), however, calls for a revolutionary change of conceptualization of English language proficiency in terms of successful communication rather than gradual application of ELF to the competence of native English speakers. This is true of the TOEIC. This important issue must be addressed to ensure that employers can make valid judgments about an employee's English skills thereby avoiding gratuitous negative consequences on the individuals, the companies, and the society. However, ELF constructs including definitions of intelligibility have not been defined with sufficient empirical studies to support the definitions (e.g., Harding and McNamara 2017). To successfully adopt ELF into language testing, more empirical studies on the definitions of ELF abilities including intelligibility need to be carried out in language testing.

12.4 Conclusion

This chapter has illustrated how English has been conceptualized over the past few decades from recognizing English varieties (WE, EIL, Tradition A of ELF) to focusing on pragmatics (Tradition B of ELF). The latter calls for changing constructs in language testing to reflect the use of English in international environments, as failure to do so may result in potentially inappropriate decisions about test takers' English language proficiency and therefore negative consequences on the test takers who take the tests, the organizations who use the test scores, and the society where a testing system is situated. To address the issues in language testing, two studies (e.g., Im and Cheng 2019; Jabeen 2016) problematized the challenges that language testing currently faces in terms of the TOEIC's potentially limited reflection of the use of English in international business workplaces in relation to Kane's (2013) extrapolation inference, and the disparity between NEST and NNEST raters' marking practices and their perceptions of assessment criteria. In addition, this chapter also provided potential solutions and suggestions to cope with the challenges. Both test developers and test users are required to pay closer attention to the current conceptualization of English (e.g., tradition B of ELF) which McNamara (2012, 2014) calls for in terms of a revolutionary change of constructs in language testing. Based on English as a lingua franca, more empirical studies are being conducted, e.g., assessing workplace listening comprehension based on English as a lingua franca and the aspects of the perceived intelligibility of second language speakers. Such empirical studies will provide alternatives to cope with the challenges in language testing for test developers and test users to ensure the validity of score interpretations and uses in language testing.

References

Bjørge, A. K. (2010). Conflict or cooperation: The use of backchannelling in ELF negotiations. *English for Specific Purposes, 29,* 191–203. https://doi.org/10.1016/j.esp.2009.04.002.

Bolton, K. (2004). World Englishes. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 369–396). Oxford: Blackwell.

Brandl, K. (2008). *Communicative language teaching in action: Putting principles to work*. Upper Saddle River, NJ: Prentice Hall.

Canagarajah, A. S. (2013). *Translingual practice: Global Englishes and cosmopolitan relations.* New York, NY: Routledge.

Charles, M. L. (2007). Language matters in global communication: Article based on ORA lecture, October 2006. *Journal of Business Communication, 44,* 260–282. https://doi.org/10.1177/0021943607302477.

Cheng, L., Im, G.-H., & Jabeen, R. (2018, July). *Whose English(es) are we assessing and by whom?* Paper presented at the 2018 International Test Commission Conference, Montréal, Québec Canada.

Cogo, A., & Dewey, M. (2012). *Analysing English as a lingua franca: A corpus-driven investigation.* London: Continuum International Publishing.

Crystal, D. (1997). *English as a global language.* Cambridge: Cambridge University Press.

Educational Testing Service. (2016a). *Examinee handbook speaking & writing.* https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_Speaking_and_Writing_Examinee_Handbook.pdf. Retrieved December 20, 2017.

Educational Testing Service. (2016b). *The TOEIC® tests—The global standard for assessing English proficiency for business.* https://www.ets.org/toeic/succeed. Retrieved December 20, 2017.

Ehrenreich, S. (2010). English as a business lingua franca in a German multinational corporation. *Journal of Business Communication, 47,* 408–431. https://doi.org/10.1177/0021943610377303.

Firth, A. (1996). The discursive accomplishment of normality: On 'lingua franca' English and conversation analysis. *Journal of Pragmatics, 26,* 237–259. https://doi.org/10.1016/0378-2166(96)00014-8.

Gramkow Andersen, K. (1993). *Lingua franca discourse: An investigation of the use of English in an international business context*. Unpublished Master's thesis, Aalborg University, Denmark.

Harding, L. (2012). *Language testing, World Englishes and English as a lingua franca: The case for evidence-based change.* Invited presentation at the Center for Internationalization and Parallel (CIP) Language Use Symposium, University of Copenhagen, Copenhagen. https://cip.ku.dk/arrangementer/tidligere/symposium_2012/Luke_Harding.pdf. Retrieved July 20, 2019.

Harding, L., & McNamara, T. (2017). Language assessment: The challenge of ELF. In J. Jenkins, W. Baker, & M. J. Dewey (Eds.), *Routledge handbook of English as a lingua franca* (pp. 570–582). London: Routledge.

House, J. (2003). English as a lingua franca: A threat to multilingualism? *Journal of Sociolinguistics, 7,* 556–578. https://doi.org/10.1111/j.1467-9841.2003.00242.x.

Im, G.-H., & Cheng, L. (2019). The test of English for international communication. *Language Testing, 36,* 315–324. https://doi.org/10.1177/0265532219828252.

Jabeen, R. (2016). *An investigation into native and non native English speaking instructors' assessment of university ESL student's oral presentation.* All Theses, Dissertations, and Other Capstone Projects. Paper 647.

Jenkins, J. (2000). *The phonology of English as an international language: New models, new norms, new goals.* Oxford: Oxford University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23,* 83–103. https://doi.org/10.1093/applin/23.1.83.

Jenkins, J. (2006). Current perspectives on teaching world Englishes and English as a lingua franca. *TESOL Quarterly, 40,* 157–181. https://doi.org/10.2307/40264515.

Jenkins, J. (2009). English as a lingua franca: Interpretations and attitudes. *World Englishes, 28,* 200–207. https://doi.org/10.1111/j.1467-971x.2009.01582.x.

Jenkins, J., & Leung, C. (2014). English as a lingua franca. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1605–1616). Chichester: Wiley. https://doi.org/10.1002/9781118411360.wbcla047.

Kachru, B. B. (1988). The sacred cows of English. *English Today, 16,* 3–8. https://doi.org/10.1017/s0266078400000973.

Kachru, B. B. (1992). Teaching world Englishes. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 355–366). Champaign, IL: University of Illinois Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73. https://doi.org/10.1111/jedm.12000.

Kang, O., Thomson, R. I., & Moran, M. (2018). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*, 1–29. https://doi.org/10.1093/applin/amy053.

Kim, H. (2012). *Exploring the construct of aviation communication: A critique of the ICAO language proficiency policy*. Unpublished doctoral dissertation, University of Melbourne, Melbourne, Australia.

Leung, C., & Lewkowicz, J. (2012). Language communication and communicative competence: A view from contemporary classrooms. *Language and Education, 27*(5), 1–17. https://doi.org/10.1080/09500782.2012.707658.

Liu, J. (1999). Nonnative-English-speaking professionals in TESOL. *TESOL Quarterly, 33*(1), 85–102. https://doi.org/10.2307/3588192.

Louhiala-Salminen, L., Charles, M., & Kankaanranta, A. (2005). English as a lingua franca in Nordic corporate mergers: Two case companies. *English for Specific Purpose, 24,* 401–421. https://doi.org/10.1016/j.esp.2005.02.003.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*(2), 173–190. https://doi.org/10.2307/3588329.

Mauranen, A., & Ranta, E. (2008). English as an academic lingua franca: The ELFA project. *Nordic Journal of English Studies, 7,* 199–202. https://doi.org/10.1016/j.esp.2009.10.001.

McArthur, T. (1987). The English languages? *English Today, 3,* 9–13. https://doi.org/10.1017/s0266078400013511.

McKay, S. L. (1992). *Teaching English overseas: An introduction.* Oxford: Oxford University Press.

McKay, S. L. (2002). *Teaching English as an international language.* Oxford: Oxford University Press.

McNamara, T. (2012). English as a lingua franca: The challenge for language testing. *Journal of English as a Lingua Franca, 1*(1), 199–202. https://doi.org/10.1515/jelf-2012-0013.

McNamara, T. (2014). 30 years on—evolution or revolution? *Language Assessment Quarterly, 11,* 226–232. https://doi.org/10.1080/15434303.2014.895830.

Medgyes, P. (1994). *The non-native teacher*. London: Macmillan Publishers.

Mollin, S. (2006). English as a lingua franca: A new variety in the new expanding circle? *Nordic Journal of English Studies, 5,* 41–57.

Paikeday, T.M. (1985). *The native speaker is dead!* Toronto, ON: Paikeday Publishing.

Pakir, A. (2009). English as a lingua franca: Analyzing research frameworks in international English, world Englishes, and ELF. *World Englishes, 28,* 224–235. https://doi.org/10.1111/j.1467-971x.2009.01585.x.

Prodromou, L. (1992). What culture? Which culture? Cross–cultural factors in language learning. *ELT Journal, 46*(1), 39–50. https://doi.org/10.1093/elt/46.1.39.

Prodromou, L. (2008). *English as a lingua franca: A corpus-based analysis.* London: Continuum.

Saraceni, M. (2008). English as a lingua franca: Between form and function. *English Today, 94*(24), 20–26. https://doi.org/10.1017/s0266078408000163.

Sato, T. (2014). *Linguistic laypersons' perspective on second language oral communication ability*. Unpublished doctoral dissertation, The University of Melbourne, Melbourne, VIC, Australia.

Seidlhofer, B. (2011). *Understanding English as a lingua franca*. London: Oxford University Press.

Swan, T. (2012). ELF and EFL: Are they really different? *Journal of English as a Lingua Franca, 1,* 379–389. https://doi.org/10.1515/jelf-2012-0025.

Taylor, L. (2002). Assessing learners' English: But whose/which English(es)? *Research Notes, 10*, 18–20. https://www.cambridgeenglish.org/images/23124-research-notes-10.pdf. Retrieved July 20, 2019.

TESOL International Association. (2014). *Explore sustain renew*. [Annual Report]. https://www.tesol.org/docs/default-source/annual-reports/2014-annual-report.pdf?sfvrsn=0. Retrieved July 20, 2019.

Trudgill, P., & Hannah, J. (1995). *International English: A guide to varieties of Standard English* (3rd ed.). London: Edward Arnold.

Widdowson, H. G. (2003). *Defining issues in English language teaching*. Oxford: Oxford University Press.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222–1255. https://doi.org/10.1111/j.1467-9922.2011.00667.x.

Chapter 13
Challenges in Developing Standardized Tests for Arabic Reading Comprehension for Secondary Education in the Netherlands

Anneke de Graaf

Abstract Since the nineties, Arabic has been one of the official languages in Dutch secondary education, which in the Netherlands ends with national examinations. The national examinations of Arabic for the different educational levels consist of a reading comprehension test about a number of (more or less) authentic texts from all over the Arab world to which students have to apply different reading strategies to be able to answer the closed and open-ended questions. When preparing these examinations, the test constructors faced and still are facing different challenges: how to apply international quality standards concerning language assessment and frameworks, such as CEFR, to Arabic, which has a long history in Dutch academic education but not in secondary education as a second language, without neglecting requirements resulting from the test specifications for national Dutch language exams? The keywords in the process are *capacity building, international standards,* and *research.* Groups of teachers are involved in every step of the test construction cycle. They develop and screen items using standardized checklists on different aspects based on international assessment literature. Outcomes of standard setting procedures prove the alignment of the examinations to CEFR, and the interpretation of data after test administration provides additional information about the quality of the items. These experiences and outcomes of research can be transferred to other language assessment contexts to tackle step by step all challenges met when a new language is introduced in an educational context.

A. de Graaf (✉)
Cito, Arnhem, The Netherlands
e-mail: anneke.degraaf@cito.nl

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_13

13.1 Introduction: Purpose and Testing Context

13.1.1 Backgrounds of the Position of Arabic in the Dutch Educational Context

Since the nineties, Arabic has been one of the official languages taught in Dutch secondary education. Traditionally, English, French, and German were considered the main foreign or second languages in the educational system, but then other languages emerged in the Netherlands which lead to the introduction of Arabic, first in primary and then in secondary education. For centuries the situation of Arabic in Dutch education was dominated by the universities where the formal variant of Arabic was taught in relation to other classical languages such as Greek and Latin. Arabic was introduced at the University of Leiden in the seventeenth century. The interest in and the importance of Arabic resulted from a situation in which Arabic was considered one of the languages that could support, firstly, Biblical studies and later, during the so-called Orientalism period, also the study of the Koran and Islam. This situation lasted for centuries and only began to change in the twentieth century.

In the 1970s, Arabic began to play a role outside the universities in the Netherlands. The reason for the introduction of Arabic at the level of primary education was migration patterns during the postwar period which affected school populations; they became more heterogeneous, especially in the bigger cities. In 1974 Home Language Instruction (HLI) was introduced by the Dutch Ministry of Education for children from different ethnic backgrounds, to promote linguistic and cultural diversity. The goal of that policy was to help children when returning to their countries of origin, as the intention was that the migrant workers and their families would stay in the Netherlands for a limited period of time. In the case of Moroccan children, Modern Standard Arabic (MSA), the formal written variety of Arabic, was promoted and taught in primary education, and not (one of) the home languages of the Moroccan migrant children.

During the 1980s, however, it became clear that many migrant families in the Netherlands were staying permanently, so a new policy for their education was needed. At that time, the focus of the official educational policy for the teaching of HLI changed from returning to integration. The goal then shifted toward tools for communicating with the family and relatives, development of their identity, a positive self-image in the host country, and identification with the school and increasing school success. Although the focus was on being able to communicate with relatives and family, the target language still was MSA which has a focus on writing and reading.

During the entire period from the 1970s until 2004, the official Dutch educational policy in the context of HLI overlooked or ignored the important feature of diglossia in the Arab world: the fact that there is a gap between MSA and colloquial language. The difference or gap between MSA and the Moroccan variant of Arabic (daarija) is considerable. Even more complex was the language background of the Moroccan children who spoke one of the Berber languages at home. The situation of diglossia and the variety in language backgrounds, should in fact have influenced the choice for the selection of the target language for HLI since the 1970s (Nielsen 2009). In primary education in the Netherlands, this issue was never addressed, and the focus stayed on MSA during these years.

In 2004 the HLI system disappeared from Dutch primary education. However, Arabic as a teaching language has remained in secondary education since then. The focus in primary education was on MSA, but there was no HLI-context anymore. As the classes for Arabic in secondary education are taken by students with different backgrounds and home languages, the focus was and still is on the standardized variety of Arabic used for writing and reading in different Arabic countries. Textbooks were developed, and final exams introduced (de Graaf et al. 2011).

13.1.2 The Final Exams for Arabic in Secondary Education

All secondary school courses in the Netherlands end with final examinations. After passing these exams, pupils gain access to different types of further education. The Dutch secondary school system has three school types or tracks. Each school type ends with a final exam that is comprised of a state and a school exam. In all of these three school types, Arabic is an official subject that has the same status as the other languages that are taught in the Dutch education system, with the exception of English, which is obligatory for all students. Students are offered the choice between French, German, Arabic, Spanish, Turkish, and Russian. Exams in Arabic are developed for the following three tracks. The first track is the prevocational level in secondary education, the so-called VMBO which has three sublevels (since 1990). Most students are 16 years old when finalizing this track. Another track is the senior general secondary education (HAVO), and most students are 17 when taking these exams. Finally, there is the pre-university education track (VWO) where most students are 18 years old. For the last two levels, final exams have been developed since 1998.

The central national examination of Arabic consists of a reading comprehension test lasting 2.5 h and containing a number of authentic texts from all over the Arab world. The students need to show understanding of texts related to daily life situations, and need to be able to apply different reading strategies to answer the questions. This national examination can be taken at three sessions during the school year—in May, June, and August. All examinees take the examination in May. The June and August sessions are for pupils doing retakes, or who were unable to take the examination in May. The examinations are marked by the pupils´ own teachers and checked by a teacher from another school. Open-ended items are 30–40% of the total number of items; the majority of the items have a closed format.

A number of organizations and stakeholders are involved in the examination process. First, there is CvTE, the Dutch National Examination Board, which represents the Ministry of Education. CvTE bears the overall responsibility for the central exams. Cito, the Dutch Institute for Educational Measurement, is a contractor hired to develop the exams. An important part of the work is done by the test construction group, a team consisting of three teachers who write exam questions. They also work as teachers of Arabic in secondary education, but spend on average five hours a week constructing test questions. The involvement of teachers active in the field is important, as they are familiar with the content of the subject and pupil ability levels. The test construction group operates under the responsibility of the Cito subject matter specialist. The validation of the exams is done by validation groups of subject specialists in Arabic and higher education representatives (CvTE). The CvTE validation group monitors the construction process (de Graaf 2011).

13.2 Testing Problem Encountered: How to Operationalize Test Specifications for Arabic?

When developing national exams for Arabic, the test constructors have been facing several challenges: Can test specifications formulated for European languages be transferred to assessments of Arabic? How could assessment of Arabic in the Netherlands evolve from Arabic being first seen as a formal and static language at university level, then as a heritage L1, and now as one of the "modern languages" in Dutch education? In this new context Arabic was considered one of the modern languages for which central exams had to be developed—exams that met all the criteria that played a role for languages with a longer tradition in the secondary Dutch educational system when it comes to the concept of reading literacy.

When test developers start working in a test construction group for a specific level of Arabic exams, they are trained in the different aspects of test development. Language testing experts offer an introduction into some key values of testing like reliability, validity, transparency, etc. (Downling and Haladyna 2006). Besides this theoretical background, there is practical training devoted to the selection of texts that can be used for the operationalization of the can-do-statements, and exercises on the application of the golden rules of item development. Checklists cover both areas and are based on international literature about the dos and don'ts (e.g., on Rodriguez 1997) and on the experiences and needs of the teachers of Arabic.

13.3 Solution/Resolution of the Problem

What are the specific challenges confronted during this process when preparing national exams for Arabic? Experience showed that the concept of reading literacy, the selection of texts for reading comprehension, and defining the level of difficulty are the main issues for which additional attention, tools, and activities were needed.

13.3.1 Reading Literacy Within the Context of Exams for Arabic

In the Dutch national exams for languages, the focus in the reading parts is on the following processes in reading (see Table 13.1):

When translating the five criteria in the test specifications into exams for Arabic, issues were encountered that could sometimes be attributed to the specific situation of the Arabic language in the Dutch educational context, where a focus on the formal aspects of this language, and a lack of development of the communicative elements of teaching and assessing, were dominant. Training the teachers who become part of one of the test development groups, helps in getting them acquainted with the encompassing concept of reading literacy. At the time of the introduction of final exams for Arabic in the1990s, most of the Arabic language teachers working in secondary education were from the so-called first generation of immigrants and had most of their education in one of the countries of the Arab world. This background had consequences when they got involved in test development: In many of these countries the focus in reading is on the technical aspects (decoding, grammar, vocabulary) and not on reading literacy in which grammar and vocabulary play a role as building blocks needed for the operationalization of reading (strategies) and achieving reading goals of the individual learner. However, test items for reading literacy should not explicitly assess this knowledge of grammar and vocabulary. Essential training for these teachers is awareness of and discussion about the fact that vocabulary and grammar are important elements, for which students need knowledge to be able to decode, read, and understand a text.

Nowadays the younger teachers working in secondary education for Arabic have had their education and teacher training in the Netherlands, so the concept of reading literacy as operationalized in Dutch education, and more specifically in the exam system, is not new to them. But since there is no specific teacher training for teachers of Arabic in the Netherlands at the moment, new members of test development groups must be trained in this aspect of assessment within a communicative context, specifically for Arabic materials. For them it is sometimes a challenge to know how to apply communicative criteria for Arabic, as this approach differs from the way Arabic is taught and assessed in most Arabic educational traditions.

Table 13.1 Reading Literacy for the Arabic Language

Candidate can select relevant information related to a certain (given) need
Candidate can indicate the main element(s) of (a part of) the text
Candidate can indicate the meaning of important elements in the text
Candidate can identify relations between parts of the text
Candidate can draw conclusions related to intentions, opinions, and feelings of the author

Cito therefore provides relevant literature on what reading literacy is in the twenty-first century and shares good practices and exercises, to align Arabic with the other language exams in the Dutch exam context. PISA (Programme for International Student Assessment) released items which are inspirational material to serve as examples of what reading literacy is considered to be in a worldwide context, taking into account all innovations and developments resulting from, for example, the digital mode of test administration nowadays. Released exams for other languages contain examples of the types of texts and items aimed at for Arabic as well. Training with daily life text types (announcements, ads, train schedules, etc.) and items not focussing on formal aspects of texts, gives insight into the international standards in the domain of reading literacy and tools on how to apply this communicative approach to reading comprehension for Arabic.

13.3.2 Selection of Arabic Texts

One of the main principles of text selection for the exams of Arabic as formulated in the test specifications, is that texts have to be from sources representing the different countries of the Arab world, and that they preferably deal with issues related to Arab societies. Texts in all exam levels should reflect the diversity of the Arab world. So, although the majority of students in secondary education have a Moroccan background, the text topics represent the different countries where Arabic is spoken. This criterion has consequences for the test development practice. In the test development groups for Arabic, Cito prefers to work with teachers from different backgrounds to guarantee this diversity. This variety in backgrounds can help in the validity aspect of the exam: From preferences in topics, specific websites or newspapers from a certain region to slight variations in the standard language, the test development group has to define which topics and influences in written texts are regional and not understandable to all students, and what level of colloquial influence is acceptable for all students taking part in the exams.

Another issue when dealing with materials that can serve as a starting point for test development is, at the beginning of their careers as test developers, the strong focus of some teachers on grammatical errors in the "original" texts. The principle is that MSA is the target language, but also for this standard variety there are generally accepted developments in vocabulary, grammar and constructions, or local influences. Discussion in the test development group is needed to define to what level these new developments in the standard language are acceptable in an exam context. The status of Arabic is generally very high among people speaking a colloquial variety of Arabic and for many Muslims in general, it being the language of the Koran. To optimize the correctness of texts, some test developers have a strong preference for texts they have written themselves or adapted for educational goals, while one of the principles in the process of text selection is that texts are authentic, so materials from textbooks or educational websites are not allowed.

The high status of Arabic influences text selection in another way: Teachers tend to focus on topics that teach students good behavior or inform them about moral issues. But texts must align as much as possible with the interests of the students of a certain age group: The exam is not the occasion to learn more about topics or attitudes, but the moment to show competencies in reading literacy in this language. Thus, a variety of text types and communication topics is needed.

More sources with interesting and relatively depoliticized texts are available nowadays, as a result of the Arab Spring and the existence of the internet. This helps with avoiding topics that excite or trigger candidates while taking the exam. In particular, children from Syria and Lebanon should not be exposed to texts about the political situation in their countries. As the availability of the internet influences many aspects of daily life in the Arab world, for the test developers for Arabic exams, it has presented more possibilities to find texts without political, religious, or other bias-causing aspects that must be avoided in exams. Freedom of the press is still a difficult issue in the Arab world, and this influences the kind of texts that can be presented in the exams directly. Sometimes it is difficult to find texts that meet the standards Cito adheres to: correct language use, relevant and daily life topics, sharing different opinions, not too many enumerations, etc.

13.3.3 CEFR Levels of the Exams for Arabic

For test developers with a strong background in the Arabic language who were sometimes educated in one of the countries in the Arab world, it was and sometimes still is difficult to imagine how hard learning Arabic can be for students living in a community dominated by the Dutch language and other language(s) spoken at home, which is never MSA, as this is nobody's first language. The script and the structure of Arabic differ from European languages. At the start of secondary education, most students have to get instructions on the Arabic alphabet. In schools the population that takes lessons in Arabic preparing for the final exams, is very mixed: Refugee children from the Middle East join these lessons, while the majority consists of immigrant children from different countries in North Africa who were born in the Netherlands. Another relatively new group in the classrooms is the children without any knowledge of Arabic or background in the Arab world when they start in secondary education. This diverse population makes it quite hard for teachers to teach all students at their specific level, while leading them to the same goals at the end of secondary education. Proceeding from learning the (new) alphabet, decoding the language, and then going toward higher order thinking skills in the four language skills take time and effort.

The issue of the CEFR level of the final exams is directly related to the level of difficulty (Council of Europe 2001): should this level align with the beginners or with the students having limited knowledge of Arabic? Or should the final exams focus mainly on students with a strong background in Arabic? Dutch policymakers preferred to follow the first option which makes the subject of Arabic theoretically accessible to all students, as other languages in secondary education are. Experience in past years shows that the stronger beginners can participate in the final exam, and that the results for the children with a strong background in Arabic are relatively good. Thus the last group can value their knowledge of Arabic acquired in a context that is not school related.

During the meetings with the different test development teams, test and item analyses from previous exams are shared and evaluated with new test developers to make them aware of the levels of Arabic and the functioning of the exams. Because of the limited number of students per level (less than 100), test and item analyses do not provide stable data. That limitation makes the role of the framework that describes language proficiency levels crucial during the process of operationalization. But how can we prove that an exam targets a certain CEFR level instead of only making such a claim, to have a more accurate idea about the level of this exam? For this reason, standard setting meetings were organized in which teachers, curriculum developers, textbook writers, and representatives from the Dutch Ministry of Education participated to assess the CEFR level of exams.

The first CEFR standard setting was organized in 2008, with the aim of making the examination requirements for Arabic in the Netherlands more transparent for both teachers and pupils, and at the same time making it possible to compare the language achievements in Arabic of Dutch pupils internationally. For the test developers it was interesting to be able to study the possibilities of developing more comprehensive CEFR-related examinations of Arabic. The steps as outlined in the Manual published by the Council of Europe were followed and lead to a scientific claim about the links of an examination to the CEFR. The outcomes of the first standard setting organized in 2008 showed that from the lowest level of VMBO to the highest level VWO, the state examinations of Arabic refer to increasingly higher CEFR levels: A minimum level of A2 was needed for the VMBO exams. For HAVO this was B1 and for VWO also B1, increasing to B2.

13.4 Insights Gained

The Dutch approach to the interpretation was rather strict at the time of the first standard setting (Feskens et al. 2014), reflected in the outcomes of the second CEFR standard setting for Arabic that was organized in 2016. The outcomes of that second session show that the minimum level of A2 was still needed for the VMBO exams. The outcomes for HAVO showed that lower results were sufficient to show a B1 level. For VWO now the B2 level is needed for the exam. This “strict” approach that some teachers tend(ed) to show, and which has become more lenient nowadays, can be seen in the results of the different CEFR standard settings that have been organized throughout the years. Another factor can be that the level of the exams became more and more aligned to the level of students without or with only limited knowledge of Arabic at the start of secondary education, seen in the fact that the effects of HLI are fading out through the cohorts of students after 2004.

13.5 Conclusion: Implications for Test Users

For all different aspects when developing central exams of Arabic, procedures that have been developed and introduced to tackle all challenges step by step can be extended to other language assessment contexts: a training at the start and after that twice weekly sessions with the item developers are held, checklists for text choice and item construction have been introduced with specific attention for all issues that have proven to cause difficulty when operationalizing test specifications for Arabic. Test developers needed a new "meta language" for Arabic, to align the exams for this language to a communicative CEFR-related approach. CEFR standard settings are needed to go from a claim to a proof concerning the CEFR level of an exam, and meetings for other stakeholders, such as teachers can contribute to the implementation of Arabic exams. All of these elements can help to improve the quality of these examinations for this relatively new language in Dutch secondary education.

References

Council of Europe. (2001). *Common European framework of references for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

de Graaf, A. (2011). *Les examens de fin d'études en langue arabe dans l'enseignement de secondaire Néerlandais: Langues en immigration.* Casablanca: Eddif Maroc.

de Graaf, A., Richters, J., & de Ruiter, J. J. (2011). The Netherlands: Arabic in education. In F. Grande, J. J. De Ruiter, & M. Spotti (Eds.), *Mother tongue and intercultural valorization: Europe and its migrant children* (pp. 49–60). Milano: Franco Angeli.

Downling, S.M., & Haladyna, T.M. (2006). *Twelve steps for effective test development.* Mahwah, NJ: Erlbaum.

Feskens, R., Keuning, J., van Til, A., & Verheijen, R. (2014). *Performance standards for the CEFR in Dutch secondary education. An international standard setting study.* Arnhem: Cito.

Nielsen, H. L. (2009). Second language teaching. In K. Versteegh, et al. (Eds.), *Encyclopaedia of Arabic language and linguistics* (pp. 146–156). Leiden-Boston: Brill.

Rodriguez, M.C. (1997, April). The art & science of item-writing: A meta-analysis of multiple-choice item format effects. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. Revised August 1997.

Chapter 14
The Conflict and Consequences of Two Assessment Measures in Israel: Global PISA vs. the National *MEITZAV*

Ofra Inbar-Lourie and Elana Shohamy

Abstract Externally standardized tests are common in most educational systems worldwide. The results of these tests are used to provide feedback for improvement, monitor teaching and learning to facilitate accountability, and to provide governments with information about the ranking of countries internationally. Research, however, has shown that the benefits of massive external testing policies are diminished in comparison with the damage caused by external tests, as is evident in the form of negative washback. In Israel, major external tests dominate the educational assessment scene. Two of these tests are discussed in this chapter: a local national test, the Indices of School Efficiency and Growth, (the *MEITZAV*), and the international Program for International Student Assessment (PISA). We examine the differential reactions toward these tests, showing that the national test is rejected to the point of considered abolishment, while administration of the international test continues uninterruptedly with limited controversy. Possible reasons for this gap are discussed, concluding with a call for a critical assessment literate perspective on the consequences of large-scale national and international test use.

14.1 Introduction: Purpose and Testing Context

14.1.1 Purpose

Externally administered standardized tests are used to monitor educational systems by providing feedback for improvement, as well as for facilitating accountability. Research over time, however, has shown that the benefits of massive external testing policies are shadowed by the damage tests cause in the form of negative washback in the schools (Shepard 2000; Shohamy 2001; Tsagari and Cheng 2017). This is

O. Inbar-Lourie (✉) · E. Shohamy
Tel-Aviv University, Tel Aviv, Israel
e-mail: ofrain@tauex.tau.ac.il

E. Shohamy
e-mail: elana@tauex.tau.ac.il

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_14

evident in the restriction of learning goals as teachers teach for the test, and in judging students solely on the basis of test performances while ignoring individual differences such as linguistic needs (Menken 2008). Shepard (2016), one of the critical voices against standardized tests, states, "Perhaps we are coming to a place finally where the negative consequences can be seen to outweigh any hoped-for good." She urges therefore that policymakers should be warned of "the corrupting effects of high-stakes accountability" (p. 119).

The purpose of this chapter is to trace the consequences of such active advocacy as it occurs in Israel regarding two external tests, one national and the other international. The negative effects of the administration of these tests have indeed outweighed the benefits, leading to detrimental consequences. However, fierce protests against these adverse consequences have focused only on the national test and not on the international one, leading to a reconsideration of the local test's format and use in the Israeli educational system. This chapter describes the impact of the two tests within the Israeli ecological assessment system, the reasons brought forth by the opponents of the tests and the emerging outcomes. The chapter concludes with implications and recommendations regarding policies for external policy testing.

14.1.2 The Testing Context

In Israel, testing and assessment serve as prime tools and catalysts for educational policy reforms and monitoring, with ongoing changes and new testing policy specifically geared to affect education. Ample research has shown that such actions result in what can be referred to as testing tyranny, where key stakeholders such as principals and teachers are trained to comply with the test demands put forth by the centralized educational system (Beller 2009; Inbar-Lourie and Levi 2020; Shohamy 2001).

In terms of languages, Israel is an immigrant multilingual country. Hebrew is the national dominant language and the language of instruction in all Jewish schools, while Arabic is used as the language of instruction in the Arab communities which make up about 21% of the total population (Israel Central Bureau of Statistics, ICBS 2019). Arabic is taught as the Second Language to Jews, and Hebrew as the Second Language to speakers of Arabic, though the teaching lacks symmetry with Hebrew teaching being more dominant (Or and Shohamy 2017). English is studied as the first foreign language in all schools from early grades on. A number of immigrant and heritage languages are also taught in some schools; the most common one is Russian, used by the immigrants coming from the former Soviet Union who since the 1990s comprise about 20% of the total population (ICBS 2019). Amharic and Tigrinya, French, Spanish, and Tagalog are also used by immigrants, foreign workers, asylum seekers, and refugees.

A number of external tests dominate the Israeli educational assessment ecology. These include the following:

1. The *Matriculation examinations* for high school graduation (along with parallel internal school exams) administered in grades 10–12.
2. *The MEITZAV* (School Efficiency and Growth) *test battery:* A test of scholastic and school indices introduced by the Israeli Ministry of Education in 2002, administered by the National Authority for Measurement and Evaluation in Education (RAMA). The *MEITZAV* comprises a battery of school achievement tests and surveys, intended to provide feedback to policymakers on the teaching and learning of core school subjects and on school climate. The achievement tests are administered annually in elementary and junior high schools to assess first languages (Hebrew or Arabic), math, science, and English as the first foreign language. The tests are administered in different content areas periodically, including the possibility of internal school-based administration when the schools are not part of the external scheme (Beller 2013).
3. *International tests*: Israel has been participating in international tests such as the test of the International Association for the Evaluation of Educational Achievement (IEA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS), since 1968. Yet, the Program for International Student Assessment (PISA), under the sponsorship of the Organization for Economic Co-operation and Development (OECD), has received the most extensive focus among government officials in the Israeli society.

In addition, multiple types of tests are administered, such as placement tests for high and low achievement groups in school, school tests, municipal test batteries, and various types of diagnostic language tests for Arabs and Jews.

14.2 Testing Problem Encountered

Testing in Israel is a central issue as Ministers of Education use tests as prime tools for educational reforms, as well as for bureaucratic control and for instigated actions. Testing issues are part of intensive debates, discussed in the media, on op-ed pages, and in public discourse. A case in point is the event that took place in 2005 whereby the Minister of Education Limor Livnat took action to create an educational reform. The main outcome of the testing and measurement committee appointed as part of the reform was the creation of the *National Authority for Measurement & Evaluation in Education.* This authority is an independent statutory unit that develops and conducts assessment and provides professional guidance to the education system on matters of measurement and evaluation (Beller 2013).

In this chapter, we focus on two of the major external testing measures administered in Israel by the testing authority: The *MEITZAV*, intended to measure the efficiency of the educational system for national purposes, and the PISA, intended to rank the achievements and skills of Israeli students on the international scale. The impact of these tests on the Israeli school system is far-reaching, creating controversies and problems, hence echoing Shepard's (2016) critical stance as cited above.

14.2.1 The MEITZAV

A review of recent articles published in the local media conjures up a number of salient points as they emerge from interviews and commentaries by educational experts and former Ministry of Education officials. In a newspaper interview, Shulamit Amichai, who served as the Director General of the Ministry of Education in 1999–2001 and in 2007–2009, and who introduced the *MEITZAV* as a large-scale test to the Israeli educational system, provided a historical account of the test (Cheruti-Sover 2016). The test was initially applied in one region of the country in science and math in response to requests made by school principals who wanted to get feedback on the quality of new programs. It was later implemented nationally in spite of warnings about its potentially harmful effects. In retrospect, Amichai regrets the decision: "It [the national *MEITZAV* test] changed from being a tool to constituting a problem, since the schools adapted themselves to the test and not vice versa. They taught *for* the test and sent home weak students whom they thought would obtain lower scores on the test. The measurement tool has become the end rather than the mean" (as cited in Cheruti-Sover 2016, p. 3). Amichai also highlighted the political angle and motives, as the test was exploited by Ministers of Education in office at the time for achieving their own agendas. Such evidence of test abuse stands in dire contrast to current understandings of learning and assessment culture, and demonstrates the power and control of tests over learning (Baird et al. 2017; Shepard 2000; Shohamy 2001).

Additional harmful impacts of the *MEITZAV* were reported in the media by principals, educational and measurement experts, teachers, and parents. Specifically, they referred to the narrowing of the curriculum, and to the fact that the teaching-learning interactions are confined to test preparation to ensure success. Principals and teachers fear low achievements by their students and are therefore reported to take drastic measures, such as devoting extra time, creating additional study groups, and excluding weak students, for the ultimate goal of high test scores. Low-level school grades lead to penalties, blaming and shaming of schools and teachers (Scope 2015). In an attempt to improve the score of a particular school in the central part of the country, the parents were informed that additional preparation classes would be conducted. The reason provided was that this initiative was intended to "ease the students' tensions" before the test, to which one of the parents reacted angrily in a Facebook post, blaming the system for these tensions as part of the "school's struggle to be ranked first" (as cited in Sapir-Weitz 2016).

The situation has accelerated lately as large-scale cheating incidents were discovered, resulting in a decision by the Ministry not to publish the results of the recent test cycle.

Parental protests have also increased with a call to boycott the test (Livnat 2019; Odem 2019). Moreover, criticism is targeted at the high annual expenditure of the test which totals 18.5 million shekel (about 5 million USD), charged mostly to outsourcing companies. This sum does not include designated additional teaching hours for preparing the students for the tests, and the ample number of class periods canceled for test administration (Detel 2019).

All of these issues had surfaced already in 2014 in the protocols of the Ministry-appointed committee which examined the shortcomings and benefits of the *MEITZAV*; the test was withheld for a year while the committee convened, and resumed when a new Minister, Naftali Bennet, took office (Detel 2019). When asked in the interview mentioned above why, despite all of these adverse findings, the test was reinstated, Amichai the former General Director of the Ministry of Education gave two reasons: (1) the political uses of the test; (2) the fact that once it has become so deeply ingrained in the public eye, abolishment could be perceived as an attempt to avoid or cover up the real situation (Cheruti-Sover 2016).

As to the language dimension of the *MEITZAV*, both Hebrew and Arabic are being tested as L1s, and as the languages of assessment in content areas of math and science. In addition, both the Hebrew- and Arabic-speaking populations are tested in their proficiency in English as a foreign language. Yet, none of the *MEITZAV* tests uses the languages of immigrants in any of its versions, nor do they provide any language accommodations for those students who are still struggling with the acquisition of Hebrew. This is despite the research about immigrant students in Israel that showed clearly that acquiring academic language in schools is a lengthy process of 9–11 years (Levin et al. 2008). Furthermore, the only option provided on the L1 test of Arabic, as well as the test in the content areas, is MSA (Modern Standard Arabic), although there are convincing reports of evidence showing that many Arab students encounter difficulties in its acquisition (Saiegh-Haddad and Spolsky 2014).

Thus, despite official positive pedagogical intentions of the use of feedback in the educational system, the *MEITZAV* as a national test is laden with significant problems. The next section will discuss the influence of another external test on the Israeli school system: the PISA.

14.2.2 The PISA

Israel has participated in the PISA tests full scale since 2002, and in the last administration in 2018 students aged 15–16 years old were sampled from 200 schools to take part in the test (http://rama.education.gov.il). The PISA test is based on different goals and contexts than those of the *MEITZAV* (http://www.oecd.org/pisa/test/). Since one of the PISA's major features is international comparability, the consequences for each of the participating nations are significant and even high stakes. In Israel, success or failure on the PISA is strongly associated with and indicative of the positioning of the country on the global/international scale. Economic relations with Europe are central for Israel's trade, especially participation in the OECD and the European Union as Israel's acceptance as a member in May 2010, was perceived by leading politicians as a major achievement. A case in point is the prioritizing of the preparation for the PISA test by Gideon Saar, the Minister of Education from 2009 to 2013, who viewed performance on the test as indicative of his positive impact on elevating academic achievements in the educational system. Yet, in spite of the different goals, the PISA has brought about negative consequences on education similar to those of the *MEITZAV*.

Specifically, in a study conducted in 2010 (Shohamy et al. 2010), evidence regarding the impact of the PISA on the educational system collected from various sources—official documents, media sources, interviews with policy experts, teachers, and principals—demonstrated extensive preparation for the test. This included allocation of extra test practice hours, development of special materials, ample training for intensive teaching for the test, as well as test preparation supervision via the recruitment of 600 expert teachers to accompany 6,000 colleagues in their endeavors to improve achievements, with a total investment of about 90 million USD (Kashti 2009). In addition, major classroom tests in each of the subject areas tested were found to replicate the PISA format, and teacher education programs included PISA item formats as part of their syllabi. Despite budget cuts in other educational areas, 30,000 teaching hours were added in the core PISA subject areas: math, the sciences, and mother tongue (Hebrew and Arabic).

In terms of language on the PISA, its test regulations require that the administration be conducted in the students' L1. For the Arab students in Israel, this means that the test creates difficulties since, as mentioned earlier, many of the Arab students are struggling with Modern Standard Arabic. Being aware of this situation and fearing that this will lower the Israeli national score, the Ministry in an uncharacteristic move, invested substantially higher resources in Arabic teaching to enhance the scores.

Further to these findings, major criticism has emerged internationally about the PISA in the past few years regarding the lack of validity of international comparisons and ranking of educational systems in different contexts based on the PISA. This is currently perceived as a thorny issue (Hopfenbeck 2018). Some of the criteria that may hamper equating learners' performance across nations are embedded in linguistic concerns, even when assessing other subject areas such as math. In a discussion on comparing performance on a math exam of pupils in Russia, England, and Scotland, Ivanova et al. (2018) note that differences on supposedly identical test versions in different languages may arise from item format and presentation, but also from the meaning conveyed due to cultural irrelevance, translation (especially of keywords), and syntactical complexity. Additionally, in a research project about the 2006 PISA, El Masri et al. (2016) point to the inherent indivisible role of language in the subject area construct, in this case science literacy, as follows:

> With language as an intrinsic part of the science construct, we contend that comparing similar versions of the same science test is strictly methodologically indefensible, as translation effects are unavoidable with bias at some level being inevitable. (p. 428)

However, despite the evidence on the problematic issues of the PISA, the administration of the test continues in 80 countries worldwide, Israel included. In Israel, protest and criticism of the PISA in public discourse and media are rare, unlike the case of the *MEITZAV*, as was shown above. The question we pose therefore is the following: Seeing that these two major tests are afflicted with negative misuses and abuses, why is it that the *MEITZAV* is openly criticized to the extent of proposed abolishment, while the PISA, with similar consequences, is immune from any criticism?

14.3 Resolutions of the Problem

Different resolutions to this question were provided by different stakeholders in a research study conducted on the impact of PISA in Israel (Shohamy et al. 2010). In interviews we conducted with school principals and high-level policymakers with regard to the significance of the two tests, principals by and large felt that for them the *MEITZAV* is more significant than the PISA, as this national test provides them with useful educational feedback about their schools' performances compared to the PISA. For example, one principal remarked:

> The PISA results are less important than the national test as they do not provide results per individuals (class, student, teacher), versus the *MEITZAV* which provides me with detailed data that I can work with. (Shohamy et al. 2010)

On the other hand, a high-ranking policymaker who is critical of the over-use of tests in the system, recognized the centrality of international tests as influencing all levels of education:

> In terms of the people at the top of the pyramid, there is no doubt as to what's more important. Definitely the international test that ranks Israeli students' achievements in comparison with students in other countries. The centrality of the PISA is a message that filters down from the top levels of administration to the inspectors, school principals and ultimately to the teachers and hence, the national exam is perceived as being less important. (Shohamy et al. 2010)

Thus, the status attributed to the PISA as a global marker of excellence among policymakers and high-ranking government officials creates a halo which underscores its detrimental consequences, unlike the case of the national test. The differential attitudes toward the two tests are elaborated on next.

14.4 Insights Gained

As was shown above, while the two external tests have different purposes, one pedagogical and the other political-economic, the impact of both tests on the educational system emerges as harmful. The *MEITZAV* reflects more accurately the learning goals of the Israeli curricula. Specifically, it is designed and constructed by local teams comprising of content and testing experts and teachers, and hence results in a content and construct valid assessment that caters to the Israeli educational system. The PISA test, on the other hand, is aimed at achieving different goals of ranking Israel on the international scale, data that can be used for macro-economic policy and planning. Nevertheless, both are high stakes measures and as such have major consequences to education that can bring about reforms and changes. However, overlooking their negative effects is detrimental.

14.5 Conclusion: Implications for Test Users

This paper builds on the work of critical language testing (Shohamy 2001) and language assessment literacy (Inbar-Lourie 2017), regarding the consequences and misuses of tests in the educational system. Tests are powerful tools that governments and Ministries of Education adopt, often with limited literacy in language assessment especially regarding the harsh consequences of the two tests. It is very rare that the introduction of tests, local or global, is accompanied by discussions with the public about the consequences and impact of the tests on students, teachers, and the educational system. The fact that there are ample objections to the *MEITZAV* versus the PISA may originate from the familiarity of parents, students, principals, and teachers with the test consequences on the local level, as they can witness from close proximity, often via anecdotes and media reports, the adverse effects of the test. With regard to the PISA, the data is mostly reported in global rankings with limited diagnostic information, which makes it harder to act upon. The information from the *MEITZAV* is "close to home" and hence capable of creating a movement that critiques the test in the public space via various types of advocacy. This is especially relevant considering the central role of tests in the Israeli society.

Yet, the power of politicians, those who committed the international test for the national ideology of getting a high score in the neoliberal economics competition, is stronger than that of the educators. For Israel, eager to become part of the international community, this factor takes over and goes beyond pedagogical considerations. Yet, ample research has shown that such goals are quite problematic, as international comparisons among nations on the premise of "neutral" knowledge that imposes universal knowledge, culture, and language, are rather naive.

Indeed, Hopfenbeck (2018) expresses the urgent need to investigate the consequences of the PISA as an international test (p. 137):

> Now, more than even before, there is a need to critically investigate the validity and reliability of comparisons using international test scores. Additionally, there is a need for monitoring and examining the tests used, how the results are interpreted, and to what extent it is possible to compare across different contexts, as these tests have implication for different education systems, educational policies and sometimes even the individual students.

References

Baird, J. A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles Policy & Practice, 24*(3), 317–350.

Beller, M. (2009*).* Israel through the global education prism: Policy implications. Paper presented at the Van Leer Education Conference "From Vision and Policy to Implementation," May 20, 2009. http://rama.education.gov.il. Accessed 20 June 2019.

Beller, M. (2013). Assessment and evaluation of the Israeli education system. RAMA. http://rama.education.gov.il. Accessed 20 June 2019.

Cheruti-Sover, T. (2016, 8 September). The MEITZAV initiative: "I was warned that it's a dangerous test but I didn't listen." *The Marker*. https://www.themarker.com/magazine/1.3057774. Accessed 12 May 2019 (In Hebrew).

Detel, L. (2019, 26 January). Failed the test: The MEITZAV will be cancelled. *The Marker.* https://www.themarker.com/news/education/.premium-1.7301498. Accessed 12 May 2019. (In Hebrew).

El Masri, Y. H., Baird, J., & Graesser, A. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice, 23*(4), 427–455. https://doi.org/10.1080/0969594X.2016.1218323.

Hopfenbeck, T. N. (2018). The global and the local in educational assessment. *Assessment in Education: Principles, Policy & Practice, 25*(2), 137–140. https://doi.org/1080/0969594X.2018.1442992.

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I., & Or, S. May (Eds.) Language testing and assessment. *Encyclopedia of Language and Education (3rd ed.)* (pp. 257–270). Cham, Switzerland: Springer.

Inbar-Lourie, O., & Levi, T. (2020). Assessment literacy as praxis: Mediating teacher knowledge of assessment-for-learning practices. In M. E. Poehner & O. Inbar-Lourie (Eds.), *Praxis and L2 classroom assessment*. Cham, Switzerland: Springer.

Israel Central Bureau of Statistics, ICBS. (2019*). Israel in Figures Selected Data from the Statistical Abstract of Israel 2018.* https://www.cbs.gov.il/he/publications/DocLib/isr_in_n/isr_in_n18e.pdf

Ivanova, A., Kardanova, E., Merrell, C., Tymms, P., & Hawker, D. (2018). Checking the possibility of equating a mathematics assessment between Russia, Scotland and England for children starting school. *Assessment in Education: Principles, Policy & Practice, 25*(2), 141–159. https://doi.org/10.1080/0969594X.2016.1231110.

Kashti, O. (2009, June 17). Teaching hours in the sciences, math and Hebrew will increase by one third in junior high schools. *Haaretz on-line.* https://www.haaretz.co.il/news/education/1.1266403. Accessed 20 June 2019. (In Hebrew)

Levin, T., Shohamy, E., & Inbar, O. (2008). *Achievements in academic Hebrew among immigrant students in Israel* (pp. 37–66). XI: Studies in Jewish Education.

Livnat, O. (2019, 12 May). The National Parents Committee's call to students' parents: "Don't send them to the MEITZAV tests." *Maariv on-line.* https://www.maariv.co.il/news/Education/Article-698348. Accessed 20 June 2019. (In Hebrew).

Menken, K. (2008). *English learners left behind: Standardized testing as language policy.* Clevedon, UK: Multilingual Matters.

Odem, Y. (2019, January 11). Parents against the MEITZAV. *Channel 12 Israeli news.* https://www.mako.co.il/news-israel/education-q1_2019/Article-5bbcd6058cb3861004.htm. Accessed 20 June 2019. (In Hebrew).

Or, G. I., & Shohamy, E. (2017). English education policy in Israel. In G. R. Kirkpatrick (Ed.), *English language education policy in the Middle East and North Africa* (pp. 63–75). Mishref, Kuwait: Springer.

Saiegh-Haddad, E., & Spolsky, B. (2014). Acquiring literacy in a diglossic context: Problems and prospects. In E. Saiegh-Haddad & M. Joshi (Eds.), *Handbook of Arabic literacy: Insights and perspectives* (pp. 225–240). Dordrecht: Springer.

Sapir-Weitz, K. (2016, March 2). Is the MEITZAV the right tool for checking our children's achievements? *Maarive daily newspaper*. https://www.maariv.co.il/news/israel/Article-528663. Accessed 20 June 2019. (In Hebrew).

Scope, Y. (2015, May 6). School principals satisfied from MEITAV results' publication. *Haaretz daily newspaper*. https://www.haaretz.co.il/news/education/.premium-1.2629999. Accessed 28 June. 2019. (In Hebrew).

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Shepard, L. A. (2016). Testing and assessment for the good of education: Contributions of AERA presidents, 1915-2015. *Educational Researcher, 45*(2), 112–121.

Shohamy, E. (2001). *The power of tests.* Harlow, England: Pearson Education Ltd.

Shohamy, E., Inbar-Lourie, O., Solomon, H. (2010). *The impact of the PISA tests in Israel.* Paper presented at the 18th Sociolinguistics Conference, Sept. 1st-4th, 2010, Southampton, UK.

Tsagari, D., & Cheng, L. (2017) Washback, impact, and consequences revisited. In E. Shohamy, I. Or, S. May (Eds.) Language testing and assessment. *Encyclopedia of Language and Education (3rd ed.)* (pp. 359–372). Cham, Switzerland: Springer.

Chapter 15
How to Challenge Prejudice in Assessing the Productive Skills of Speakers of Closely Related Languages (the Case of Slovenia)

Ina Ferbežar and Petra Likar Stanovnik

Abstract The study explores potential sources of bias in rating writing performance of test takers who are speakers of languages closely related to the Slovenian language, one of the South Slavic languages. One would expect that—due to positive language transfer—speakers of other South Slavic languages would be able to pass the exam set at the A2 level of the CEFR. However, more than 40% fail, mainly because of a low score in writing and/or speaking. There are two possible reasons for this: poor test taker performance in productive skills, and incorrect rating. This chapter focuses on the latter. Research shows that some raters have higher expectations regarding the language performance of former fellow citizens than they do from others, and consequently, they rate test takers' performance unfavorably. Many raters were found to argue that speakers of languages other than South Slavic ones should be rewarded also for the fact they have to make far greater efforts to learn Slovenian. They therefore suggest different rating criteria for these "others." Taking that view into consideration, this chapter addresses the questions of "foreignness" and prejudice in a broader context of fairness in (language) testing.

15.1 Introduction: Purpose and Testing Context

15.1.1 Purpose

Fair rater-candidate interactions are a key concern in language testing practice and are therefore an important area of research. In language testing, it is crucial to assure fair rating procedures for all test takers. Fairness is a very complex concept and can be defined in many and various ways. Kane (2010) suggests two general conceptions,

I. Ferbežar (✉) · P. Likar Stanovnik
Centre for Slovene as a Second and Foreign Language, University of Ljubljana, Ljubljana, Slovenia
e-mail: ina.ferbezar@ff.uni-lj.si

P. Likar Stanovnik
e-mail: petra.likarstanovnik@ff.uni-lj.si

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_15

procedural and substantive fairness, both closely connected with validity. In this chapter, we will focus on procedural fairness, which “can be viewed as a lack of bias for or against any individual or group” (p. 178). Such fairness “is concerned with how we treat test takers, in particular with how consistently and fairly we treat them, and is therefore largely under our control” (p. 179). In language assessment that involves human raters, fairness must be of central interest since “human ratings are often associated with more or less severe consequences for those being rated” (Eckes 2017, p. 443).

Rater bias is one of the rater effects (together with rater severity/leniency, central tendency, halo effect) (Eckes 2017, pp. 444, 445), and it is manifested in reliability—or rather unreliability—of rating. Reliable and consistent (human) rating is one of the most challenging issues in language testing, and it has important implications for test quality.

The purpose of our study was to explore potential sources of bias in rating writing performance of test takers who are speakers of languages closely related to the Slovenian language (which is one of the South Slavic languages). We assumed that rater bias might stem from the very personal relationship or specific attitude Slovenian first language (L1) speakers have toward their own language, and such bias might also have its origin in stereotypes and prejudice. To make this hypothesis more clear, it is necessary to familiarize readers with the broader historical and linguistic context.

15.1.2 Historical and Linguistic Context

With a population of some two million, Slovenia is among the smallest members of the European Union. Before independence, Slovenians lived in multinational and multilingual states. Most of the time, therefore, the Slovenian territory was marked by a (more or less hidden) bilingual linguistic situation (with German having higher status in the Austrian Empire, and Serbo-Croatian in the Socialist Federal Republic of Yugoslavia). In the so-called Yugoslav times

> the issue of the status of the Slovenian language, on the one hand, seems to have been formally resolved, while, on the other, remaining largely open… and politically sensitive, in certain periods even politically or ideologically unacceptable. (Stabej 2010, p. 45)

In comparison with what was then called Serbo-Croatian, Slovenian was “a language with a weaker status and prestige” (Stabej 2010, p. 171; see also Požgaj Hadži et al. 2009).

Throughout history, Slovenian has had a strong identification role for Slovenians. Moreover, it has always been “linked also with political desires, and after 1848 it was often at the forefront of political events as an argument for and goal of political action” (Stabej 2010, pp. 44–45). Space does not allow for detailed delving into the complex political and social processes in which the attitude of Slovenians toward their own language was formed; however, the following comment by Stabej should serve as an introduction:

> The rich Slovenian experience with ideological and linguistic normative interventions…, which arise from the centrality of the identification role of the Slovenian language in the history of the Slovenian national movement, is still mirrored in the relation of Slovenian speakers to their language and to thinking about it. (p. 55)

To put it simply, rather than a functional view of language, a distinctly personal, sometimes even emotional, one developed. This view especially holds true for those who are professionally involved with the language, for example, copyeditors or proofreaders (Stabej 2010), and for teachers of Slovenian. To some extent, Slovenians have an ambivalent attitude toward the Slovenian language: While they feel their less widely spoken language is therefore less valuable, they also often believe that "proper" Slovenian cannot be learned by many native speakers, let alone by foreigners (pp. 62, 168, 198).

A consequence of such thinking is also the ambivalent attitude of Slovenians toward "others." On the one hand, there is a feeling of inferiority toward Western Europe, and, on the other, one of superiority toward the "Balkans" and the "East," to which they do not wish to belong (Šabec 2007, pp. 113–115; Šabec 2006, p. 127; Debeljak 2004, pp. 96, 129).

15.1.3 Testing Context

Slovenian independence in 1991 saw Slovenian constitutionally determined as the official language. This step, it seems, entailed that all the historical goals of Slovenian (linguistic) political activities had been fulfilled (Stabej 2010, p. 46). But in the light of the migration processes and European integration, new issues of public communication and thus the status of the Slovenian language have come to the fore. A number of legislative acts have been introduced which regulate the status of Slovenian (see Ferbežar 2012; Stabej 2010; Ferbežar and Stabej 2002), the Citizenship of the Republic of Slovenia Act (1991) being one of the first laws to require knowledge of Slovenian. Consequently, the system of testing and certifying knowledge of Slovenian was introduced.

These historical circumstances obviously influenced "the degree of conviction about the power of Slovenia to unite and to give its speakers a sense of (national) unity" (Ferbežar 2012, p. 32). In this, however, Slovenia was not alone.[1]

Knowledge and use of Slovenian also functions as a sign of belonging, even allegiance, to the Slovenian nation and to "Slovenianness" (Ferbežar 2012, p. 32). Specifically, only "good" knowledge, especially in public communication, counts (Stabej 2010, pp. 204–205). Expectations for foreign speakers of Slovenian are (or were) therefore very high. Nevertheless, in the naturalization processes, the Slovenian

[1] In the mid-1990s, language exams were introduced in Estonia and Latvia for the purposes of naturalization. The reasons for introducing these exams "were clearly tied to the view that the Baltic Republics were illegitimately occupied … during the Soviet period" (Hogan-Brun 2009, p. 40).

state puts forward attainable language requirements: basic knowledge of Slovenian, i.e., knowledge that suffices for everyday communication, without knowledge of society (unlike, for example, the Baltic and some other European states) (see Extra et al. 2009).

In 1994, the Centre for Slovene as a Second and Foreign Language (hereinafter the Centre) at the University of Ljubljana was appointed by the government for testing and certifying knowledge of Slovenian. Since then, various testing systems have been in place. In 2014, the exams were aligned to the Common European Framework for Languages (2001; hereinafter the CEFR) and the education program Slovenian as a Second and Foreign Language (2014) was published; this program defines and determines the exams at Basic, Intermediate and Advanced levels. The program provides a detailed description of levels, specifies standards of knowledge and defines the structure of the exams.

Exams on all three levels are administered by the Centre. Additionally, the Basic Level Exam is administered by external testing centers (their number varies; currently there are 16).[2] The Centre issues the certificates, collects and keeps all data for statistical analysis[3] and research purposes.

It is important to note that the vast majority of test takers sit the Basic Level Exam. Most sit the exam for naturalization purposes, with most coming from the area of what was once a common state.[4] This has important implications: Most test takers are speakers of languages closely related to Slovenian.[5]

The balance of this paper focuses on the Basic Level Exam.

15.1.3.1 Basic Level Exam

According to the program Slovenian as a Second and Foreign Language (2014, revised 2020), the Basic Level Exam is set at A2 and B1 level of the CEFR. It is a paper-based exam consisting of four subtests covering reading, listening, writing and speaking. Reading and listening are composed mainly of text-based selected response tasks and are marked according to a marking scheme. The writing and speaking subtests are performance tests (guided writing and speaking) and are double-rated

[2] See "the Centre's web page" (n.d.).

[3] Analysis of candidature and exam results for each year is available in "The Centre's annual reports" (n.d.).

[4] Slovenia has always been economically attractive for immigrants, and after independence the influx to Slovenia increased. Since Slovenia joined the EU in 2004, the influx of immigrants has been growing (Eurostat n.d.); although more immigrants are now coming from other countries, citizens from former Yugoslav republics still represent the majority (Republic of Slovenia, Statistical Office n.d.).

[5] In 2018, 91% of test takers sat the Basic Level Exam, 75% of them were applicants for Slovenian citizenship, 86% of them were speakers of one of the languages closely related to the Slovenian language, i.e., Bosnian, Croatian, Serbo-Croatian, Serbian, Montenegrin, Macedonian (see "The Centre's annual reports" n.d.).

All data are systematically collected in the sign-up for the exams, but not all are obligatory; the proportion might therefore be likely higher than that cited.

according to rating scales. The cut score for each subtest is 60% for A2 and 85% for B1, enabling flexible profiling of test takers' skills. For the purpose of this study, it is important to present the writing subtest in more detail.

The writing subtest consists of two guided tasks: Test takers (1) write a practical text on a hypothetical situation, and (2) reply to a short private letter. For each task, the expected response is a simple coherent text, 40–50 words in length.

Each text is rated according to an analytical scale, the categories rated are text content (0–3 points), vocabulary (0–3 points), accuracy (0–3 points), and coherence and style (0–1 point). Each score reflects a certain level of performance (e.g., 3 points = B1, 2 points = A2 and so on). A maximum of 10 points can be awarded for each text, with a maximum of 20 points for the whole subtest. The cut score has been set at 12 points for the A2 level, and at 17 points for the B1 level.[6]

The first step of the rating procedure is rater standardization (see further in this section). When rating, raters record their scores on the rating sheet and have to comment on their decisions. Writing performance is double-rated. If administered by external testing centers, the first rating (hereinafter 1st rating) is provided by external raters, and the second rating (hereinafter 2nd rating) by the Centre's raters. The final score for writing the subtest is an average of both ratings. The score is then converted into the corresponding CEFR-level. In the event of major discrepancies,[7] a third rater is brought in. The third rating (hereinafter 3rd rating) is also provided by the Centre's raters. In the case of a 3rd rating, the two ratings which follow the rating scale most closely are taken into consideration.

The final grade for the entire exam is assigned centrally.

15.1.3.2 Raters

Raters are selected according to the Minister of Education Regulation on the Educational Program Slovenian as a Second and Foreign language (2014). The raters are all teachers of Slovenian, preferably with some experience in Slovenian as a second language (L2). Additionally, they have to be trained for administering the Basic Level Exam, and for rating writing and speaking performance. Their work is regularly monitored by the Centre.

In 2018, the Centre kept records on 93 raters: 19 of them were the Centre's raters, and 74 were external ones. It is important to note that external raters are appointed by external testing centers. Most of them work as teachers of Slovenian as L1, while some of them also have experience in teaching Slovenian as L2. The Centre's raters, on the other hand, are all experienced teachers of Slovenian as L2.

[6] Cut scores for the writing subtest have been set at the benchmarking seminar in the process of aligning exams to the CEFR. At the same time, prompts for the two tasks were also linked with the CEFR scales, and they were placed between the A2 and B1 levels of the CEFR (see Ferbežar et al. 2014).

[7] Major discrepancy means a difference in the number of points awarded (if this is greater than 3/20), or the line between pass/fail (below A2) or the line between the levels A2 and B1.

To ensure consistency/precision in rating, and consequently reliability of tests, the Centre provides the following measures:

- 16-hour introductory examiner and rater training;
- regular training for raters (standardization seminars);
- a book of regulations on administering and rating exams;
- standardization prior to rating speaking and writing performance[8];
- rater monitoring (visits to external centers, and regular monitoring of marking and rating);
- regular checking of intra- and inter-rater agreement.

Although these measures should ensure fair decisions "regardless of individual test takers' group membership" (Bachman and Palmer 2000, p. 32), we have noted some inconsistencies.

15.2 Testing Problem Encountered

One would expect that—due to positive language transfer which enables the main message to be communicated despite linguistic inaccuracy—speakers of languages closely related to Slovenian would be able to pass the Basic Level Exam. However, many of them fail, mainly because of a low score in writing and/or speaking. These results invite the following questions: (1) Is the test takers' performance in productive skills insufficient? (2) Is rating being carried out inaccurately?

If we briefly consider the first question, from collected data (see "The Centre's annual reports" n.d.), we can draw some conclusions regarding potential reasons for test takers' insufficient performance in productive skills. According to these data, about half of the Basic Level Exam test takers have acquired Slovenian unsystematically. Even though they have the opportunity to participate in free language courses (60–180 h), their motivation for systematic learning seems to be quite low. In addition to the main reason, i.e., lack of time to attend a course (which is often related to financial reasons), there are other reasons: the similarity of languages and/or a lack of awareness that linguistic proximity is no guarantee of good language skills (the interlanguage used by the test takers may suffice for comprehension and speaking, but not for writing); relatively low education (many test takers have low functional literacy in their L1), etc. (Ferbežar 2012, pp. 40–41). There may also be a lingering historical reason: In the former common Yugoslav state, fellow citizens from other republics did not have to learn Slovenian, while Serbo-Croatian was a compulsory subject in Slovenian schools.

[8] Benchmark ratings "are usually assigned to a range of typical performances that raters may encounter during operational rating sessions" (Eckes 2017, p. 447). In the case of the Basic Level Exam, benchmarks are available as a recorded sample of performance by a minimally acceptable person (Angoff 1971, as cited in Council of Europe 2009, p. 65), and samples of writing performance at the levels below A2, A2 and B1.

However, the focus of this paper is the second question, i.e., as to whether the low score on the Basic Level Exam is a consequence of inaccurate rating.

Undoubtedly, rating should reflect a test taker's ability rather than factors unrelated to that ability, such as rater bias. However, as stated by Eckes (2017, p. 444),

> It is commonly acknowledged that raters do not passively transform an observed performance into a score using a rating scale, but actively construct an evaluation of the performance…. These constructions are based, for example, on the raters' professional experience, their understanding of the assessment context, their expectations about the performance levels, and their interpretation of the rating scale categories.

As stated, in the case of Slovenian language exams, writing performance is double-rated: the 1st rating is provided by external raters, and the 2nd rating is provided centrally. In the Centre, we have identified discrepancies between external and central rating. We have observed that in some cases external raters "display particular patterns of harshness or leniency in relation to only one group of candidates, not others" (McNamara 1996, p. 123). To be concrete, external raters have higher language expectations toward speakers of languages closely related to Slovenian than toward others, and rate them more severely. This practice leads us to the conclusion that in the case of Slovenian language exams, not all test takers are treated "in essentially the same way" (Kane 2010, p. 178). If we see test fairness as linked directly to validity (Xi 2010), this would mean that the validity of our exams might be compromised.

The Slovenian Language Exams are high-stakes exams. Especially critical are therefore test reliability (consistency of measurement) and validity as essential measurement qualities (Bachman and Palmer 2000).

15.3 Review of Literature

The increasing popularity of (supposedly) authentic performance testing entails rater judgment. A considerable number of studies have explored performance test assessment, and they have addressed various aspects of rater behavior, while suggesting various methods for evaluating rating quality (a systematic overview of methods is provided by Wind and Peterson 2018; see also Eckes 2017). Of special relevance for the purpose of this study are those overviews that focus on rating writing performance. Here, we are referring to some more recent contributions.

On the basis of patterns recognized in rater behavior—i.e., how raters perceive rating criteria, grammar being perceived as the most important category—Eckes (2012) defined six rater types (for rater-category interaction see also Brown 2010, Eckes 2008, Schaefer 2008). Similarly, Baker (2012) explored the impact that individual differences in cognitive style may have on rater decisions.

Numerous studies have focused on rater behavior in relation to particular subgroups of test takers. Kondo-Brown (2002) thus revealed that a "much higher percentage of significantly biased interactions was found for the candidates with extremely high or low abilities" (p. 24). Similarly, Schaefer (2008) reported about raters' tendency "to show more severe bias towards the highest ability writers and more lenient bias towards the lowest ability writers" (p. 486). Wind's survey (2019) highlighted "differences in the rater interpretation of test-taker achievement when evaluating female and male test-taker compositions" (p. 18). As opposed to human scoring, which, although unreliable, is still "the gold standard," Brown (2010, p. 281) suggested alternative approaches to testing and rating essays, and presented the advantages of machine essay scoring as opposed to human scoring.

The research focusing on rating speaking performance has come to similar findings. Lynch and McNamara (1998) revealed that in rating speaking skills, "certain raters are more severe or more lenient than others for certain persons" (p. 166) and warned about the risks of single rating practices (p. 170). Although rating speaking skills is beyond the scope of this chapter, it seems necessary for the context of our study to draw attention also to the findings of Winke, Gass and Myford (2013). Their study is one of the few exploring a potential *source* of bias. It reveals that familiarity with a particular accent affects rating (not necessarily in favor of the speakers of the familiar accent). Listeners might stereotype foreign accents, and "speakers of certain foreign accents may be stereotyped by some individuals as having a lower social status" (p. 232).[9]

As language testers we obviously "do not expect professional raters to have varying attitudes towards different accents or if they do at any level, these attitudes should not affect how the raters evaluate test takers' speech" (Winke et al. 2013, p. 233). Nevertheless, Huei-Lien Hsu (2016) reported on the negative attitudes listeners tend to hold toward speakers of non-standard English and a tendency to judge them accordingly.

Many of the studies cited here have revealed that, despite training, rater bias is "resistant to change" (Eckes 2012, p. 274; see also McNamara 1996, p. 127, Baker 2012 and Schaefer 2008 for an overview) and confirmed scoring writing performance (i.e., essays) as being "notoriously unreliable" (Brown 2010, p. 278; see also Wind and Peterson 2018, p. 178). Of course, this does not mean omitting rater training. On the contrary, rater behavior can and must be controlled, and rater training and monitoring are essential measures for assuring such control (see also Engelhard et al. 2018).

It can be argued that the findings of the cited studies may have limited generalizability and cannot be simply transferred into the context of this paper, that is, into a Slovenian context. This is due to the fact that they relate exclusively to large-scale assessment that allows for the use of such research methods as Many-Facet Rasch Analysis and nonparametric item response theory. Generally, much of the research

[9] In the Slovenian context, Balažic Bulc (2009) provided evidence that untrained listeners stereotype native speakers of different languages, expressing more negative attitudes toward native speakers of South Slavic languages.

has been, as stated by Engelhard et al. (2018, p. 47), "dominated by a psychometric perspective."[10] Furthermore, studies focusing on writing performance relate to rating of essays,[11] mostly in a very specific academic setting, and they are not necessarily concerned with L2 performance (Wind and Peterson 2018, p. 181). They have neither addressed less widely taught languages and/or small-scale testing contexts nor addressed issues of testing speakers of closely related languages. Ferbežar and Stabej (2013) raised some questions regarding rating speakers of closely related languages. However, they simulated a rating situation focusing on text comprehensibility and acceptability that can have implications for rater decisions rather than rater bias.

It is worth noting that quantitative approaches in large-scale rater-mediated assessment (which is therefore somewhat technical in nature) allow researchers to recognize variability in rating and to highlight different aspects of rater behavior. However, they observe rater judgment mostly outside the broader social context, and the actual *reasons* for or *sources* of such variability therefore remain undetected. In any case, they all discuss issues that are of key importance for any high-stakes assessment.

15.4 Methodology

The purpose of our study was

- to recognize discrepancies between external and central rating,
- to explore potential inconsistencies in rating two different groups of test takers regarding their first languages,
- to identify reasons for such inconsistencies and to detect potential sources of bias in rating writing performance.

The findings should have implications for further rater training.

A two-part study was carried out: (1) We analyzed data from a live test, i.e., the Basic Level Exam administered in March 2018 at external testing centers. (2) Additionally, we examined raters' attitudes toward rating procedures through a questionnaire launched in May 2018. We used a mixed quantitative and qualitative approach to analyze the data.

[10] In order to evaluate the quality of rater judgments and consequently to improve rater-mediated assessment, Engelhard et al. (2018) suggest combining two theoretical perspectives: psychometric and cognitive.

[11] Writing an essay is, compared to writing a short practical text, a very specific skill not only in terms of organization and length but, above all, complexity (i.e., reasoning and logic).

15.5 Findings

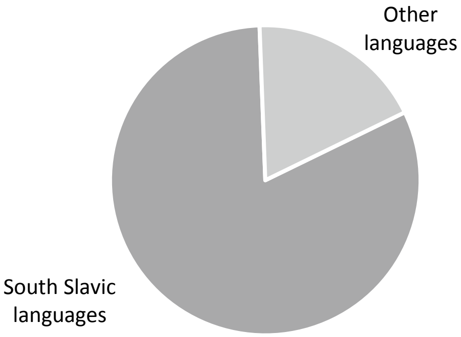
15.5.1 Rating Analysis

15.5.1.1 Test and Test Takers

In May 2018, 686 test takers sat the Basic Level Exam. Figure 15.1 shows the breakdown of test takers according to first language (L1).

Figure 15.2 shows test takers' grades on individual subtests. A total of 53% of test takers passed the exam. As can be seen, the test takers' performance in the receptive skills is high, as most of them performed at level B1.

Fig. 15.1 The test takers' L1 (N = 686)



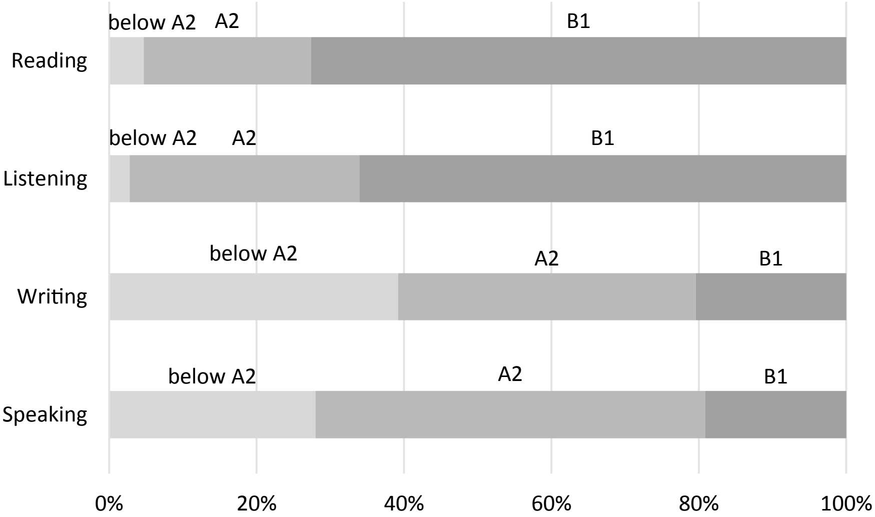


Fig. 15.2 Test takers' grades on the individual subtests (N = 686)

All test takers took the same test, with two writing tasks in the writing subtest (see Sect. 15.1.3.1). According to statistical analysis,[12] both the test and the writing subtest have high reliability (Cronbach α 0.93 for the test and 0.82 for the writing subtest).

The mean score for the writing subtest was 12.11 points (out of 20), with SEM 0.18 and difficulty index 0.59 (0.59 for task 1 and 0.60 for task 2).

15.5.1.2 Raters

In the rating procedure, 41 raters (out of 93) were employed to rate 681[13] writing performances (here we understand one writing performance as a response of one test taker to both writing tasks). There were 34 external raters that provided a 1st rating, and seven Centre raters to provide a 2nd and, in the case of discrepancies, a 3rd rating. Within the span of three weeks, external raters rated between 2 and 58 writing performances (with a mean of 20), whereas the Centre's raters rated between 34 and 280 writing performances (with a mean of 115.1), providing 806 ratings altogether. This means that major discrepancies appeared in 125 cases, and a 3rd rating was needed. We can therefore conclude that the agreement between external and central rating was relatively high (82%). In what follows, we will focus on the remaining 125 ratings to examine discrepancies.

In the rating procedure, the Centre collected the following data: raters' name, scores for each writing performance and level awarded (below A2, A2, or B1).

15.5.1.3 Analysis

The analysis was performed in the following steps:

- we excluded from further analysis ratings where the change in the score[14] for the writing subtest did not influence the final grade; there were 44 such cases;
- we analyzed the remaining 81 writing performances where a change in the score influenced the final grade for the writing subtest; we observed ratings according to rater leniency/severity, and according to the test takers' L1. Finally, we checked whether such a change only influenced the test taker's level of writing performance achieved (i.e., A2 or B1) or, more importantly, whether the test taker would pass or fail the exam.

[12] In post-examination analysis, classical test analysis and IRT analysis are applied on a routine basis for the whole test population. At item level, indices for difficulty, discrimination, standard errors of measurement are calculated, and at test level indices for difficulty, reliability, internal consistency and standard errors of measurement are calculated.

[13] The difference in the number of test takers is due to the fact that 5 candidates did not attend the writing subtest.

[14] E.g., the 1st rating of writing performance was 12 points and 2nd rating was 16 points, both indicating the level A2.

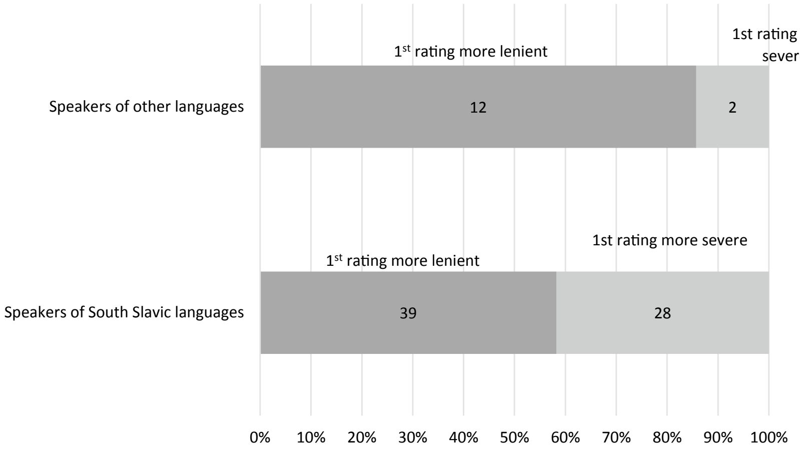


Fig. 15.3 Difference in scoring between 1st and 2nd/3rd ratings (N = 81)

Breakdown of test takers' writing performances (step 2) is summarized in Fig. 15.3.

We can conclude from Fig. 15.3 that the 1st rating generally was in favor of the test takers (1st raters award higher scores than those providing 2nd and 3rd ratings). From the test takers' point of view, these discrepancies can be seen as positive; as a high-stakes exam, the Basic Level Exam has serious implications for their personal lives. However, inconsistent rater decisions affect test validity.

This finding raises the question whether the central raters are too severe or the external raters are too lenient.

More detailed analysis revealed that central rating was consistent (98% agreement).[15] It also revealed that external raters were "overusing the middle categories of a rating scale" (Myford and Wolfe 2004, p. 198), i.e., a central tendency effect. Thus, the 1st ratings did not actually reflect the actual test takers' writing ability in Slovenian in its whole range: Test takers who used complex structures and were thus more likely to make more linguistic errors, were not appropriately rewarded for taking risks. Of concern, however, is the fact that lower grades were systematically given to speakers of South Slavic languages.

Firstly, seven test takers (out of 81), all South Slavic language speakers, would not have passed the exam. In terms of the whole test population, this number is very small (1%), but it clearly represents a rater behavior that systematically disadvantages a group of test takers, i.e., differential rater severity (Myford and Wolfe 2004, p. 213).

Secondly, in 42 cases (out of 81), a difference in leniency meant a difference between whether the test takers achieved level A2 or B1. Although this difference was not decisive for them (they passed the exam), the analysis shows that the South Slavic language speakers' productive skills were often rated lower than they should

[15] Benchmarking seminars that the Centre regularly holds to check cut-off scores and grade boundaries confirm this claim.

be, with the 1st ratings often being too severe (21 out of 36 speakers of South Slavic languages, as opposed to 2 out of 6 speakers of other languages).

We can conclude that not all test takers were treated in the same way, and that rater bias influenced rating writing performance.

15.5.2 Attitudes Toward Rating Criteria

In the second part of our study we tried to identify potential sources of rater bias by means of a questionnaire on raters' attitudes toward rating procedures.

15.5.2.1 Questionnaire

The questionnaire consisted of four dichotomous questions that required a short explanation with the purpose of verifying the relevance of the rating procedure:

- Question 1: Do benchmarks[16] help you to standardize before rating? Please explain.
- Question 2: Should speakers of Slovenian as a L1 be the reference for rating criteria? Please explain.
- Question 3: Should non-Slavic speakers be rated according to different criteria? Please explain.
- Question 4: Should the test taker's "effort" or "path travelled" be taken into account when rating? Please explain.

The questionnaire in Slovenian was accessible through an online application "One Click survey 1KA" (n.d.)[17] from 16 April to 16 June 2018, and it was anonymous.

The answers to dichotomous questions were analyzed quantitatively, while the short explanations were analyzed qualitatively.

15.5.2.2 Participants

All 93 raters from the Centre's records were invited to answer the questionnaire. A total of 87 raters replied—19 of the Centre's raters (100%), and 68 external raters (92%).

We sent separate links to the questionnaire to the Centre's raters and to external raters in order to identify potential differences in their attitudes.

[16] See note 8.

[17] OneClick survey 1KA is available in Slovenian and English.

15.5.2.3 Analysis

Question 1: 60 external raters (88%) answered that pre-exam standardization helps when rating. The remaining 8 (12%) mostly commented that it is difficult to compare benchmarks with “live” performances, since each performance is different.

Question 2: 14 external raters (23%) answered that speakers of Slovenian as a L1 should be the reference for rating criteria. In their comments, they stated that native speakers know the language best and are most suited for setting the norm; non-native speakers will never be capable of speaking Slovenian at the level of a native speaker. Some also highlighted the social and symbolic role of the Slovenian language.

Question 3: 15 external raters (25%) answered that non-Slavic speakers should be rated according to different criteria. They commented that Slavic speakers have the advantage of not having to put much effort into learning Slovenian and can be better understood on account of linguistic proximity. Although South Slavic language speakers were not referred to in the question, the raters mentioned them explicitly as being “privileged.” Even among those who responded negatively to this question, 6 (9%) commented on having a favorable inclination toward non-Slavic speakers, because it is harder for them to learn Slovenian. Therefore, the proportion of raters who actually agree with having different rating criteria for different groups of test takers is more than one-third (21; 35%). We did not ask the raters how the rating criteria should differ, but it can be concluded from the answers that they think the rating criteria should be less severe for non-Slavic speakers.

Question 4: 18 external raters (31%) answered that the test taker’s “effort” or “path travelled” should be taken into account when rating. In the comments they mentioned the far greater efforts required of non-Slavic speakers. They also pointed out that the test takers who speak closely related languages do not prepare for the exam and thus show their dismissive attitude toward the Slovenian language and culture. Similar comments were also found among those who answered negatively; two (3%) further emphasized that South Slavic language speakers have no desire to learn Slovenian even after living in Slovenia for 10 or more years.[18] Therefore, the number of raters who actually agree with the consideration of “effort” is 20 (34%).

Compared to external raters the Centre’s 19 raters seemed to be more supportive toward the existing rating procedures: 18 (95%) found pre-exam standardization helpful, two (13%) would prefer native speakers as the reference for rating criteria, two (13%) were in favor of different criteria, and none would take into account the test taker’s “effort” when rating. Unlike the external raters, the Centre’s raters commented on rating procedures rather than expressing personal views.

These results partially confirm our concerns that rater bias might have origins not only in raters’ specific attitude toward their own language, but also in stereotypes and prejudice (Winke et al. 2013). Even though rating speaking was not the subject of our

[18] These comments probably refer to one of the conditions for obtaining Slovenian citizenship: 10 years of living in Slovenia, including 5 years continuously before applying for citizenship (Citizenship of the Republic of Slovenia Act 1991).

research, here, we would like mention analysis on rating speaking[19] that provided similar results. It revealed that a considerable number of raters (50%) rated spoken production of a Macedonian speaker more severely, and consequently placed it at a lower level than the speaker actually performed at. Moreover, the vast majority of raters (82%) rated spoken production of a Dutch speaker more leniently and placed it at a much higher level. A possible explanation is that the raters were disturbed by the Macedonian's "foreign accent," but also rewarded the Dutch speaker's efforts to learn Slovenian. It is thus clear that some raters "have a tendency to over- or underrate a test taker or class of test takers" (McNamara 1996, p. 123). Given the common history with South Slavic language speakers, and also the relation of Slovenians to the Slovenian language (described in Sect. 15.1.2), we hypothesize that rater bias is a result of ethnic prejudices and possibly also historical grudges.

15.6 Insights Gained

The hypothesis that rater bias is present as a consequence of ethnic prejudices seems to be confirmed. In this regard, we refer to Allport's (1966, p. 9) seminal definition:

> Ethnic prejudice is an antipathy based on a faulty and inflexible generalization. In may be felt or expressed. It can be directed towards a group as a whole, or towards an individual because he is a member of that group.

Our research has shown that some prejudices are still explicitly expressed (i.e., old or traditional racism) (see Ule 2005). Even those raters who in the questionnaire disagreed on principle about test takers being treated differently, often relativized their answers in the comments by adding a "but" and some general or stereotypical explanations.

The analysis of writing performance ratings shows that prejudices can also be expressed in a disguised, indirect way, i.e., new racism (see Ule 2005)—namely, to "punish" errors and not to "reward" good language performance. In our study we did not explore rater-category interactions to confirm that claim. Ferbežar and Stabej (2013, p. 349) have provided some insights, stating that

> stricter testers tend to put an emphasis on accuracy rather than on other categories, whereas more lenient testers seem to consider other categories more important. But generally it seems that an inaccurate text is more 'disturbing' when produced by a speaker with a South Slavic background than speaker of language other than South Slavic.

Thus, it seems that the emotional attitude toward Slovenian continues to emerge, reinforcing the stereotype that Slovenian cannot be truly learned by foreign speakers (see Sect. 15.1.2); moreover, expectations toward former fellow citizens are higher than toward others (Ferbežar and Stabej 2013).

[19] The analysis referred to here was a part of broader analysis of interrater agreement and rater consistency (see measures for determining validity and test reliability in Sect. 15.1.3.2). It was carried out in 2016–17; all Basic Level Exam raters (N = 92) took part in it; the difference between external and Centre's raters was not observed.

When compared to predominant research with large-scale assessment, the figures we refer to here are negligible. They do not allow for the use of statistical methods for detecting and measuring rater effects (Sect. 15.3). In our context, however, the figures do allow for some generalizations, since we analyzed data pertaining to the administering of the exam as a whole and almost all raters responded to our questionnaire. The findings lead to important implications. The most important one is that the vast majority of raters adequately rate the test takers' writing performance and are aware that the use of the same rating criteria for all is crucial for ensuring fairness and test reliability. But there still seems to exist hidden bias as a result of deeply rooted prejudices—especially among teachers of Slovenian as a L1. As already pointed out, this bias, which manifests itself through treating individual groups of test takers differently when rating their production, is not coincidental.

Our research does not provide any unexpected findings; it only confirms those from other areas (e.g., social psychology). However, it does invite sensitive questions of discrimination with broad implications. It is therefore necessary for language testers to research and report on such bias more often.

15.7 Conclusion: Implications for Test Users

We are well aware of several limitations of the study presented in this chapter: Factors that may also influence rating, such as experience in L2 teaching, experience in rating, raters' L2 background, interpretation of rating criteria, i.e., rater-category interaction, were not taken into account. Nor did we analyze whether different ratings had come from individual raters or whether there really was a systemic problem.

Moreover, combining the findings of both parts of our research is methodologically questionable: In rating analysis, only 44% of raters were participating, whereas 94% of raters responded to the questionnaire. Since the questionnaire was anonymous, we do not know how individual raters responded. Therefore, these findings must be interpreted with some reservation.

The Centre's task is therefore to conduct further research in this area, including a longitudinal study that would provide enough data to enable us to use appropriate analytical tools and to discern other rating errors. Special attention will have to be paid to rating speaking performance.

In any case, notice will have to be paid to prejudices. Although the described measures to ensure fair decisions (see Sect. 15.1.3.2) should be sufficient, our research shows that it is not always the case. It is necessary to constantly raise raters' awareness of prejudices and to help them recognize possibilities for the constant development of intercultural competence, which we understand as cognitive curiosity and the ability to "empathize and experience" (Debeljak 2004, pp. 52, 102); in other words, it is a matter of sensitization regarding foreign cultures in general and of the ability to overcome stereotypes and prejudices, while respecting equality.

Although tentative in nature, this study has important implications also for test takers.

Referring to prejudices, they are

> informal institutions… that translate certain real relations of inequality, dominance and subordination between social groups into the realm of everyday life, and vice versa, [and] establish certain relationships between different groups in the everyday world in a general framework of valid social norms, values and institutions. (Ule 2005, p. 28)[20]

Therefore, it appears that completely eliminating prejudices is impossible. As stated by Baker (2012, p. 226), "there continues to be unexplained variability that resists training," with prejudice being one of the variables.

But in order not to increasingly entrench, legitimize or—in the light of growing and institutionally supported intolerance in modern Europe—even normalize these prejudices, they must be constantly addressed; it is a matter of social justice, and in the language testing context, a matter of fairness. From test takers' perspective, addressing prejudices is crucial since rater decisions, potentially influenced by such prejudices, have direct consequences for their lives.

Acknowledgements Main findings of this study were presented at the 54th ALTE Meeting and Conference Day, Ljubljana, Slovenia, November 8, 2019. The early version of this chapter was written in Slovenian. English translation of that version, including all quotations from the references in Slovenian was provided by Dr Jason Blake. We would like to thank him for his great effort and assistance with later versions. Also, we would like to express our gratitude to the editors, especially to Dr Betty Lanteigne for her valuable comments and guidance through the process of revision.

References

Allport, G. W. (1966). *The nature of prejudice*. Addison-Wesley Publishing Company.

Angoff, W. H. (1971). Scales, Norms and Equivalent Scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bachman, L. F., & Palmer, A. S. (2000). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baker, B. A. (2012). Individual differences in rater decision-making style: An explanatory mixed-method study. *Language Assessment Quarterly, 9*(3), 225–248. https://doi.org/10.1080/15434303.2011.637262.

Balažic Bulc, T. (2009). Odnos do tujih jezikov in njihova prepoznavnost v slovenski družbi. In V. Požgaj Hadži, T. Balažic Bulc, V. Gorjanc (Eds.). *Med politiko in stvarnostjo. Jezikovna situacija v novonastalih državah bivše Jugoslavije* (pp. 181–194). Ljubljana: Znanstvena založba FF UL.

Brown, G. T. L. (2010). The validity of examination essays in higher education: Issues and responses. *Higher Education Quarterly, 64*(3), 276–291. https://doi.org/10.1111/j.1468-2273.2010.00460.x.

CEFR (2001) = *The common European framework of reference for languages: Learning, teaching, assessment* (2001). Strasbourg: Council of Europe. https://rm.coe.int/1680459f97. Accessed 26 July 2018.

Citizenship of the Republic of Slovenia Act. (1991). http://www.legislationline.org/download/action/download/id/6579/file/Slovenia_Citizenship_Act_1991_am2002_eng.pdf. Accessed 26 July 2018.

[20] See also van Dijk (1987).

Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR). A manual*. Strasbourg. https://rm.coe.int/1680667a2d. Accessed 27 July 2018.

Debeljak, A. (2004). *Evropa brez Evropejcev*. Ljubljana: Založba Sophia.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. https://journals.sagepub.com/doi/pdf/10.1177/0265532207086780.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behaviour. *Language Assessment Quarterly, 9*(3), 270–292. https://www.testdaf.de/fileadmin/Redakteur/PDF/Forschung-Publikationen/Eckes_LAQ_2012.pdf.

Eckes, T. (2017). Guest editorial: Rater effects: Advances in item response modelling for human ratings—Part I. *Psychological Test and Assessment Modelling, 59*(4), 443–452. https://www.testdaf.de/fileadmin/Redakteur/PDF/Forschung-Publikationen/03_Eckes.pdf.

Engelhard, Jr, G., Wang, J., & Wind, S.A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessment using accuracy ratings. *Psychological Test and Assessment Modelling, 60*(1), 33–52. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/3_PTAM_Engelhard__Wang___Wind__2018-03-10__1855.pdf.

Eurostat. (n.d.). https://ec.europa.eu/eurostat/web/population-demography-migration-projections/data. Accessed 30 August 2019.

Extra, G., Spotti, M., & Van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives on integration regimes*. London, New York: Continuum.

Ferbežar, I. (2012). "Izrekam zvestobo moji novi domovini Republiki Sloveniji …": testiranje znanja slovenščine kot drugega in tujega jezika v Sloveniji. *Jezik in slovstvo, 57*(3–4), 29–45.

Ferbežar, I., & Stabej, M. (2002). Slovenian as a second language: Infrastructure and language policy. *Strani jezici, 32*(3–4), 235–243.

Ferbežar, I., & Stabej, M. (2013). Slovene or not Slovene? Issues on testing speakers of closely related languages. *Strani jezici, 42*(4), 338–353.

Ferbežar, I., Pirih Svetina, N., & Lutar, M. (2014). The Common European Framework of Reference: a reference for Slovene. *Linguistica, 54*(1), 277–291. https://revije.ff.uni-lj.si/linguistica/article/view/2615/2738.

Hogan-Brun, G. (2009). The politics of language and citizenship in the Baltic context. In G. Extra, M. Spotti, & P. Van Avermaet (Eds.), *Language testing, migration and citizenship. Cross-national perspectives on integration regimes* (pp. 37–560). London, New York: Continuum.

Huei-Lien Hsu, T. (2016). Removing bias towards world Englishes: The development of a rater attitude instrument using Indian English as a stimulus. *Language Testing, 33*(2), 367–389. https://journals.sagepub.com/doi/pdf/10.1177/0265532215590694.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182. https://journals.sagepub.com/doi/pdf/10.1177/0265532209349467.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3–31. https://journals.sagepub.com/doi/pdf/10.1191/0265532202lt218oa.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158–180. https://journals.sagepub.com/doi/10.1177/026553229801500202.

McNamara, T. (1996). *Measuring second language performance*. London, New York: Longman.

Minister of education regulation on the educational program Slovene as a second and a foreign language. (2015). http://pisrs.si/Pis.web/pregledPredpisa?id=ODRE2325. Accessed 31 August 2019.

Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2). http://jimelwood.net/students/grips/tables_figures/myford_wolfe_2004.pdf.

OneClick survey 1KA. (n.d.). https://www.1ka.si/d/en. Accessed 31 August 2019.

Požgaj Hadži, V., Balažic Bulc, T., Miheljak, V. (2009). Srbohrvaščina v Sloveniji: Nekoč in danes. In V. Požgaj Hadži, T. Balažic Bulc, V. Gorjanc, V. (Eds.). *Med politiko in stvarnostjo. Jezikovna situacija v novonastalih državah bivše Jugoslavije* (pp. 27–40). Ljubljana: Znanstvena založba FF UL.

Republic of Slovenia, Statistical Office. (n.d.). http://www.stat.si/statweb/News/Index/7485. Accessed 27 July 2018.

Šabec, K. (2006). *Homo europeus. Nacionalni stereotipi in kulturna identiteta Evrope.* Ljubljana: Fakulteta za družbene vede.

Šabec, K. (2007). Kdo je čefur za kranjskega Janeza: stereotipi in kulturne razlike v sodobnem evropskem kontekstu. In I. Novak Popov (Ed.). *Stereotipi v slovenskem jeziku, literaturi in kulturi* (pp. 102–116). Ljubljana: Filozofska fakulteta.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465–493. https://journals.sagepub.com/doi/pdf/10.1177/0265532208094273.

Slovenian as a second and foreign language [Slovenščina kot drugi in tuji jezik]. (2014). https://centerslo.si/wp-content/uploads/2015/10/Slovenscina_kot_drugi_in_tuji_jezik.pdf. Accessed 31 August 2019.

Stabej, M. (2010). *V družbi z jezikom.* Ljubljana: Trojina.

The Centre's annual reports. (n.d.). https://centerslo.si/en/books/annual-reports. Accessed 31 August 2019.

The Centre's web page. (n.d.). https://centerslo.si/en. Accessed 31 August 2019.

Ule, M. (2005): Predsodki kot mikroideologije vsakdanjega sveta. In V. Leskošek (Ed.), *Mi in oni. Nestrpnost na Slovenskem* (pp. 21–40). Ljubljana: Mirovni inštitut.

van Dijk, T.A. (1987). *Communicating racism. Ethnic prejudice in thought and talk*. Newbury Park, Beverly Hills, London, New Delhi: Sage Publications. http://www.discourses.org/OldBooks/Teun%20A%20van%20Dijk%20-%20Communicating%20Racism.pdf.

Wind, A.W., & Peterson, M.E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161–192. https://journals.sagepub.com/doi/full/10.1177/0265532216686999.

Wind, S.A. (2019). A nonparametric procedure for exploring differences in rating quality across test-taker subgroups in rater-mediated writing assessment. *Language Testing*, April 2019, 1–22. https://journals.sagepub.com/doi/pdf/10.1177/0265532219838014.

Winke, P., Gass, S., & Myford, C. (2013). Raters L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231–252. https://journals.sagepub.com/doi/pdf/10.1177/0265532212456968.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170. https://journals.sagepub.com/doi/pdf/10.1177/0265532209349465.

Part III
Learning from Program-Level Language Tests

The six chapters in this partfocus on learning from program-level language tests, with experience-based papers in chapters 16–19 and data-based papers in chapters 20–21. These chapters are as follows:

- Chapter 16 "EFL Placement Testing in Japan," by Carpenter and Matsugu, addresses *Filiopietism*, which is defined by the authors as the uncritical adherence to making decisions as they have been made in the past. This chapter suggests the need for considering filiopietism and its role in determining the reliability, validity, and practicality in program level language tests.
- Chapter 17 entitled "TEFL Test Practices at a Ukrainian University: Summative Test Design Through Teacher Collaboration," by Kvasova, describes the many problems faced by teachers untrained in language testing when they are required to produce a summative test (of grammar and receptive language skills, but also of speaking performance) in a Ukrainian university. The author recommends a solution through conducting a series of workshops and demonstrates how this can be accomplished.
- In Chapter 18 "Designing a Multilingual Large-Scale Placement Test with a Formative Perspective: A Case Study at the University of Grenoble Alpes," by Cervini and Masperi, the authors describe how designing the SELF test in six languages presented a number of challenges, including maintaining the same communicative construct across languages, enhancing item writers' abilities, dealing with logistical and technical issues, and coordinating the varied teams involved.
- In the next chapter (Chapter 19) Santavicca addresses "The Relationship Between English Placement Assessments and an Institution: From Challenge to Innovation for an Intensive English Program in the USA." In this chapter, the author discusses how stakeholder relations, student language skills, assessment development, and testing innovation all present challenges for institutional placement testing. Assessment design and administration principles and procedures for dealing with such challenges are also shared.
- Chapter 20 "Placement Decisions in Private Language Schools in Iran," by Razavipour and Firoozi, examines placement testing issues in Iran especially with regard

to the content areas tested, test taker characteristics, institutional issues, test users, and issues of power. To address these issues, the authors suggest: raising stakeholders' awareness of the hidden agendas involved in English language testing and teaching; enhancing the assessment literacy among all stakeholders; and establishing national and local standards for language placement testing and decision making.

- In Chapter 21 Figueroa and Zimányi examine the "Perceptions of (Un)Successful PET Results at a Private University in Mexico," focusing on the difficulties faced by the B2-C2 high-school students related to institutional test score requirements. The authors suggest re-evaluating institutional policies especially in test selection to account for the standard error of measurement as well as consequential validity.

With the focus on program-level language tests, five areas of challenge emerge as topics in the chapters in this part: filiopietism in Japan, untrained teachers who are required to develop summative language tests in Ukraine, designing a test in six languages in France, institutional placement testing in the USA and Iran, and institutional policies in test selection in Mexico.

All of the chapters in Part III deal with relatively high-stakes assessments with four chapters focusing on placement testing, one on summative testing and another on the use of the PET, an assessment from the Cambridge suite of exams. The global reach of these chapters covers North American contexts like the USA and Mexico (Chapters 19 and 21), Europe (Chapters 17 and 18), Asia (Chapter 16) and the Middle East (Chapter 20). Of the four chapters that share the theme of placement testing, issues of large-scale placement in six different languages is the focus of Chapter 18, institutional placement testing issues are examined in Chapters 16 and 19, and the challenges faced in private language school placement are under study in 20.

A common link in all the chapters of Part III is the need for more enhanced levels of assessment literacy for all stakeholders in the assessment process.

Chapter 16
EFL Placement Testing in Japan

James Carpenter and Sawako Matsugu

Abstract In many university-level foreign language programs throughout the world, teachers and administrators struggle to develop valid and reliable assessments. Foreign language test developers are trained to use statistical methods to help ensure the quality of their assessments. Nevertheless, many foreign language programs face a variety of institutional constraints that influence their ability to apply "best practices" in test design. In the testing literature, such constraints are often grouped under the broad category of "practicality." This chapter, however, proposes that universities, like any social system, operate according to an internal logic. The systematic elements of this internal logic are often based on decisions made in the past that continue to unconsciously guide decision-making in the present. Filiopietism refers to this kind of uncritical devotion to "the way things have always been done." This chapter will describe the influence that filiopietism has had on the placement testing practices in an EFL program at a Japanese university. To do this, the chapter analyzes articles published in the program's in-house journal over a 20-year period. This chapter will demonstrate that testing practices in foreign language programs can arise organically from the internal logic guiding institutional decision-making. Filiopietism, in other words, plays an outsized role in determining how validity, reliability, and even practicality are factored into test design.

16.1 Introduction: Purpose and Context

This paper discusses the placement testing practices in an English as a Foreign Language (EFL) department at a university in Japan from a critical discourse perspective. Our goal in adopting this perspective was to understand the power relations and ideological processes in the discourse surrounding the creation of the placement test

J. Carpenter (✉)
Rikkyo University, Tokyo, Japan
e-mail: james.carpenter@rikkyo.ac.jp

S. Matsugu
Meiji University, Tokyo, Japan
e-mail: smatsugu@meiji.ac.jp

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_16

(Fairclough 1989). Arguably, discussions about EFL testing practices within the field of applied linguistics tend to focus on the presence (or absence) of “best practices,” with particular attention paid to concepts such as validity, reliability, and practicality. In the field, it may be widely assumed that when each of these concepts is thoroughly and systematically addressed, the test can be considered sufficiently effective. While this perspective is justifiably pervasive because language test designers are primarily concerned with creating good tests, we argue that such concepts represent only one level of analysis. Another level of analysis is that of the distinctly social and political dimensions that guide the creation of the tests (Pennycook 2001).

This study focuses on the single case of an EFL department at University X, a mid-tier, private university in Japan. We chose this university because, we argue, University X represents a “typical case” of such programs in Japan (Yin 2003). Our discussion draws heavily from the test evaluation reports generated by the EFL department’s assessments committee over a twenty-year period. These reports were chosen because we were interested in how individual assessment committee members dealt with practical, social, and political administrative constraints.

16.1.1 Testing Context

The EFL department at University X primarily teaches in the Freshman English (FE) program. Students are sorted into their FE class levels using the Freshman English Placement Test (FEPT). The FEPT is administered at the beginning and at the end of each academic year. The current version of the FEPT contains two parts—a listening section and a reading section. In the current version of the test, there are 74 questions consisting of 39 listening questions and 35 reading questions. Based on the FEPT scores, students are sorted into classes by faculty, with a maximum of 20 students per class. This means that the 20 highest test scores are placed in the level one class, the next 20 in the level two class, and so on. Of the five faculties that sort students based on the FEPT, the law faculty has 22 levels, the economics faculty has 16 levels, the business faculty has 21 levels, the business hospitality faculty has six levels, and the civil engineering faculty has eight levels.

16.2 Testing Problem Encountered

In the first chapter of their book *Language Assessment in Practice*, Bachman and Palmer (2010) state the following:

> The practice of using the same test year in and year out, simply because “it works,” or of mimicking whatever test method is currently in widespread use, provides no basis for justifying test use if and when the developer is held accountable by stakeholders, including students, teachers, and administrators. (p. 9)

The FEPT was first conceived in 1996 (Sinnot 1996), and first administered to students at University X in 1997 (Forster and Kearney 1997). Since this time, the test has been revised in a number of different ways, for a number of different reasons (Hull 2012). However, as Carpenter (2016) noted: "…the design principles first envisioned by Sinnot (1996) and Forster and Kearney (1997) have not been seriously reconsidered since the FEPT was first administered" (p. 8). This is so evidenced by the lack of any serious attempt at explaining the validity of the FEPT in any of the papers written about the test since 1996 (e.g., Sinnot 1996; Forster and Kearney 1997; Ridge and Matsuta 1999; Ridge 2000; Wilson and Hansford 2001; Hansford 2004; Messerklinger 2007, 2008; Hull 2012, 2013; Hull and Brennan 2014; Hull et al. 2015). Conventionally, validity has been defined as "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (Bachman and Palmer 1996, p. 21). In other words, validity is a test developer's argument for why their test should be taken seriously as an accurate measurement. However, the meaningfulness of this argument, and, indeed, the accuracy of the measurement, is situated within the particular people, situations, and interactions that constitute the testing context (Mislevy 2018). As a result, psychometrically invalid measures can come to be considered valid from the point of view of the stakeholders in a particular context.

In the case of the FEPT, this issue was partially exemplified by such findings as those of Messerklinger (2008), who stated that "the fact that there is no connection between [FEPT] scores and grades shows very clearly that the test has little to do with the curriculum, and the test's weaknesses demonstrate that it is not a valid assessment of language ability" (p. 14). That the test has both no connection to the curriculum, and no explanation for why this is so means that, from a psychometric perspective, it impossible to justify the continued administration of the FEPT year on year. Yet, according to Bates (2018), the same problems persist with the current version of the FEPT. Therefore, some aspects of the situated factors that shape the context of University X must contribute to the continued recycling of an invalid instrument.

Conventional explanations for the uncritical recycling of the same test, including those given in Bachman and Palmer (2010), tend to focus on the importance of test literacy: if EFL teachers were better trained in test design and construction, major threats to test validity and reliability could be sufficiently addressed. Our analysis however, is that while test literacy is important, it is not the primary reason for the placement testing practices observed in EFL departments like University X. As Shohamy (1997) has proposed, "the act of language testing is not neutral. Rather, it is a product and agent of cultural, social, political, educational, and ideological agendas that shape the lives of individual participants, teachers, and learners" (p. 2). The question, then, is what various agendas are shaping and have shaped the development of the FEPT?

16.2.1 Filiopietism

Universities tend to follow conserver models of organizational behavior (Crow and Shangraw 2016). Raadschelders and Vigoda-Gadot (2015) define conserver models in terms of the "desires to maintain security and…not take risks" (p. 154). While universities are not traditionally thought of in terms of bureaucracies, all institutions tend to operate according to an internal logic dedicated to maintaining the status quo (Crow and Dabars 2015).

One term for the excessive veneration of the status quo is *filiopietism* (Crow and Dabars 2015). Menand (2010) defines filiopietism in this way: "The system gets internalized. It becomes a mind-set. It is just 'the way things are,' and it can be hard to recover the reasons why it is the way things are" (p. 116). Broadly speaking, filiopietism refers to an administrative status quo that continues to guide decision-making processes because relevant stakeholders feel that they have no choice. Such organizational structures, which serve as a backdrop to the educational and research goals of most universities, eventually constrain the very goals they are assumed to support (Crow and Dabars 2015).

In the Japanese context, one of the more serious effects of filiopietism is isomorphism. This concept refers to the tendency for organizations in general, and universities in particular, to "emulate one another and become increasingly homogeneous" (Crow and Dabars 2015). We identified two organizational features of the EFL department at University X that seem to particularly represent institutional isomorphism: (1) the hiring practices of the EFL instructors, and (2) the influence of the TOEIC test. We discuss how each of these features influence test development below.

16.2.1.1 Hiring Practices

Native-Speakerism The relationship between EFL instruction and the Japanese education system is complex. In brief, the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) is currently planning to implement major educational reforms at every level of the education system from the year 2020. These reforms include an expansion of "English education corresponding to globalization" (MEXT 2014), and the introduction of productive-skills tests into the entrance examination system (The Mainichi 2018). These proposals have not been without controversy. A committee known as the "Japanese Government Revitalization Unit" (GRU) working under the Cabinet Office wrote a 2010 proposal to review the Japan Exchange and Teaching (JET) program. As Hashimoto (2013) documents, the goal of this proposal was to investigate the ambiguous relationship between English education on the one hand, and international exchange on the other. This ambiguity about the role of the foreign-born educator extends back to the Meiji Restoration (ca. 1868), when Japan invited thousands of non-Japanese to accept the position of *gaikokujin kyoshi*, or foreign instructor. As Heimlich (2013) has pointed out, this newly formed occupational category served as a template for how non-Japanese would be integrated into the educational system at every level. Yet, as the GRU proposal clearly illustrates, major stakeholders in the Japanese education system have become dissatisfied with early definitions of "foreign instructor" as it relates to EFL educators. As a result, the term has largely been abandoned in favor of "native speaker," because this term more clearly indicates the difference in status and responsibilities between the foreign instructors and the tenured faculty (Heimlich 2013).

Employment Status The institutional practices at University X are similar to those described in Rivers' (2013) case study. All of the teachers in the EFL department are employed under one-year four-times renewable contracts. Regardless of their contributions to the institution itself, or the quality of their teaching, all teachers are terminated after this five-year period. According to the Japan Association of College English Teachers' (JACET) 2018 survey, the number of Japanese EFL instructors working under short-term contracts has been increasing. Therefore, while some of the employment practices described in Rivers (2013) do not apply exclusively to foreign EFL teachers across Japanese universities, they play a major role in how EFL programs are managed, and how those same programs evolve over time.

16.2.1.2 Influence of the TOEIC

The influence of the Test of English for International Communication (TOEIC) on English education at the university level is extensive. Since many Japanese companies began using TOEIC scores as a basis for hiring, promotion, and overseas assignments, Japanese universities have felt increasing pressure to produce graduates with high TOEIC scores (Takahashi 2012). The result for Japanese universities has been that (1) the TOEIC is often used for admissions and placement decisions (Weaver 2016), (2) some academic credits are awarded on the basis of TOEIC scores alone (see In'nami and Koizumi 2017), and (3) universities use their students' average TOEIC scores as a promotion tool. According to the Institute for International Business Communication (IIBC) (2016), of the 751 Japanese universities that responded to their survey, 427 universities use the test as a part of making admissions decisions, and 378 universities used the test for awarding credits to students with a high score; which exempts those students from taking certain classes.

At University X, the influence of the TOEIC is obvious to the extent that the FEPT was clearly designed to resemble the TOEIC. Like the TOEIC, the listening section of the FEPT includes (1) picture identification, (2) question-response, and (3) dialogues-based listening tasks. In addition, the reading section of the FEPT contains (1) sentence completion, and (2) reading comprehension tasks. While the FEPT also has sections that are not in the current version of the TOEIC, the overall format of the test, the test booklet and answer sheet, and the format of the audio file resemble the TOEIC in substance if not in every particular.

16.3 Solution/Resolution of the Problem

As we have argued that the influence of filiopietism on EFL testing practices in University X is systemic, simple solutions are not possible. Yet, the necessity of finding a solution to similarly systemic problems in Japanese society as a whole is becoming increasingly obvious, and may be cause for some hope. In Tokyo, the percentage of foreign-born workers continues to increase precipitously. In 2018 the percentage of foreign nationals in Shinjuku-ward, a well-known commercial and administrative center of Tokyo, was 45.8% (Fulford 2019). As Japanese society becomes more and more cosmopolitan, the internal logic that has guided the hiring policies in EFL programs like that of University X may change.

Whatever the future holds, filiopietism continues to impact the quality of testing at University X. As the discussion above indicates, while the FEPT has been revised many times in the last 20 years, the underlying problems with the validity of the test remain (e.g., Carpenter 2016; Mabe 2017; Bates 2018). Also, as the discussion of the employment practices at University X indicate, every person who has worked to revise the test lost their job after their five-year contract period ended. While the test-analysis reports mentioned in this chapter represent some aspects of these test designer's thinking at different points in time, the documents cannot capture the larger thought processes that went into these revisions. In this way, whatever these educators hoped to accomplish as members of the assessments committee has been lost.

Interestingly, one solution to the relative isomorphism surrounding the use of the TOEIC in Japanese universities would seem to be allowing individual EFL departments to construct their own placement tests. Yet, as the discussion above indicates, the strong tendency for universities to resemble one another over time resulted, in the case of University X, in a so-called "homemade test" that still strongly resembled the TOEIC. Despite this, a solution to the institutional overreliance on the TOEIC is, in fact, imminent. In addition to the forthcoming curricular changes proposed by MEXT discussed above, the Japanese National Center for University Entrance Examinations has announced that it will approve seven commercially available tests for use in Japanese universities (The Mainichi 2018). The TOEIC was originally included on this list, but the Institute for International Business Communication, which oversees all administrations of the TOEIC in Japan, withdrew the test from consideration (The Mainichi 2019). This may force Japanese universities to reevaluate their use of the TOEIC. The extent to which Japanese universities will be able to respond to these and similar external calls for change remains an open question.

16.4 Insights Gained

Our goal in this paper was to discuss filiopietism with the hope of contributing to an improvement in the placement testing practices at Japanese universities. However, as we hope our discussion has illustrated, it is not possible to "solve" EFL placement testing practices without considering the larger administrative context in which they appear. It is highly problematic, for example, to employ instructors with only a limited contract (e.g., 1, 3, 5 years), while simultaneously expecting testing practices (to say nothing of teaching practices) to improve over time. To put it bluntly, it is our contention that people do not work well when they cannot foresee any kind of stable future.

Relatedly, because EFL instructors do not have the option to continue working beyond a fixed period of time, the accumulated knowledge about the institutional context and the student population itself, is lost with each passing generation. As a result, the thought processes that guided the successive revisions to the FEPT discussed in this paper were also lost. Test validity, if understood as "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (Bachman and Palmer 1996, p. 21), is therefore impossible to maintain.

More generally, because evidence suggests that instructors with limited-term contracts tend to be younger than their tenured colleagues (JACET 2018), they may in fact possess a more sophisticated level of test literacy. Indeed, one aspect of filiopietism not discussed in this paper that clearly also has a pervasive influence on EFL testing practices in Japanese universities is the clear division of responsibilities between tenured professors and short-term contract instructors. In Japan, tenured professors whose academic expertise is related to English are typically involved with major decisions about the university's EFL program (i.e., test and curriculum development). However, as the results of the JACET (2018) survey of English education referenced above clearly indicate, these same tenured professors do not necessarily have a background in applied linguistics, much less in language testing. As a result, the influence of filiopietism in Japanese universities may be extended to include a mismatch between the tenured faculty's expertise and their responsibilities. One unfortunate result occurring from this may be that test analyses and item banking are hardly done for entrance examinations (Mizumoto et al. 2017).

This last point highlights the importance of test literacy, which, while not the focus of this paper, remains central to this discussion. The wide dissemination of quality information from professional language testing organizations like the International Language Testing Association (ILTA) or the Japan Language Testing Association (JLTA) would increase the likelihood of sensible testing practices being incorporated into the bureaucratic structure of Japanese universities in the future. To the extent that there is at least one instructor with language testing experience in a department, some short-term improvement may still be possible. If that same person could also offer professional development seminars to receptive colleagues, the improvements may even be durable in departments exclusively employing short-term contract instructors.

16.5 Conclusion: Implications for Test Users

Finally, in Japan, the relationships between the self and the group, and the continuity between the current group's actions and those of the past, can serve as the primary motivators for continuing to do many things (Doi 1971). Accordingly, the absence of a precedent is often sufficient reason not to try anything new. This aspect of Japanese culture has, perhaps, perpetuated the influence of filiopietism in educational settings: in particular, with regard to assessment. In this way, test designers can, potentially, become trapped between competing cultural discourses. On the one hand, according to our quote of Bachman and Palmer (2010) above, continuing to use the same test again and again just because "it works" is not sufficient. This quote, and the larger body of language testing scholarship it represents, assumes that a series of objective best practices about test design both exist, and can be abstracted from one context and applied to another. Japanese educational practices on the other hand, like Japanese culture in general, may be considered *particularistic* in the sense that it is those people with the most history in a given context who decide what so-called best practices should be (Brown 1993). While educators of either persuasion would likely agree with the statement that students have the right to be fairly assessed, the very meaning of *fairness* remains at issue. Perhaps like many of the authors in this volume, we would argue that applying well-researched, general principles of test design constitute an important aspect of fairness. In Japan, however, fairness is not necessarily about measuring the differences between individuals within groups in some objective sense, but rather ensuring that the circumstances across different groups remain essentially the same (Kang 1990). Put another way, even if a test is poorly designed, as long as everyone has to take it, the test may still be considered, essentially, fair. From this perspective, an educational context heavily influenced by filiopietism is essentially fair, too. The challenges are significant for those who disagree.

References

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* New York, NY: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. New York, NY: Oxford University Press.

Bates, D. (2018). An analysis and review of the 2017 Freshman English placement test at Asia University. *CELE Journal, 26,* 1–11.

Brown, D. M. (1993). *The Cambridge history of Japan: Ancient Japan.* New York, NY: Cambridge University Press.

Carpenter, J. (2016). Past, present, and future assessment practices at CELE: A view from 2015. *CELE Journal, 24,* 52–78.

Crow, M., & Dabars, W. (2015). *Designing the new American university*. Baltimore, MD: Johns Hopkins University Press.

Crow, M., & Shangraw, R. F. (2016). Revisiting "public administration as a design science" for the twenty-first century public university. *Public Administration Review, 74*(5), 762–763.

Doi, T. (1971). *The anatomy of dependence: The key analysis of Japanese behavior*. Tokyo, Japan: Kodansha International Ltd.

Fairclough, N. (1989). *Language and power*. Harlow, UK: Longman Group UK Limited.

Forster, D. E., & Kearney, M. (1997). Writing the freshman English placement test. *ELERI Journal, 5,* 144–157.

Fulford, A. (2019). Back-tracking: A path to the future in regional Japan. *Japan SPOTLIGHT, 223,* 51–54.

Hansford, V. (2004). *CELE's OPI system. CELE Journal, 12,* 95–102.

Hashimoto, K. (2013). The construction of the 'Native Speaker' in Japan's educational policies for TEFL. In S. A. Houghton & D. J. Rivers (Eds.), *Native-speakerism in Japan: Intergroup dynamics in foreign language education* (pp. 159–168). Bristol, UK: Multilingual Matters.

Heimlich, E. (2013). The meaning of Japan's role of professional foreigner. In S.A. Houghton & D.J. Rivers (Eds.), *Native-speakerism in Japan: Intergroup dynamics in foreign language education* (pp. 169–183). Bristol, UK: Multilingual Matters.

Hull, J. (2012). Modifying Asia University's freshman English placement test. *CELE Journal, 20,* 1–11.

Hull, J. (2013). Review and analysis of Asia University's freshman English placement test. *CELE Journal, 20,* 1–11.

Hull, J., & Brennan, J. (2014). Review and analysis of Asia University's 2013 freshman English placement test: Transition from version 2.4 to version 2.5. *CELE Journal, 22,* 35–62.

Hull, J., Brennan, J., & Wells, L. (2015). Review and analysis of Asia University's freshman English placement test: Transition from version 2.5 to version 2.6. *CELE Journal, 23,* 21–49.

IIBC. (2016). TOEIC tests nyugaku shiken/tanni nintei katsuyo jokyo [Report on the use of TOEIC for entrance examinations and awarding credits]. https://www.iibc-global.org/toeic/official_data/lr/search.html. Accessed 1 February 2019.

In'nami, Y., & Koizumi, R. (2017). Using EIKEN, TOEFL, and TOEIC to award EFL course credits in Japanese universities. *Language Assessment Quarterly, 14,* 274–293.

JACET. (2018). Daigaku eigokyoiku no ninaite ni kansuru sogoteki kenkyu [Comprehensive research on prospective educators of university English education]. http://www.jacet.org/wp-content/uploads/実態調査委員会報告書WEB掲載版20180904-1.pdf. Accessed 29 September 2018.

Kang, T. W. (1990). *Gaishi: The foreign company in Japan.* New York, NY: Basic Books.

Mabe, K. (2017). Review and analysis of Asia University's 2016 freshman English placement test: The need for major or minor changes? *CELE Journal, 25,* 1–16.

Menand, L. (2010). *The marketplace of ideas: Reform and resistance in the American university.* New York, NY: Norton.

Messerklinger, J. (2007). The new freshman English placement test. *CELE Journal, 15,* 14–22.

Messerklinger, J. (2008). Results of the 2007 FEPT. *CELE Journal, 16,* 6–16.

MEXT. (2014). *English education reform plan corresponding to globalization.* http://www.mext.go.jp/en/news/topics/detail/__icsFiles/afieldfile/2014/01/23/1343591_1.pdf. Accessed August 1 2018.

Mislevy, R. (2018). *Sociocognitive foundations of educational measurement*. New York, NY: Routledge.

Mizumoto, A., Wakita, T., & Nabei, T. (2017). An analysis of English language entrance examination data from Kansai University: Applying language testing theory. *Bulletin of Data Analysis of Japanese Classification Society, 6,* 21–29.

Pennycook, A. (2001). *Critical applied linguistics: A critical introduction.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Raadschelders, J., & Vigoda-Gadot, E. (2015). *Global dimensions of public administration and governance: A comparative voyage.* Hoboken, NJ: John Wiley & Sons.

Ridge, P. (2000). Results of the 1999/2000 FEPT and the 1998/1999 I-TOEFL tests. *CELE Journal, 8,* 63–65.

Ridge, P., & Matsuta, K. (1999). Results of the 1998/1999 FEPT and 1997/1998 I-TOEFL tests. *CELE Journal, 7,* 75–77.

Rivers, D. J. (2013). Institutionalized native-speakerism: Voices of dissent and acts of resistance. In S. A. Houghton & D. J. Rivers (Eds.), *Native-speakerism in Japan: Intergroup dynamics in foreign language education* (pp. 75–81). Bristol, UK: Multilingual Matters.

Shohamy, E. (1997, March 10). *Critical applied linguistics and beyond*. Plenary address to the American Association of Applied Linguistics, Orlando, FL.

Sinnot, L. (1996). What's behind the proposal for the new ELERI freshman English placement exam? *ELERI Journal, 5,* 177–184.

Takahashi, J. (2012). An overview of the issues on incorporating the TOEIC test into the university English education curricula in Japan. *Tama University Bulletin, 4,* 127–138.

The Mainichi. (2018, March 27). Eight private English tests accepted for future university admission system. *The Mainichi.* https://mainichi.jp/english/articles/20180327/p2a/00m/0na/013000c. Accessed 29 September 2018.

The Mainichi. (2019, July 2). TOEIC withdrawing from university admission system. *The Mainichi.* https://mainichi.jp/articles/20190702/k00/00m/040/062000c?pid=14517. Accessed 29 September 2018.

Weaver, C. (2016). The TOEIC IP test as a placement test: Its potential formative value. *JALT Journal, 38,* 5–25.

Wilson, R., & Hansford, V. (2001). Results of the 2000/2001 FEPT and 1999/2000 I-TOEFL tests. *CELE Journal, 9,* 175–181.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

Chapter 17
TEFL Test Practices at a Ukrainian University: Summative Test Design Through Teacher Collaboration

Olga Kvasova

Abstract This chapter presents a case study of the development of summative test design in a Ukrainian university. The increased role of assessment of learning, after Ukraine joined the Bologna Process, posed multiple challenges to university foreign language (FL) teachers, as most of them lacked training in language testing and assessment (LTA). The author of this chapter, who received special training in LTA, shared her knowledge and experience with fellow teachers in the TEFL department, acting as an LTA teacher trainer and an observer of the team's cooperation in building assessment skills. She conducted a series of workshops in the department to help teachers raise their awareness of assessment cornerstones, which resulted in the collaborative development of achievement tests for particular units, as well as increased assessment literacy of the faculty members. Summative assessment involved the development of tasks to test receptive and grammar skills, as well as speaking performance.

17.1 Introduction: Purpose and Testing Context

The introduction of the European Credit Transfer and Accumulation System (ECTS) in Ukrainian higher education in 2005 prompted the redesign of the FL curricula in accordance with the ECTS. This development elevated the role of assessment for accountability. A test format was deemed appropriate to tackle the frequent assessment of large numbers of students. This entailed a considerable shift of teachers' workload from delivering instruction only to additionally developing summative assessments, thus posing multiple challenges.

Following the introduction of ECTS, in the mid-2000s, the Ukrainian government also introduced the Unified External School-leaving examination in major school subjects including FL. Responding to the urgent need for preparing a dedicated group of people to develop language tests, the British Council Ukraine launched the project *Assessment in ELT*. The project involved training the participants, English

O. Kvasova (✉)
Taras Shevchenko National University of Kyiv, Kiev, Ukraine

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_17

teachers from across the country, in LTA fundamentals, test design, and item writing. Since the majority of the participants were university teachers, they applied their knowledge and skills in LTA in their workplace to assist fellow teachers in handling the challenges of the ECTS-compatible assessment for accountability. The author of this chapter participated in the project and further enhanced her assessment literacy through independent self-study and cooperation with international LTA experts.

This chapter addresses the testing practices of the author's fellow teachers' group in their workplace—one of TEFL departments at the University of Kyiv. It focuses on the development of the mid-term and end-of-course tests for students of linguistics through teacher collaboration.

17.2 Testing Problems Encountered

The implementation of ECTS-compatible, summative written mid-term tests, which includes reporting the results against a 100-point scoring system, raised some concerns. The main problem was that although the authorized curriculum was implemented, assessment was not uniform—topics, testing methods (e.g., tests, dictations, compositions, translations into English), and scoring were at the discretion of teachers. This prioritized the development of unified tests to elicit reliable evidence regarding students' achievements.

Initially, all teachers were to be engaged in the test preparation procedure since "Assessment is a collegial activity" (Coombe et al. 2007, p. 13) and a good test cannot be written by one person however knowledgeable and experienced that person is. Involving everyone proved to be a difficult task as teachers required advanced training in test development in order to be efficient. Aware of this limitation, a group of teachers working on the same unit, agreed to compile a mid-term test following the test specifications and utilizing tasks provided by international exam systems, mainly of the First Certificate in English test. The necessity to administer the written summative test within a typical 80-minute class required essential adaptation of the initial summative test specifications. Based on their teaching experience, the group of teachers managed to decide on a realistic number of tasks for each part of the summative test: Listening (1 task), Reading (1 task), Vocabulary (2 tasks), Grammar (2 tasks), and Writing (1 task). Once the first draft of the test was compiled by the author of this chapter, the group members were invited to pilot it in order to finalize the keys, check time allotments, and come up with suggestions for improving the test, if necessary. However, only four of the group members supported the initiative, did the trialing, and discussed the test quality. All other group members made their remarks on the test quality through informal discussions after test administration. Collaboration on the summative test design was progressing although slower than expected.

At that point in time, listening and reading test tasks were selected regarding their relevance to a particular topic area studied during the mid-term. However, this principle could not ensure the consistency of the subskills tested in a selected test task and the subskills taught and expected to be assessed. As the testing process evolved, such mismatches were not infrequent. To address the mismatches identified in the reading and listening test tasks, teachers decided to develop their own reading test tasks, adapt listening tasks without modifying the recordings, as well as tailor test tasks of grammar, vocabulary, and writing to the curriculum. Therefore, tailoring the ready-made test tasks to the needs of the Ukrainian university classes was the first challenge encountered by the department. Writing their own quality items and test tasks in order to use them as valid and reliable instruments for measuring students' achievement was yet more critical. Both problems revealed the need for teachers to receive on-the-job training to learn the essentials of item/task/test writing instead of relying on experience and intuition.

Another issue arose in relation to the assessment of productive skills, i.e., speaking and writing. Traditionally, in the Ukrainian educational framework, the dominant way of testing speaking skills involves providing examinees with a list of topics to be studied before the exam and then requiring them "to speak about the topic." The topics are formulated in quite an abstract fashion, such as "A question of health" or "Modern wonders of the world." The students have to prepare and merely reproduce the pre-prepared script at the exam. The teacher's function is to listen to the delivery and occasionally ask questions, thus imitating discussion. There are no clearly stated criteria to assess such prepared monologues, and teachers assess the speaking samples by judging content, linguistic (mostly grammatical) accuracy, and fluency.

In line with administering a summative assessment of oral production as a test, an attempt was made to alter the "speak about the topic" format following the pattern of acclaimed speaking tests. The faculty members in the department, however, were neither engaged in discussing or trialing the tasks, nor provided with the basic theory or practical tips. Instead, they were asked to develop cards with three types of tasks (a guided interview, comparing and contrasting/describing pictures, and an individual long turn) as in Sample Card 1 (Appendix 17.1a). At first glance, the new speaking tasks seemed more sophisticated than the previous format. However, after some training sessions on assessing speaking, the team agreed that employing three types of tasks on the same theme was impractical and redundant. Also, a concern was raised about the task "describing pictures," as teachers were not trained in selecting pictures that could serve as communicative stimuli; as a result, the pictures performed a decorative rather than a stimulating function.

With respect to assessing writing skills, it was mostly intuitive on the part of the teachers in the department. Criteria such as "content," "grammar," and "vocabulary" were the most popular among the teachers, although a specific scoring method based on those criteria was not identified. The most frequent practice was negative marking by deducting marks for errors. Needless to say, the students did not receive proper feedback, were not aware of how their written papers were judged, and were not told what could be done to improve their writing in the future.

17.3 Resolution of the Problems

To bridge the gap in teachers' assessment literacy which hindered efficient test design, a series of workshops was conducted at that time by the author acting as a teacher trainer. During the workshops, the participants were introduced to the cornerstones of LTA, received hands-on experience in identifying the purposes of particular test tasks, and learned how to discriminate between good and faulty test items, write appropriate instructions for test tasks, and construct different types of items/tasks. Further workshops were aimed at learning how to test receptive skills, including grammar and vocabulary.

Collaborative work on test design offered a new perspective that eventually helped tackle effectively the issues that occurred. Some examples of effective collaborative test task design are provided below.

Testing receptive and grammar skills. Meeting the widely spread preference of teachers for multiple choice questions (MCQs), two consecutive workshops were dedicated to training teachers on constructing and validating MCQs for reading. Apart from receiving training on constructing MCQs, the teachers were encouraged to collaborate on all of the phases of the testing cycle. The specific objective was to enable teachers to conduct an analysis of item difficulty and descriptor efficiency. The procedures were adapted to the classroom context based on the materials offered by Coombe et al. (2007).

Below is an example (Fig. 17.1) of a reading MCQ item that underwent several stages of collaborative preparation:

The item was initially trialed by two colleagues and further piloted by three others in the classroom environment, with data collected from 60 students. Then, item difficulty and distractor efficiency analyses were conducted.

The item indexes convinced teachers that the correct option (C) had good distractor efficiency and, therefore, was efficient, whereas option (D), with 0.00, was inefficient and therefore useless. Guided by the trainer, the teachers modified the item by shortening the stem and making it more focused, also placing the word "father," which was common to all options, in the stem. In addition, the options were reworded in plain language, the shades of meaning became clearer in the options, and, finally, the

Read the passages below and answer the questions choosing the best answer (A-D).
(1) "Sean (the father) peered over their (the doctors') shoulders watching his new born baby. 'She's perfect,' he said, turning to me (the mother), but the words curled up at the end like a puppy's tail, looking for approval…
…Perfect babies didn't sob so hard that you could feel your own heart tearing down the center…" [Picoult J. "Handle with Care," pp.5-6].
1. Which of the following is an unstated assumption made by the author?

A	The father admired his new born daughter, finding her beautiful.	0.11
B	The father feared that his wife did not share his admiration.	0.22
C*	The father was apprehensive that something might be wrong with the baby.	0.67
D	The father couldn't decide if his baby was really a beauty.	0.00

Fig. 17.1 MCQ item

According to the text, the father ....

A admired the beauty of his newborn daughter	0.15
B doubted that his wife shared his admiration	0.11
C* felt something was wrong with the little girl	0.55
D was aware that the newborn was fatally ill	0.19

Fig. 17.2 Modified MCQ item

State the difference in meaning between the sentences (1 point for each correct answer).
1.a) I know! I'll ask my boss for a pay raise tomorrow.
b) I've arranged to see my boss tomorrow. I'm going to ask for a pay raise.

Fig. 17.3 Produced response task

options were modified to become parallel and of similar length. The modified item with new indexes of distractor efficiency is provided below (Fig. 17.2):

For testing grammar, which traditionally favored MCQs, the teachers began to seek more varied formats. Below (Fig. 17.3) is an example of a produced response task included in the summative test:

When doing a similar activity orally in the classroom, the students were taught to follow the pattern: "The Future Simple is used to denote a promise or a decision made on the spot. The structure 'going to V' is used to denote an arranged action." However, the instructions on the written test task did not contain an example of how responses should be formulated. Therefore, some of the written responses were close to a full oral answer, e.g., "The form is used to talk about an on-the-spot decision" or an interpretation of the usage, e.g., "We can see that you are willing to ask for a pay raise," whereas others simply indicated "to promise" and "to have plans" or just named the verb form "future tense." It is obvious that, although the keys were supplied (they were formulated in a full, rule-like fashion), when scoring the items, the examiners attempted to understand what each of the test takers meant. Some of the answers were never deciphered. Not only were test transparency and the practicality of the test task violated but validity was as well. As a result, even though some students possibly knew the correct answer, they could not express it comprehensibly.

Building on this experience, the teachers decided to provide a concise sample response making the item more testee- and tester-friendly (Fig. 17.4).

Testing speaking. In addition to enhancing the quality of the written summative test, the team of teachers worked on the oral summative end-of-term test. Adhering to the required format (Appendix 17.1b), they rearranged the tasks so as to focus

State the difference in meaning between sentences A and B. You will receive 1 point for each correct answer. See the example. Write your answers on the answer sheet.
Example *(0)*: A She arrived late for the meeting. A not in time
B He hasn't been feeling well lately. B recently

Fig. 17.4 Modified produced response task

Explain what makes an advertising slogan efficient and persuasive.

Fig. 17.5 Draw-a-card task

Explain what makes an advertisement efficient and persuasive:
- what are the ingredients of a good advert?
- what makes some slogans more memorable than others?

Share your team's experience of creating and presenting an advertisement.

Fig. 17.6 Modified draw-a-card task

on different aspects of the topic and speaking subskills (Appendix 17.1a). Thus, "describing a picture" served as a lead-in, "answering questions" aimed to develop the ideas offered by visual prompts and elicited students' knowledge of content, whereas the "long turn" invited test takers to go into more detail and express their own ideas.

Despite admitting the improved efficiency of this format, the teachers suggested that the task "describe pictures" should not be included in the card. Instead, they decided to focus on students' coverage of the content points studied during the half-term in a real-life like communicative situation. During the test, the students were asked to draw a card with a task on it, take 30–45 seconds to plan the utterance, and then talk for at least two minutes. The element of spontaneity was preserved by suggesting exactly what students should do: discuss, share impressions, provide reasoning, prove/disprove, etc. One such task is the example below (Fig. 17.5):

The task underwent one more modification so that it could be used as a tool in end-of-term assessment. After modification, the exam task (Fig. 17.6) included instructions about what to do (*to explain*), notes on the content that was to be presented in a coherent oral text, and a directive to *express personal views on the issue*:

Therefore, the students received tasks with clear instructions to cover main content points and rhetoric functions, leading to reliable assessment results. Besides, verbal prompts helped students control test anxiety and perform efficiently. Having covered the key content points, they had the freedom to demonstrate their creativity and mastery of the language. On the other hand, poor content knowledge reduced the chances of passing the test. From the assessor's perspective, the degree of revealing key content points and the use of relevant rhetoric functions served as a benchmark against which they were able to make fairer decisions about candidates' task achievement, i.e., sufficiently or insufficiently full/logical/coherent. As a result, the teachers voiced the need to resort to rating scales, with a strong preference for analytic scales.

17.4 Insights Gained

In retrospect, the development of teacher-constructed test tasks for summative assessment yields interesting insights. There are limitations to the collaborative process presented in this study. The initial written test was developed almost totally by one person and included a number of violations of the test design procedure, such as insufficient pre-testing and analysis of test quality, the preparation of only one variant of the test, and its administration to different groups on different days. Nevertheless, the overall insights gained at this stage of assessment literacy development are quite important as teachers learned to make judgments about the appropriacy of using various ready-made test tasks and to prefer the tasks with immediate relevance to the language skills to be tested (thus attempting to adhere to content validity). Moreover, the teachers noted the right way to formulate instructions to tasks, the layout of test and answer sheets, and the agreement on and checking of keys. While discussing students' scores on the test and their fairness, the teachers admitted that the test format of summative assessment allowed for efficient feedback to learners. Moreover, the procedure for scoring tests and documenting the results helped the teachers adjust better to the recently introduced ECTS.

The procedure for constructing MCQ items was very beneficial to the teachers. First, they were convinced to adhere to the testing cycle and work collaboratively on all of its stages. Second, although being initially skeptical about the feasibility of carrying out statistical analysis in practice, the teachers admitted that this first-hand experience allowed for the vivid discrimination of good and inefficient options, suggesting reasons for misuse and ideas for modification. Third, the teachers, who previously constructed MCQs to check for understanding of details only, learned to shift the focus of such items to checking the gist and inferencing. The intuitive construction of MCQs and reliance on the teachers' own practical wisdom gave way to a theoretically grounded vision of writing this item type.

As far as the grammar test tasks described above are concerned, after a year of practicing such tasks for formative assessment and familiarizing students with them, the teachers experienced improved efficiency in their rating. The main lesson learned in the process of test development was avoiding formats unfamiliar to students, especially those missing examples of a response, in a summative test as they may violate test transparency and put the validity of scoring at stake.

The development of tasks and scales to assess productive skills was greatly influenced by the teachers' participation in the workshops conducted in Kyiv by international LTA experts between 2015 and 2018. The workshops were organized by the Ukrainian Association for Language Testing and Assessment (UALTA). The association was founded by the members of the department, with the author of the chapter being its founding president. The workshops were reported on the UALTA website (http://ualta.in.ua) and contributed to the enhancement of theoretical knowledge and the development of the practical assessment skills of the participating university teachers.

Currently, the development of assessment skills in the department is mainly driven by the need to develop context-specific instruments to assess students' FL writing. A collaboration with the Centre for Research of English Language Learning and Assessment (CRELLA), the University of Bedfordshire, UK within the Erasmus + Staff Mobility Program engaged the project team in the development of rating scales or, rather, modifications to existing scales. The project work included rater training (10 teachers in the department) followed by ratings of students' writing performances which were collected during the end-of-term exam. The results offer necessary information for scales modification and finalizing.

17.5 Conclusion: Implications for Test Users

This study presents assessment skill building for teacher-created assessments in a TEFL department through teacher collaboration. The progress in test construction ability was tangible enough to encourage further stages of test design evolution.

The involvement of a small group of teachers in test development (only five out of 22 working in the department) probably reflects the teachers' difficulty to perform additional duties because of the existing workload. It should also be noted that it would be beneficial for all teachers to be trained in LTA fundamentals to gain at least moderate assessment literacy. However, this could be realized only with the help of education managers, i.e., through organizing in-service LTA teacher training. So far, the number of participants in UALTA workshops has been increasing but participation in workshops cannot replace solid training.

With respect to who should undertake the development of summative tests in each department, the experience from the project described above shows that the presence of even a small team of committed teachers is a big asset to any department. While training all university teachers to be equally effective in LTA is not possible, it is quite realistic to provide advanced training to enthusiasts who will then contribute to the development of good summative tests.

As to the quality of teacher-constructed tests compared to standardized tests, Green's (2014) so-called "assessment wars," i.e., the relationship between teacher-made and external tests written by LTA professionals, the experts seem to find it good enough. Green, in particular, refers to Harlen (2007) who asserts, "[i]f scoring criteria are clearly specified, training provided and teachers' work moderated, the results over the extended period are at least as reliable as those from standardized tests" (cited in Green 2014, p. 212).

The above encouraging outcomes motivate further those who are committed to improving teachers' assessment competence to design quality summative tests at least on a local level. The significant interest of the LTA community in classroom-based assessment and teachers' assessment literacy testifies to an optimistic future for assessment literacy development.

Appendix 17.1a

Sample Card 1

Part One

What makes a language truly alive? When does a language die?

Why are so many languages in danger of dying out now?

Is it inevitable for minority languages to die? Why?/Why not?

Will English survive another thousand years? Give one reason why it will and one reason why it won't.

Part Two

Look at the pictures and tell what do you think is the difference between "a living language" and "a dead language."

Part Three

Dwell on the topic "Languages alive and dead"

Appendix 17.1b

Sample Card 2

Look at the pictures and comment on them.

Answer the questions:

- Why do communication difficulties between males and females occur?
- What is typical of male and female communication?

Provide some examples of miscommunication between men and women. Explain how misunderstandings could be prevented.

References

Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners.* Ann Arbor, Michigan: University of Michigan Press.

Green, A. (2014). *Exploring language assessment and testing: Language in action.* New York: Routledge.

Harlen, W. (2007). *Assessment of learning*. London: Sage.

Chapter 18
Designing a Multilingual Large-Scale Placement Test with a Formative Perspective: A Case Study at the University of Grenoble Alpes

Cristiana Cervini and Monica Masperi

Abstract An interdisciplinary team composed of more than thirty people has been engaged in the process of designing, developing, and validating an online placement test with a formative perspective, called SELF (*Système d'Evaluation en Langues à visée Formative*). SELF is a large-scale assessment system validated according to the ALTE cycle using both quantitative (Classical Test Theory and Item Response Theory) and qualitative methods (questionnaires, interview, and focus group), and has been developed within the framework of the ANR IDEFI project Innovalangues.(ANR-11-IDFI-0024—*cf.*⟨hal-02004250⟩) Today, SELF has already placed around 140,000 students in six different languages. Designing a multilingual test with these features is a very demanding and long process. The most challenging aspects concern (1) keeping the same communicative construct for the six different languages; (2) improving item writers' skills in psychometrics and, more broadly, spreading high-quality evaluation culture; (3) infrastructural and technical demands; and (4) coordination of a large, heterogeneous team over a long period of time (six years). These difficulties required adoption of specific strategies to reach our goal, e.g., careful organization of the working team composed of a scientific manager, team coordinators, and item writers; the decision to start with two pilot languages, Italian and English, followed by the other four; drafting and sharing common documents to guarantee interlinguistic transfer; in-house design of a multi-task platform serving as an authoring tool, a piloting and pre-testing repository, and a large-scale administration system to track, archive, and disseminate the final results.

C. Cervini (✉)
University of Bologna, Bologna, Italy
e-mail: cristiana.cervini@unibo.it

M. Masperi
Université Grenoble Alpes, Grenoble, France
e-mail: monica.masperi@univ-grenoble-alpes.fr

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_18

18.1 Introduction: Purpose and Testing Context

The IDEFI-ANR Innovalangues project (Masperi 2011) at the *Université Grenoble Alpes* is concerned with research into innovative pedagogical approaches in the field of teaching and learning second languages. Its main objective is to make a significant contribution to the improvement of language teaching and training practices. One of the central axes of the research is the creation, scientific validation, and development of an online formative language assessment system, called SELF (*Système d'Evaluation en Langues à visée Formative*) (Cervini and Jouannaud 2015). SELF is a large-scale assessment system that currently assesses six different languages (English, Italian, Chinese, Japanese, Spanish, and French). It is composed of a set of assessment modules that gauges students' language level based on the Common European Framework of Reference for Languages (CEFR). This specialized system, available to the entire educational community, integrates different functions in the same platform: player, task authoring tool, results manager, and test session organizer.

SELF[1] arises from the recognition (and evidence) at national level in higher education in France (Masperi 2011, p. 8) of the inadequacy of operational solutions for formative assessment. The main shortcomings observed include: (1) the closed (non-dynamic) nature of application software; (2) the lack of transparency in the calibration of items used to assess linguistic ability; (3) the absence of tracking of student work; (4) the summary nature of the information provided without any diagnostic assessment that would allow an effective learning response. The evidence of these shortcomings, which are found in all language teaching across the country, encouraged us to propose the ambitious design of a multilingual system to provide guidance and reliably assess the strengths and weaknesses of French-speaking students and so facilitate and provide an incentive for the creation of groups with similar levels and needs (targeted needs-based training).

18.1.1 SELF Conceptual Foundations

The design of an assessment system like SELF must be based on a wide-ranging consideration of the language and skills model to be proposed. In testing, this consideration means defining the construct of the test. In this respect, the CEFR is an important, if not central, point of reference, but insufficient as a guide to designers in the creation of assessment tasks within a communicative approach that respects the level descriptors. From this point of view, the realization of tasks and items must be duly supported by explicit and rigorous procedures that are not set up a priori, but are developed through constant interaction with the academic discipline, the foundations of which have been laid for many years, and a research-action-development approach that operates in a precise area of application.

[1] SELF—Système d'Evaluation en Langues à visée Formative.

Specifically, SELF is a teaching tool conceived as a hinge for training that aims to place the student unquestionably at the center of the learning processes. The system is based on the need to adopt the same methodological approach for all target languages in terms of structural coherence, content, assessment processes, and visualization of results. An equally fundamental need is that of realizing a technical and pedagogical system that is both flexible and adaptable, while taking into account the needs of all players involved in learning assessment within institutions, both in teaching (researchers, teachers, and students) and in administration.

18.2 Testing Problems Encountered: Communicative Constructs and Standardized Language Tests

The design of SELF is based on a response to a series of key linguistic, pedagogic, and organizational questions. The main challenge in the development of the system was to reconcile our pedagogical aims—designing a valid and reliable multilingual communicative test—with the practical constraints linked to standardization and computer-based assessment. A test can be defined as "communicative" if it conveys *meaningful communication exchanges* in *authentic situations* (Brown 2005). Besides these two key points, a real communicative test should have *unpredictable and/or creative language inputs* and *outputs* where *integrated skills* are simultaneously stimulated, as is the case in real life. The features of unpredictability and creativity are the most difficult to reproduce through self-correcting online tasks. An in-depth definition of the construct and its operationalization can be a valid way to avoid the risk of its under-representation in standardized tests.[2]

Before describing the SELF construct in detail, it is important to highlight other relevant constraints in our design. The construct was supposed to be the same for all six languages used, despite different second language acquisition and consolidated testing traditions, which could significantly differ for non-European languages such as Japanese and Chinese (Higashi et al. 2017) compared to Italian, French, Spanish, and English, the other four languages included in the system. The first aim of the test system is to guarantee valid and reliable placement in a language course but, given its formative nature, SELF should also provide information and guidance for students and teachers.

Another contextual factor concerns the practicality of the test, which is part of the richer and broader concept of *usefulness* of a test. A test is required to be useful (for

[2] "Standardized assessment makes a serious effort to capture crucial aspects of the component abilities of comprehension. Drawing on these assumptions for standardized test construction, […] standardized reading assessment should seek to translate (aspects of) the reading construct into an effective reading test (fluency and reading speed; automaticity and rapid word recognition; search processes; vocabulary knowledge; morphological knowledge; syntactic knowledge; text-structure awareness and discourse organization; main-ideas comprehension; recall of relevant details; inferences about text information; strategic-processing abilities; summarization abilities; synthesis skills; evaluation and critical reading") (Grabe 2009, p. 357).

institutions, for students, for society in general), and to be useful it should satisfy six requirements: validity, reliability, authenticity, interactivity, impact, and practicality (Bachman and Palmer 1996). Practicality within SELF consisted in the design of a durable system for large-scale assessment (more than 140,000 candidates[3] evaluated in around four years), easy and safe to use in an institutional environment. In this specific case, practicality refers not only to the available resources for development and administration (human, economical and organizational), but also—for test candidates—to the reasonable period of time required to complete the test (not more than one hour), considering its low-stakes- context of exploitation.

SELF's communicative constructs are focused on three abilities—listening, reading, and limited production—which means that the principle of interactional or situational authenticity is alternatively based on an oral (just audio, such as a phone call exchange or a radio broadcast, or audio-visual, such as TV news, ads, lessons, etc.) or a written input (e.g., taken from magazines, post-its, newspapers, etc.).

Considering the formative nature of SELF, it is clear that limiting the exploration of language competence to the macro ability does not provide sufficient information for either students or teachers. For this reason, we have expanded some facets of receptive or productive ability through items that we have called "linguistic focalizations" and "cognitive operation(s)." The concept of linguistic focalization partially covers that of a sub-skill, whereas that of "cognitive operation" refers to the process that a candidate is supposed to activate in order to resolve items or to reply to questions. In terms of linguistic focalization, test items concern three main facets of language competence: grammatical knowledge (morphology and syntax), vocabulary (including collocations and idioms), and socio-pragmatic aspects. Due to the multidimensionality of human linguistic expression and of texts, these focalizations often coexist in the same item. In some other specific cases, some items could be more focused on phonetic discrimination, on textuality (coherence and cohesion), or on metalinguistic reflections.

Cognitive operations refer to functions of a subject's cognitive activity, i.e., to the mental processes (understanding, inference) that he/she needs to activate to respond to the item. A cognitive operation also refers to what a candidate is called on to do (complete, interact, correct) with a text that is read/heard. Tracking all these features makes a significant contribution (1) in defining/observing the degree of complexity of the language task and (2) helping to clarify the multidimensional construct of communicative competence. Section 18.5 will describe the technological measures that we have adopted in order to enhance students' centrality in the testing process and to develop the concept of a formative perspective within SELF.

[3] Around 80% of the administrations were in English, whereas the remaining 20% were more or less equally distributed among Spanish, Japanese, French, Italian, and Chinese.

18.3 Solution of the Problem: The Testing Cycle for a Good Culture in Evaluation

When applied to language testing, the concept of validity has evolved in the last decades in the direction of the study and observation of its social impact on all stakeholders (students, teachers, institutions, and society as a whole). Therefore, we have sought to anchor SELF to the best practices in language evaluation, both to afford our system maximum scientific legitimacy and to spread a positive culture in the field of assessment at the University of Grenoble Alpes and within its connected networks. Indeed, validation is not a process to be undertaken on the spur of the moment. It involves a series of different steps which are intertwined and iterative, from quantitative to qualitative and vice versa. For this reason, it is very important to plan validation well in advance, because the organizational effort required is enormous, particularly in the piloting and pre-testing phases.

These objectives have resulted in some necessary operational choices: (1) invest energy, time, and economical resources in acquiring new, specific skills in the field of item writing and psychometrics; (2) increase, through individual and group responsibility and motivation, the team's appreciation of being part of a project with long-term goals to produce a durable system; (3) improve the team's awareness of the risks of subjectivity in language evaluation and, consequently, of its unethical impact on institutions and society.

The main qualitative validation phases at the beginning of the SELF test cycle were (1) content re-reading and peer correction, and (2) think-aloud protocoling to fine-tune the effectiveness of the software interface, whereas at the end of the test cycle, we considered (3) standard setting and post-test qualitative evaluation with teachers and candidates. Generally speaking, "identifying the score which corresponds to achieving a certain level is called standard setting. It inevitably involves subjective judgement, as far as possible based on evidence" (ALTE 2011, p. 44). Different standard setting methodologies exist (focused on learners' corpora, on candidates' performance, on test contents), but for SELF the most adequate was the *bookmark method* (Hsieh 2013), which enabled direct discussion and debate among language teachers regarding features of content (clear and bias-free formulations) and task difficulty based on student competence. The application of the bookmark method for standard setting and post-test analysis encouraged triangulation between intuitive (i.e., assign a level of difficulty to the items during the conception phase), quantitative (large-scale pre-tests and statistical analysis to establish items' psychometric values), and qualitative methods (final validation by experts after reaching a general consensus).

Post-administration analysis was conducted with both students and teachers through questionnaires and interviews. The aim of this analysis was, on the one hand, to discover if students who had been placed in a specific language class on the basis of SELF results felt that they had been placed in the correct group (in terms of proficiency) and, on the other, to assess whether the class group was sufficiently homogeneous, thus making teaching of the class easier for the teachers. In the case of the Italian version of SELF, this qualitative survey proved that the system had a slight tendency to overestimate French students' competence in Italian. This side effect was relatively predictable, because it is a self-corrective test with a strong component based on the evaluation of receptive abilities. This tendency was promptly corrected through two different measures: (1) an increase in the threshold levels for limited production, which was the most discriminating ability in the linguistic combination "Italian for French candidates," and (2) a reduction in the number of reading comprehension items, which proved to be less discriminating than listening and limited production.

Regarding the quantitative methods, it is crucial to remember the fundamental importance of pre-testing both the tasks and the assembled version of the final test. We organized the quantitative validation in two main stages: a *pilot test* on a target corpus of around 50 participants, mainly aimed at improving the quality of content preparation and at providing a first look at item discrimination indexes (we applied the Classical Test Theory through the use of the TiaPlus software), and *pre-testing* on a target group of 250 candidates (this large-scale trial allowed us to apply the Item Response Theory and, in this case, we used the Winstep software). As shown in the ALTE testing cycle, pre-testing was preliminary to item calibration, which, again, occurred before the standard setting phase. Through pre-testing all the items are calibrated and put in order of difficulty but threshold levels have not yet been defined. Therefore, this last step can be accomplished thanks to the new involvement of language teachers or of linguistic experts in the standard setting discussion which is a very interesting process from a cultural and intercultural point of view: teachers and linguistic experts are requested to explicitly uncover and share their vision of language competence with others. Even if competence descriptors are the same for all participants, their interpretation is often very subjective because it reflects individual teaching and learning styles and habits. For this reason, "definition of the threshold scores is probably the part of psychometrics most associated to cultural, political and artistic issues" (Cizek 2011, p. 5).

This very enriching experience revealed the fundamental importance of including direct and indirect users in the test validation cycle, not just to benefit from the natural increase in the social acceptance of the test, but also in order to neutralize bias and other critical issues.

18.3.1 SELF: An in-House Conception of a Multi-Task Platform

SELF is a complete system—player, task authoring tool, results manager, and test session organizer—that is fully operational and designed to respond to specific needs of research and teaching. Depending on user status (student or editor/administrator), SELF presents two different interfaces. Specifically, the SELF interface enables (1) designers and editors to create tasks and items that they can then assemble into

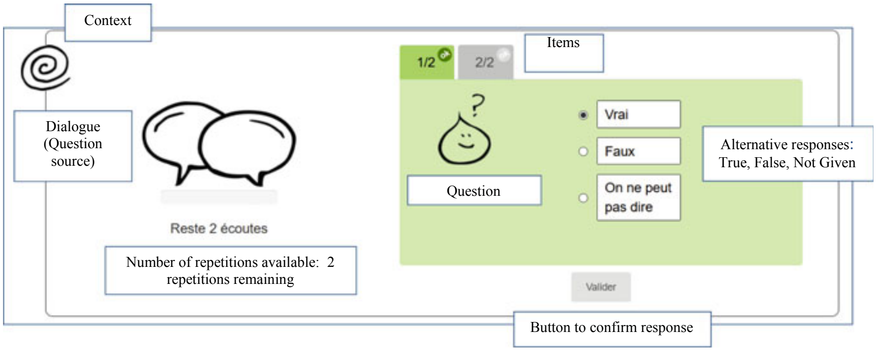

Context
Dialogue (Question source)
Reste 2 écoutes
Number of repetitions available: 2 repetitions remaining
1/2
2/2
Items
Question
● Vrai
○ Faux
○ On ne peut pas dire
Alternative responses: True, False, Not Given
Valider
Button to confirm response

Fig. 18.1 Example of self layout (for an oral comprehension task)

tests, and (2) administrators to manage test sessions and export the results. The different elements that make up a task are shown on the screen always with the same layout and labeling (called the "task grammar"). Shown on the left of the screen (see Fig. 18.1) are (1) the context, (2) the question source (i.e., the input from which the question is formulated), and (3) the number of repetitions (for listening comprehension questions). On the right-hand side, there are: (1) one or more items set out in sequence in tabs, (2) the question, (3) the possible responses, and (4) the button to confirm a response. The combination of all these elements, called "task grammar," has the same features for the three abilities (listening, reading, limited production).

One of the main strengths of the software is the considerable flexibility in integrating different types of resources. All the fields (context, question source, etc.) are *all media compatible*, so they can accept any audio, visual, image, or text source. Regarding task conception, the editor has access to different types of exercise depending on the ability to be tested and the objective of the task.[4] The system flexibility is also linked to the independency of the different item banks. Each test refers to a specific language item bank composed of the validated items, but all six banks (English, Italian, Chinese, Japanese, Spanish, and French) are conceived and technically structured in the same way. This feature of the SELF system assures that each language team can easily work in autonomy. In addition, the authoring tool allows editors (1) to integrate ad hoc feedback, (2) to retrospectively add to the set of possible responses to a construct "short written expressions" taken from students' responses that are correct but were not initially envisaged by the editor. Finally, tracking of individual and group activity is a powerful added value for researchers and teachers. Even if, at present, only a small part of the information collected with SELF can be exploited to provide feedback to students and teachers, the tracking system used reflects this methodological approach and its relevance for diagnostics and training.

[4] It should be noted that the variety of protocols, which was initially sought to make the test more attractive and enhance user attention, was substantially reduced following the first pilots. A smaller number of standard exercise types did in fact give greater control over results.

The system that tracks and manages results can generate files with different features and objectives. Alongside management and statistical files (the former for administrators involved in organizing groups, and the latter for analysis with statistical software), there are full export files providing a broad range of information relevant to language teaching and learning. This includes some biographic data (e.g., first language(s) and other reference languages in addition to the first language for each candidate), the time required to complete the full test and the time spent on each task before confirming the response, and the level of perceived difficulty (again for each task). Regarding these last two points, correlations in the data collected could give rise to more wide-ranging and highly informative considerations. For example, we could compare the actual difficulty of an item (expressed by the psychometric index of difficulty) with candidates' perceived difficulty and the response (correct or incorrect) given to the question.

At the same time, it is important to note that the diagnostic and training approach of SELF is supported by a further tracking tool that we have designed and developed, and entitled *identity card*. The identity card associated with every item and every task tracks essential linguistic and didactic information regarding the characteristics of the written and spoken texts, the specific qualities of individual items, and the psychometric indices. All these factors may influence both task complexity and the strictly individual relationship created between a candidate and the task presented. This is the way in which such a meticulous tracking system can open the door to studies of the diagnostic and training perspective of SELF.

18.4 Insights Gained: Looking Back at Process and Choice

The development of a multilingual assessment system founded on a common methodological and didactic framework for use by adult French-speakers was a challenge determined by key, local factors. Today, the widespread dissemination of SELF in academic institutions in France is proof of the need for the tool. However, the results obtained have never been taken for granted. In our opinion, the large-scale adoption of this training tool is based on four joint factors: (1) the quality and stability of the staff involved in the process, (2) the rigorous documentation of the process, (3) the thorough quality control, and (4) the intrinsic nature of the product itself.

First, the work assigned to the designers and editors was of a high professional level (Cervini 2014). From a technical point of view, exceptional linguistic competence must be accompanied by an excellent command of procedures that require a specific training background. However, the process must also include a creative component both in identifying sources and in creating original texts. The role of the *performant item writer* therefore combined a rather uncommon dose of perfectionism and inventiveness. Finally, the collaboration with programmers also had been mediated through a technical and pedagogic professional who defines the profile of the technological, IT, and ergonomic specifications.

The second essential aspect was the development and provision of valid and substantial working tools: a reference bibliography, clear and exhaustive methodological and didactic guidelines (regarding text creation, editing of items, and psychometric analyses), interlingual glossaries and didactic memoranda (keywords and definitions, types of protocol, question banks), and clearly stated procedures regarding the activities related to the preparation of tasks (studio recordings and use of authentic resources).

In line with the methodology adopted, the third element—quality control of the SELF project—played a role for the six target languages at two process levels: during the creation of the tests and when they were delivered. The measures adopted were of three types: (1) the methodological support for the researcher and editor team provided by international experts in the sector[5]; (2) the product maturation envisaged by the testing cycle and undertaken following the required stages of validation and psychometric analyses; and (3) the compilation of questionnaires during piloting, as well as the ex-post use of qualitative research protocols applied to results collected from the students tested. Moreover, the service offered to the universities using SELF is shown to be appreciated in annual ad hoc questionnaires.

Finally, the question of the transferability of innovative teaching practice varies in function of the nature of the product itself. SELF fills an evident gap in the field of learning assessment of which we were fully aware. Moreover, we assume that broad, consensual adoption of the system might be determined by the fact that SELF is a finished, "turn-key" product that is non-invasive and not in competition with other solutions. SELF might act as a lever to define a university's language policy, but it leaves institutions maximum freedom to decide on the use of their own teaching tools (communicative, action, and thematic approaches; classroom, blended, and holistic systems, CLIL), and the way in which these tools interact with the assessment system.

18.5 Conclusion: Implications for Test Users

SELF is a multilingual assessment system with a formative and diagnostic aim. It is available in six languages at state higher education institutions in France and is used to quickly assess a student's level in three skill areas. SELF is designed to respond to institutional needs to guide students toward a training path that is suitable to their linguistic profile and thus to facilitate the adequate development of existing expertise. The strengths of the SELF system can be summarized as follows:

Academic Solidity and Interlinguistic Coherence. SELF is based on design procedures and academic validation that are rigorous, from both a quantitative

[5] In the first stages of development (2013–2015), research methodology was based on suggestions and training provided by CIEP. More recently (2016–2018), the project has enjoyed the expert support of James Purpura (Columbia University) and CITO (Department of Psychometrics and Research), The Netherlands.

(piloting and psychometric analyses) and qualitative perspective (analysis of references, cross-checked revisions between item writers, standard setting, post-assembly piloting). The task banks are designed and produced following a common methodology for all six languages and aim to guarantee customized teaching. Each task has an “identity card” to categorize both the tasks and the items, indicating the linguistic and pragmatic focus and so serving as a precursor to the diagnostic framework.

Conscientization of Prior Learning and of Perceived Difficulties in a Formative Perspective. In testing each chosen ability, SELF seeks, in spite of the objective limitations of automatic feedback, to propose an assessment context that is as close as possible to an authentic communicative situation. In this respect, we have chosen to assess oral comprehension with a fully oral-based approach without any written support, and to include in the bank of possible responses to written expression questions, any correct responses given by students that were not contemplated by the item editor (Cervini 2016). The formative dimension is further supported by the decision to present results in the form of a “recommended learning path” (e.g., *en route vers…*).

Multifunctionality of the Underlying Technical Structure. SELF offers flexible and efficient editing and delivery that can adapt to the needs of different institutions (division of students into groups by academic year, by discipline, or by department) and interface easily with training paths set by institutional policy. The system’s IT platform, which is currently experimental and could be extensively enhanced, is already able to serve a large number of simultaneous accesses (approximately 500). The technical set up is also designed to serve research (editing tools, analysis and categorization of tasks, archiving of data on user actions and results, tracking of item behavior, etc.) and to respond to developments suggested by the data collected for each language.

The design of a system such as SELF must always be considered as a work in progress and subject to continual improvement, not only as far as the obvious need to update content is concerned, but also regarding verification of the usefulness (validity and reliability) of the test for a body of candidates in continual evolution. Experience has shown that psychometric results from the quantitative assessment can be enriched and complemented with qualitative information collected through interviews, focus groups, and questionnaires.

The methodological framework that has been established for the development of SELF’s diagnostic and formative perspective is specifically based on modeling this information to the benefit of language students (self-awareness, motivation, customized learning paths) and language teaching staff—teachers and tutors—who can more easily design remedial work appropriate for student needs.

References

ALTE. (2011). *Manuel pour l'élaboration et la passation de tests et d'examens de langue.* Division des Politiques Linguistiques. Strasbourg: Conseil de l'Europe, DG II—Service de l'éducation.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Brown, J. D. (2005). *Testing in language programs.* New York, NY: McGraw-Hill.

Cervini, C. (2014). La valutazione multilingue nel contesto dei dispositivi formativi: Il sistema 'SELF' per il posizionamento e la diagnosi delle competenze linguistiche. *LEND—Lingua e Nuova Didattica. Periodico di Linguistica Applicata e Glottodidattica, 1*, 16–26.

Cervini, C. (2016). Approcci integrati nel testing linguistico: Esperienze di progettazione e validazione in prospettiva interlinguistica. In C. Cervini (Ed.), *Interdisciplinarità e apprendimento linguistico nei nuovi contesti formativi. L'apprendente di lingue tra tradizione e innovazione*. Bologna: Quaderni del CESLIC. http://amsacta.unibo.it/5069/1/Volume%2520CeSLiC.pdf. Retrieved February 18, 2019.

Cervini, C., & Jouannaud, M. P. (2015). Ouvertures et tensions liées à la conception d'un système d'évaluation numérique multilingue en ligne dans une perspective communicative et actionnelle. *ALSIC—Apprentissage des langues et systèmes d'information et de communication. Numéro spécial 'Des machines et des langues', Alsic, 18*, 2. http://alsic.revues.org/2821. Retrieved February 18, 2019.

Cizek, G. J. (Ed.). (2011). *Setting performance standards: Foundations, methods, and innovations.* New York, NY: Routledge.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Higashi, T., Shirota, C., & Nagata, M. (2017). *Developing a Japanese language test for a multilingual online assessment system: Towards an action-oriented approach to Japanese instruction in Europe* (pp. 236–245). Alte 6th International Conference Proceedings, Bologna.

Hsieh, M. (2013). Comparing yes/no Angoff and bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly, 10*(3), 331–350.

Masperi, M. (2011). *Innovalangues: Innovation et transformation des pratiques de l'enseignement-apprentissage des langues dans l'enseignement supérieur*. MESRI. ANR. Investissements d'avenir. https://hal.archives-ouvertes.fr/hal-02004250. Retrieved February 18, 2019.

Chapter 19
The Relationship Between English Placement Assessments and an Institution: From Challenge to Innovation for an Intensive English Program in the USA

Nicholas Santavicca

Abstract The chapter discusses the institutional processes, assessment development, and English language proficiency in relation to university expectations of international students. Balancing the relationship between best practices for achievement for English skills, campus academic programs, and a pathway provider is a whole-institution initiative. Higher education decision-making utilizing assessments is subject to controversy, since they are at risk of operating unfairly for students expecting uniform assessment treatment and institutions expecting uniform indications of linguistic readiness. The chapter highlights issues emerging from practices identified from the past five founding years of a university-based ESL program in the USA and international pathway provider. The issues highlighted include: (1) stakeholder relations, (2) student language skills, (3) assessment development, and (4) testing innovations. Rather than a cure-all for the complexities and maladies, the chapter presents details of assessment design and implementation for dealing with the challenges that emerge, in order for other institutions to develop deeper insights into their own language testing relationships that in turn determin*e* student trajectories, institutional connections, and missions of programs.

19.1 Introduction: Purpose and Testing Context

Intensive English Programs (IEPs) inside universities face different assessment challenges than those of their academic and administrative colleagues from other disciplines. IEP assessments differ from other disciplines in multiple ways. First, students take intensive English either as a form of skills training or a pathway into degree programs. Students do not graduate with a major or minor specialization, and they often study English full-time and exclusively. In addition, IEP programs in postsecondary institutions exist, at least in part, to generate revenue. As such, they often have

N. Santavicca (✉)
University of Massachusetts Dartmouth, Dartmouth, MA, USA
e-mail: nsantavicca@umassd.edu

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_19

separate tuition structures, admissions policies, and budgets (Norris 2016; Richards 2017), meaning that assessment approaches within IEP programs are evaluated from a different set of standards than others on a postsecondary campus.

In this specific assessment context, there are three main functional programming units to consider: (1) The IEP provides coursework for international students to attain a level of English proficiency in lieu of having an official TOEFL or IELTS score upon admission. For this context, the IEP is known as the American Language and Cultures Institute (ALCI). (2) The outside pathway provider recruits and places international students into the IEP and other university programs. Providers typically contract with a university via a corporate model to recruit and admit students to share revenue. The outside pathway provider has the purview to place students into academic courses without clear student English language proficiency levels assessed and/or vetted directly by the ALCI. (3) University coursework is attained when a student has reached level 5 (high-advanced) in the IEP. Upon achieving this level of academic English proficiency, students will begin their university career by enrolling in university courses at the 100 level, for example, History 101, Math 101.

Many institutions like the specific university context described above, unknowingly have adopted an English language proficiency policy that reifies languages as static, bounded, and evaluated according to a narrow canon of rules, and it also reifies social identities in terms not of language use but of nationality (Banjong 2015; Schlaman 2019). This ideology supports a limited view of linguistic and social communication in which the ideal speaker is thought to be a monolingual native speaker of a social variety of English. Many times over, Basic Interpersonal Communicative Skills (BICS) (Cummins 1981) are evaluated by English language assessment for higher education decisions. BICS and CALPS are the names given to two broad registers of language, Basic Interpersonal Communicative Skills (BICS) and Cognitive Academic Language Proficiency (CALP), in the 1970s by Canadian educator Jim Cummins (1981). BICS language is sometimes called social language or even survival language, takes only three to five years to develop (Cummins 1981). BICS can be acquired informally, at least in part, through social interactions or in social media. Before educators understood the difference between BICS and CALP language, many students were exited from language programs before they were ready.

Social language discourse practices (BICS) in relation to university course readiness had a firm hold on the evaluation of English proficiency readiness from an international recruiting and marketing perspective led by the university context described. This specific campus context contributed to the IEP having little control regarding the students being recruited and placed into English and academic courses where (CALPS), academic discourse is prevalent. The English proficiency assessment issues could be resolved, if the outside pathway provider was thoroughly probed by campus leaders for a better understanding of the provider's English proficiency testing practices and academic quality standards of students recruited. For their part, in order to rectify the problem of misplaced students, the IEP faculty needed to innovate assessment and testing measures to ensure academic readiness for the international students learning on campus that were placed by the outside provider into ALCI coursework and other academic courses without proper language testing.

The IEP faculty grounded their innovations for language assessment in Assessment for Learning Principles and Design (Assessment Reform Group 2002) to aid in supporting the IEP's mission to support language learning through best practices in second language acquisition and reach all English levels of the international student population.

An example of a BICS assessment (without CALPS) was provided from our University International Recruitment Office. The office liaison had described an official campus correspondence where many of the students in the new cohort of international students were described to be English proficient due to the fact that the liaison had spoken to them personally on the phone. This anecdote stresses the necessity for a clearer and more comprehensive understanding of BICS campus wide. Here is the assessment challenge: Students were considered academically prepared for the study of university-level content on the basis of BICS-types of assessments sanctioned by the university and the outside pathway provider. Examples of student assessments reviewed consist of phone calls, email correspondences, and the ability to complete an application for university admission. The assessments occurred without language testing expertise or collaboration with English proficiency experts and researchers on campus. Thus, many students began to flounder, for their social English skills could not support their academic endeavors.

Developing learners' communicative competence is a large part of the ALCI program. The five-level program is based on aspects of communicative competence, including linguistic, strategic, discourse, and sociolinguistic areas. The program and placement system for coursework demonstrates the abilities of students to master all four aspects of communicative competence to create a skillful language user. The program is comprised of five levels from beginner to high advanced. Once a student places into level 5 in the IEP/ALCI, a student is considered to have reached the level of English proficiency required for university coursework. Students take 20 hours of face-to-face coursework, and students are sequenced into courses that specialize in specific content areas, cultural events, reading, writing, listening, speaking, and grammar.

Recruited international students should be placed into the IEP based on grade point average (GPA), high school coursework, and English proficiency assessments (among other factors) set by the university. However, the students on campus are placed via admissions and the outside pathway provider based on provider business practices, GPA, and individual student English capabilities set by the provider, typically outside the realm of academic rigor.

19.2 Testing Problem Encountered

The BICS's effect on the assessment of multilingual students in university IEPs, particularly creates a financial burden for students and risk of academic probation that results from a student's delaying fulfillment of the first-year English requirement for three semesters. However, a financial gain for the international pathway provider and the university is created. On one side, the university would gain from the assessed placement into the lowest level of the three-course ESL sequence, for it requires a university-sponsored pathway for the English proficiency program to garner more revenue and support students' English proficiency due to the longer duration of time spent in the program by students. On the other side, the outside pathway provider markets and promises quicker entry into the university course work not based on English performance. Thus, English assessment and subsequent placement are often overlooked or considered not a priority. The course sequence consists of a basic introduction to writing for international students, a course to support academic presentations, and a Freshmen Year English writing course designed for an international population.

The responsibility of the IEP for making decisions regarding international student language skills has proved to be complex, especially when multilingual learners are seen to need support services from the university like tutoring or extended class times. For example, the IEP can offer individual assessments for students, occasionally circumventing the ineffective placement process developed via staff on campus. Such forms of language support structures, helpful though they may be and easy to accomplish given this autonomy, allow non-native English-speaking students to be placed outside of the normal university curriculum, rather than to be supported across the curriculum by means of an inclusive placement structure for English assessment. Thus, unintentionally, the IEP constructed its position as an unofficial campus gatekeeper that may help English language learners navigate outside the structure of the university. A greater limitation for the university develops where students work around the curriculum and shelter their linguistic skills from view of the academic community.

Unfortunately, the IEP program is seen as peripheral, non-academically dense, and expendable. This view of the IEP and English proficiency was constructed in part due to the strong influence held on the campus by the outside pathway provider contracted by the university. The outside provider's business-minded goals do not coincide with student support for English because students are promised a quick entry into university programming based on GPA, educational background, and high school courses. English skill is not stressed or seen as essential for student success by the provider. The outside provider was established on campus before the IEP, and the provider molded many of the academic and English policies and beliefs in existence.

19.2.1 Pathway Programming Challenge

Outside pathway programming or the "bridge program" for recruitment of international students has had a profound influence on the university and its ability to assess and support an international student population. The suggestion here is that the social language ideology for university readiness (BICS) has such a firm hold on the practices of the campus community that the IEP has little control over the students recruited and placed into English and other academic courses. This situation might be resolved if the outside pathway provider were vetted for English proficiency practices and academic quality of students recruited for the campus. Both the IEP and the pathway provider work with faculty across disciplines to effect changes regarding international populations and how the relationships between campus and international community are viewed. However, the pathway provider position is housed outside of the university structure and chain of command, while the IEP is housed in the English department. The pathway provider is for profit and earns funds from the students; the IEP program is part of the larger non-profit side of the university. The rest of the university, then, is symbolically absolved of responsibility for educating multilingual populations. The university has unintentionally constructed the IEP's position as the place where the English of international students is policed and debated on campus. Many "problem English students" are sent to the IEP, and the IEP is provided with limited knowledge of pathway recruited students' English level or academic background.

19.2.2 Assessment NEED

This assessment challenge developed out of a necessity to support students that were evaluated outside of the expertise of ESL professionals and the utilization of BICS assessments to support the stance of academic readiness. Currently, all of the ALCI student population share the same schedule of courses. However, each individual student is provided with an individualized path of language study due to the varied levels of English proficiency of recruited students admitted. A shared multilevel classroom creates a language learning environment equivalent to a "one-room school house." Inside the walls of this "contemporary one-room school house," the instructors and pioneers of inclusive placement-intensive English instruction found a need for a single multilevel assessment using the same source and/or material (e.g., TV episode) to support all of the students within the four walls. The assessments support the varied linguistic levels of the students in the program. The assessments include beginning and end-of-term assessments, and other formative/summative assessments to enhance instructor knowledge of students' language skills. The IEP has innovated a new means of testing to alleviate some of the stress of multiple levels of proficiency in one class, for both students, faculty, and the campus.

The student population hails from all parts of the globe. In the past three years, the IEP has hosted students from India, Pakistan, Jordan, Iraq, China, Japan, Colombia, and Vietnam. The international student population age range is from 18 to 25. The initial English proficiency for students recruited for the IEP is high-beginner/novice based on ACTFL proficiency levels. The program is designed for students to spend no more than three semesters in the intensive English path of study.

19.3 Solution to the Problem

19.3.1 Adoption of AFL

AFL assessments were implemented by the IEP as a multilevel solution to the problem of students being misplaced into academic course by the outside pathway provider. Designing the multilevel assessments for English proficiency is founded in Assessment for Learning Principles (AFL) (Lee 2011; Lee 2017; Lee and Coniam 2013). AFL is an educational framework built around 10 principles seeking to assess students in a way that creates awareness of their current skills and knowledge gaps, that provides the ability to map future learning and goals. In 2002, the Assessment Reform Group released 10 principles to consider when incorporating AFL in the classroom:

- Is part of effective planning
- Focuses on how students learn
- Is central to classroom practice
- Is a key professional skill
- Is sensitive and constructive
- Fosters motivation
- Promotes understanding of goals and criteria
- Helps learners know how to improve
- Develops the capacity for self-assessment
- Recognizes all educational achievement

The IEP's implementation of this assessment idea is founded upon authentic language use and extending the concept of BICS to CALPS. Walking around the campus, IEP instructors frequently heard students discussing events from their favorite TV shows or movies. This observation led the program to employ TV shows and movies as frequent topics of conversations due to the fact that in all cultures, television creates an authentic language learning medium. Therefore, one show/genre was selected for the entire term, and an episode was shown each class.

AFL enhances learning in the classroom by treating assessments as a process where learners display their knowledge and skills and then analyze their responses to map out future learning (William 2011). Therefore, it is not just the students participating in the assessment, but also the instructors. Instructors, in tandem with their students, analyze the assessment results and decide where learners are in their learning, where they need to go, and how best to get there (Assessment Reform Group 2002). As shown above, AFL design principles are complex and cannot be realized in isolation; instructor/student collaboration is key to identify and account for interrelationships between teaching, learning, and evaluation. This process, when applied appropriately, is crucial in developing students' confidence and motivation for language and culture acquisition, for both summative and formative testing situations. In the end, AFL, for the purposes of this testing selection, illustrates the pivotal role assessment plays in reinforcing and extending learning and learner autonomy in language learning settings (Dann 2014; Lee 2017).

These AFL-backgrounded assessments serve as placement tests for the beginning/end of term. However, to prevent students from being stressed about their performance (key to AFL), students are evaluated according to AFL principles on the entire learning process. Therefore, if the students perform to their best ability and complete all testing sections, test performance will not negatively impact grades. AFL emphasizes motivation without negatively impacting students who progress at a slower rate. For example, students are graded on their classroom participation, two-three smaller assessments, and their classwork as a whole. These newly introduced AFL-based tests function as a final assessment that highlights the students' skills and serves as a placement test for the following term. The tests only have a negative impact on the students' grade if they put in little to no effort (scoring lower than their current level). The tests place significant emphasis on the writing process: students' knowledge of the ability to revise work using resources. AFL stresses the importance of a continuous feedback loop between instructor/student to foster oral and written academic work. The students must demonstrate that they are capable of both skills before being placed at the university-level coursework and performing with native-English-speaking peers. Students are given the level of their performance following the test while the information and process are still fresh in their minds. Immediate feedback allows the students to ask more specific questions about their performance and plan ways to move forward effectively.

19.3.2 Assessment Descriptors

Using AFL tests for varied proficiency levels, instructors show episodes of an American sitcom (a situation-comedy show from television) for students to review throughout the semester/term. The sitcom functions as a focal point in and out of the classroom for activities, content area focus, and assessment. A thirty-minute episode of a sitcom serves as a basis for student-generated and accessible knowledge during a class for evaluation. The consistent and familiar scaffolded content and contact with specific characters, social situations, accents, cultural phenomena, etc., provide more equity and balance in the classroom for introduction to knowledge and skill sets.

In regard to content or material, the IEP faculty chose American sitcoms because they are generally 30 min and provide 2–4 storylines each episode. This structure allows for multiple examples and activities to be taken from the show based on each storyline. Additionally, as students watch more of the show, students complete language-specific assignments focusing on season-long plot lines (especially the more advanced students). Students are asked to perform lesson or test tasks immediately following the episode, to practice their ability to intake new information and material and then reflect on it in speech or writing, as they would in an academic course.

Assessment focuses on four key aspects:

- A familiar TV show and set of characters.
- Parts 1–2: closed-book answers on the material/plot. The length and complexity of these parts are based on the level of the student. For example, the high beginners are given content-specific questions and only required to answer in complete sentences. The low-intermediate students are asked to perform the task of writing a summary, using the writing process.
- Part 3: open-book revision. Students are given the chance to check over their work and make corrections in colored pens.
- Reflection. Students are asked to identify what they did well, what they struggled with, and what sources they used (ranking the sources for helpfulness on a scale of 1 to 10).

Test Protocol:

- Explain vocabulary words for the chosen episode (included on test)—10 min;
- Watch episode (or short clip if needed for time constraints)—30 min;
- Complete Parts 1 and 2 with closed books and notes—1–1.5 h;
- Using colored pens, complete Part 3: revision with open resources—30 min–1 h;
- Complete a self-reflection questionnaire and turn in—5 min.

19.4 Insights Gained

The multilevel AFL assessments were designed to enhance English proficiency evaluation, discover curriculum improvements, and find the knowledge gaps of the international student population, in addition to assigning accurate level placement. Traditionally structured tests (using test item formats such as cloze or fill-in-the-blank) were not giving accurate representations of students' abilities to produce and understand English in a university setting, which led the program to introduce more open-ended and performance-based test items in placement exams. The tests presented here were adapted from a series of classroom activities that received high levels of interest from students. The assessments produced increased student participation and production of spontaneous English (both written and spoken). The tests led to increased peer dialogue, in-class discussion, and analysis of the TV show. Furthermore, instructors could summarize how well students understood what they were watching and hearing, and synthesizing information from recent episodes, when speaking or writing in class. Language learners deal with multiple complexities during the assessment process of coding and decoding messages from the classroom to the sitcom. Even for native speakers, the process of forming thoughts and ideas and expressing them coherently through language is not a simple endeavor. The assessment presented supports students' "strategic competence" to employ a number of strategies to communicate in and out of the classroom. Moreover, this assessment process focuses on competence strategies that have traditionally received little attention in language learning settings, and serve a more pervasive role in and out of the

Table 19.1 Assessment content review guide

Review category	Parts of test	Aspects of BICS & CALPS	Rationale
Material Comprehension	Parts 1 & 2	• Main plot points • Sub-plots • Themes • Time sequencing • Storytelling • Description • Summary • Analysis • Inferences	• The ability to follow main and minor plot points is a useful benchmark in comprehension • The ability to summarize and sequence events shows a good understanding of storytelling tactics and events • Inferences, analysis, and, themes show advanced understanding of the topic
English Skills (Unrevised)	Parts 1 & 2	• Sentence structure and variety • Appropriate and varied verb tense • Vocabulary • Word form	• Sentence structure and verb tense variety allow students to give detailed information in more concise and efficient ways • Using new terms, academic vocabulary, and the ability to adapt word forms demonstrate understanding of the appropriate discourse and terminology
Revision Skills	Part 3	• Ability to find errors in work (using guide) • Ability to correct the errors • Ability to use and navigate various resources according to need	• Revision is an essential skill for Academic English. Students need to become accustomed to checking over all their work and develop familiarity with the multiple sources available to them

classroom. Students acquire strategic competency strategies through AFL testing that include: confirmation checks, avoidance, and commands. The strategies are meant to be thought of as fluid and spontaneous parts of a student's language acquisition capabilities and use (Ellis 1997; Lee 2017).

As a final insight, incorporating the entire writing process (outline, write, and revise) into the three parts of the test helped the students realize the effectiveness of the AFL on the quality of their written work. Instructors review each assessment designed for content, using our innovative guide that is focused on three main categories: Material Comprehension, Unrevised English Skills, and Revision Skills. See Table 19.1 for a breakdown of each category:

The AFL multilevel assessments follow the same procedures used in class activities to maintain a comfortable situation for the students, and emphasize key understanding of the requirements of the language tasks. The assessment ranks each category on a 5-level basis to reflect the ALCI/IEP structure of courses. A passing level would be considered the current level or above. If a student scores below their current level, then a one-on-one meeting will address whether the low score is due to misunderstanding or lack of attention on the student's part. This process fosters student motivation (per AFL) by allowing the student to focus on content and production rather than grades (Lee 2017). By comparing respective student performance to the level expectations (both their current level and the exit level), students are able to see improvements and gaps of these particular skills in a meaningful and constructive manner (Lee 2011). Additionally, by scoring the tests according to level placement rather than a fixed score, instructors help students remain focused on the overall goal of graduating from the IEP and building their English skills.

19.5 Conclusion: Implications for Test Users

The IEP and AFL assessment model shown by the assessments presented here, provides a clearer examination of students' English proficiency to perform BICS and CALPS successfully by the students enrolled in our campus IEP. This approach creates much-needed transparency for the process of evaluating students' language levels and readiness for university coursework evidenced by the student context studied. Most importantly, the AFL assessment created supports authentic language use, mimicking the real-life study skills that students need for academic achievement in university-level courses (Lee 2011; Lee 2017). The possibility of further research and inquiry exists to investigate the AFL assessment model with a myriad of different contexts, student populations, language proficiencies, and instructional practices.

AFL is a holistic process and is not achieved by individual educators, university staff, outside programs, and a campus working in isolation. Instead, it is paramount that everyone involved in this AFL process, and international programming, work collaboratively to review curriculum and plan a comprehensive program that takes into account the interrelationships between teaching, learning, and assessment for international student language support. The campus can then develop strategies to support BICS and CALPS with all stakeholders involved. To implement AFL, a campus needs to define and communicate goals and expectations clearly to international students, provide them with opportunities to engage in language learning rather than reduce them to passive examinees, and prompt them to take responsibility for learning. AFL should be considered a key professional skill for instructors in Intensive English Programs, and a consideration for continuing professional development for internationalizing a campus through language.

References

Assessment Reform Group. (2002). *Assessment for learning: 10 principles*. Port Melbourne, VIC: Cambridge University Press.

Banjong, D. N. (2015). International students' enhanced academic performance: Effects of campus resources. *Journal of International Students, 5*(1), 132–142.

Cummins, J. (1981). Empirical and theoretical underpinnings of bilingual education. *Journal of Education, 163*(1), 16–29.

Dann, R. (2014). Assessment as learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice, 21*(2), 149–166.

Ellis, R. (1997). *SLA research and language teaching*. Oxford: Oxford University Press.

Lee, I. (2011). Bringing innovation to EFL writing through a focus on assessment for learning. *Innovation in Language Learning and Teaching, 5*(1), 19–33.

Lee, I. (2017). Assessment for learning in the L2 writing classroom. *Classroom writing assessment and feedback in L2 school contexts* (pp. 105–122). Singapore: Springer.

Lee, I., & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing, 22*(1), 34–50.

Norris, J. M. (2016). Language program evaluation. *The Modern Language Journal, 100*(S1), 169–189. https://doi.org/10.1111/modl.12307.

Richards, J. C. (2017). *Curriculum development in language teaching*. Cambridge, UK: Cambridge University Press.

Schlaman, H. (2019). Designing structures and pathways to support language development and content learning for English learners: Dilemmas facing school leaders. *International Multilingual Research Journal, 13*(1), 32–50. https://doi.org/10.1080/19313152.2018.1531675.

William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3–14.

Chapter 20
Placement Decisions in Private Language Schools in Iran

Kioumars Razavipour and Tahereh Firoozi

Abstract Despite being highly frequent, placement decisions and how they are made do not frequently feature in the language testing literature. The current study surveyed 30 language institutes in Iran on their placement testing policies and practices. Particularly, we probed into content areas tested, test taker characteristics considered, institutional issues in connection with placement decisions, test users and issues of power in assigning language learners to course levels. Descriptive statistics showed that oral skills and the sub-skills of grammar and vocabulary are often tested for placement purposes. Reading, writing, and translation are tested with less frequency. In addition, apparently construct-irrelevant variables like gender and age were found to moderate placement decision making. It was also found that institutional interests seem to affect decision making about assigning students to course levels. Moreover, it was revealed that in making placement decisions, the power almost exclusively is in the hands of the institutions, and stakeholders have little, if any, influence in the process. Finally, though participants claimed otherwise, expertise in language testing and assessment seems to be lacking on the part of those in charge of making placement decisions. Findings carry implications for placement decision making in Iran and in other EFL/ESL contexts.

20.1 Introduction: Purpose and Testing Context

With the exception of the realist school of thought in validity theory (Borsboom and Mellenbergh 2007), in current thinking about validity, validation is about the uses and consequences of a test and not about a test per se (Bachman 2005; Bachman and Palmer 2010; Messick 1989). Despite this, the use of tests for making placement decisions in numerous language programs across the globe has not attracted proportionate research attention (Hudson and Clark 2008; James and Templeman 2009; Plakans and Burke 2013). One possible reason for the Cinderella state of placement testing might be that placement decisions are deemed to be low-stakes

K. Razavipour (✉) · T. Firoozi
Shahid Chamran University of Ahvaz, Ahvaz, Iran
e-mail: razavipur57@gmail.com

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_20

because unlike, say, high-stakes selection decisions, placement decisions made are not often irreversible (Green 2013). This reasoning, however, might not always be true because first, there is little evidence that once placement decisions are made, they are reversible in all contexts. Secondly, learners' language education pathways are likely to take dramatic turns in the wake of wrong placement decisions. Many a learner has given up language learning for being wrongly placed in language classes and many teachers may face challenges because of misplaced language learners in classrooms. Parents of misplaced learners are also charged both emotionally and financially for wrong placement decisions. Placement tests are therefore of substantial consequence for a range of stakeholders. It seems that the neglect of placement testing might be part of the collective neglect of the test taker, lamented decades ago by Brown (1981 as cited in Brown 2008):

> In this plethora of theoretical activity, the student himself often seems to be forgotten. . . . There just seems to be a general lack of follow-up on what actually happens to a student after we have affected his life by placing him at one level or another in our ESL classes. (p. 276)

Writing about the use of writing tests for placement, Crusan (2002) states that assessing writing is an "act laden with pedagogical, ethical, political, psychometric, and financial implications" (p. 18). Given the consequences that placement decisions have for various groups of stakeholders such as students, teachers, language programs, and parents, studies into how placement decisions are made are warranted (Plakans and Burke 2013). In addition, the sheer number of placement decisions that are made adds to the significance of studies on placement testing. With the rise of English as the language of international trade and communication and the increasing demand for English across the globe, millions of students seek to improve their English language skills. Every one of these students should be placed in a language course, which shows how highly frequent placement decisions are.

The existing literature on placement testing is mainly about the validity of placement tests (Alderson et al. 1995; Johnson and Riazi 2015, 2017; Kokhan 2012, 2013). Far less attention has been given to how final placement decisions are made. Plakans and Burke's study is perhaps the exception in this regard. Using a grounded theory approach, Plakans and Burke (2013) studied how international students at an American university are placed into four levels of language ability. They found that in making placement decisions, test scores constitute only one factor. Test taker factors, test user factors, and programmatic factors were found to interact in the process of placement decision making. Though Plakans and Burke (2013) provide a rich picture of how students are placed in one tertiary institution, it remains to be seen how placement decisions are made across diverse contexts. In particular, to improve generalizability, studies addressing placement decision making across a larger number of institutions are needed. The present study is aimed at narrowing the noted gap. Within this spirit, this chapter addresses the following research questions.

Research question 1: What language skills or sub-skills are tested for placement purposes?

Research question 2: How far do micro-political considerations affect placement decision making?

Research question 3: Which learner variables other than language proficiency are considered in making placement decisions?
Research question 4: Who makes the placement decisions?
Research question 5: How assessment literate are those who make placement decisions?

In Iran, the fever for English language learning is so high that in some cases the onset of English language learning even predates the learner's birthdate. We have seen middle class expectant mothers with no knowledge of English who, despite the complications associated with pregnancy, would sit idly among a bunch of kids in an English class in the hope that mere exposure to English during pregnancy would help the unborn to pick up some English before they formally make it to life on planet earth. This shows the true scope of hunger for English in the Iranian society, where English language schools have not spared even the villages. In some small cities, the number of private language schools is comparable to that of grocery stores of the city.

This appetite for English brings millions of language learners to thousands of language schools across the country annually. Considering the fact that learners arrive at the schools with different backgrounds in English, they have to be placed in classes of the right level. Thus, placement decisions have to be made about millions of learners across thousands of language institutes around the country.

20.2 Testing Problem Encountered

A number of constraints limit the options available to private language schools for proper placement testing. For one thing, the logistics and resources available bear on the magnitude of evidence that can be collected for placement purposes. Whereas big institutes can afford to make placement decisions within age groups because of their large pool of applicants, this is not feasible in small scale language schools. In small private schools, the small number of language learners renders the act of placement testing futile because even if they spread learners across proficiency levels based on scores from some testing procedure, there are not corresponding classes at the right level in the school for each proficiency level diagnosed in placement testing. For the noted reasons, the context of making placement decisions in local language schools is far messier than where placement decisions can be made based on scores from the administration of a language test. As a result, in such contexts many seemingly construct-irrelevant variables are likely to be involved in the decision-making process. Therefore, the act of assigning students to classes in many language institutes might be only partly about language proficiency. It is a social act influenced by various aspects of human interaction discourse, as well as by individual and institutional interests. The fact that there is no official body regulating or overseeing the functioning of local language schools in Iran further renders the situation of research interest.

20.3 Review of Literature

Depending on the sense we make of the term *place*, placement tests are of two types (Alderson et al. 1995; Brown 2005; Green and Weir 2004). When we intend to create groups of learners of roughly the same ability, a general proficiency or an aptitude test is used. The use of external tests for placing incoming students at levels of language ability is in fact quite widespread (Kokhan 2012). Research findings about the effectiveness of external tests for making placement decisions are mixed. Adopting a grounded theory approach, Fox (2004) found evidence for the effectiveness of an external EAP test for placing students across English ability levels. Yet the study, as is the case with grounded theory studies, was conducted within a single academic institution, and Fox warned against generalizing the findings. Similarly, Wang et al. (2008) argued for the validity of TOEFL iBT for placement purposes.

Some remotely construct-relevant variables are often used in making placement decisions about students. "Seat time (i.e., amount of classroom exposure to the language)" is one method of making placement decisions, which has been shown in some cases not to be satisfactory "since seat time is not a particularly good indicator of ability—given the great disparity in secondary school Russian programs, teachers, and students" (Larson and Murray 2000, p. 49).

Focusing on a web-based Spanish placement test in the context of tertiary education, Long et al. (2018) investigated, in addition to the internal consistency and content validity of the test, the validity of placement decisions made based on the test. It was found that placement decisions made were consistent, and the test showed high internal consistency. The authors concluded that the test was valid for its intended placement uses.

Designing and administering placement tests are known to be resource intensive (Harrington 2018) given that they should frequently be administered to a limited number of participants. Therefore, some studies have investigated the potential of short and easy to administer tools such as lexical facility tests (Clark and Ishida 2005; Harrington and Carey 2009; Harsch and Hartig 2015; Lam 2010). In a couple of studies in two language schools in Australia and Singapore, Harrington and Carey (2009) found that the tests enjoyed good correlations with in-house placement tests, though not equally informative because of the limited construct Yes/No which lexical facility tests tapped. Yet another strategy for making placement testing practical is the development of computerized adaptive placement exams (CAPEs). For instance, Brigham Young University has developed a suite of CAPEs in a number of languages including Spanish, Russian, German, and French (Larson and Murray 2000).

Another line of research in placement testing concerns involving the learner in the process. This is often accomplished via self-assessment. Such assessments have a practicality advantage in that they reduce the costs associated with placement. In the context of placement in academic settings, however, scholars have warned against the use of self-assessments because of the diverse backgrounds which candidates come from (Wall et al. 1994). Others have recommended the use of self-assessments in combination with other measures. Comparing scores of applicants on a standardized test, an in-house test, and a self-assessment survey, Ferris et al. (2017) found a legitimate role for self-assessment. Yet, they cautioned that placing large numbers of students of multilingual backgrounds might not be effectively done using only self-assessments. Ruecker (2011) compared the attitudes of resident, domestic ESL, and international students toward their placement and found that placement is about labeling students, which affects and shapes their identities: "Students do pay attention to the way they are labeled and placed" (p. 108).

Another major challenge in placement testing is determining the cutoff points. One strategy to tackle this challenge is to have students already studying at different levels in an institution take the test to be used for placement purposes, and decide on cutoff points in light of the average scores of students at each level, as used in the development of CAPEs at Brigham Young University. This is in fact a case of predictive validity, where a newly developed test is judged based on its agreement with a measure that is already in operation. The problem is that in doing so, the validity of the old test is taken for granted. The circularity of the predictive validity argument is what led scholars in educational and psychological measurement to look for other validation procedures (Newton and Shaw 2014).

To decide on the levels, some scholars see a role for learners' corpora. They suggest that it is possible to identify levels of proficiency by deriving and categorizing errors in students' corpora. For instance, Taylor and Baker (2008) maintain that "Learner corpora can help identify typical errors at a given proficiency level which can inform the focus of test items or tasks for a particular test-taker population as well as test preparation publications" (p. 246).

Regarding test format, indirect, multiple choice tests are frequently used in making placement decisions (Crusan 2002). Despite sufficient evidence suggesting that discrete point grammar items overestimate the productive skills of EFL and ESL learners, grammar sections constitute a constant component on such tests.

In developing a placement test of reading, Green (2013) notes that test purpose, theoretical consideration, and practical limitations must be taken into account. This is of course true of all test design programs; yet, the latter practical constraints might take on more importance in placement testing because of the numerous factors that must be taken into account and the limited budget and resources that are allocated to the design and maintenance of placement tests.

Based on this brief review, it seems that the body of research on placement testing is mainly about the validity of different language proficiency tests used for placement. There is less research on whether and the extent to which test and non-test factors inform placement decisions. Perhaps the one exception is Plakans and Burke (2013), who found that in addition to test scores, other variables such as student factors, program factors, and test user factors were taken into consideration in making placement decisions. Regarding students, factors such as affect, diligence, and motivation featured in the process of decision making. The knowledge and experience of test users also influenced their interpretations of scores, assessments, and all other considerations in placement decisions. Test users' "knowledge about the tests, test takers, classes, levels, and outcomes impacted the use of tests and the decisions made in the test use process" (p. 126). Programmatic factors most frequently referred to in the placement process were class size and the range of language abilities within each level. In several cases, initial placement decisions were changed to create classes of equal sizes. Other programmatic issues featuring in the process were teachers, teaching materials and the curriculum. Yet, Plakans and Burke's (2013) study is rather limited in scope, given that it addressed placement decisions in only one institution of higher education. In addition, the body of existing research on placement is mostly about college placement for international students applying to universities in English-speaking countries. There is little, if any, research on how placement decisions are made in numerous language schools where English is learned as a foreign language. The present study contributes to narrowing this gap.

20.4 Methodology

20.4.1 Participants

Thirty-three participants involved in making placement decisions across 33 language institutes in Iran contributed data to this study. Three participants were managers who were interviewed for generating an item pool for the design of the questionnaire we used in this study. Two interviewees held PhD degrees in TEFL and one had a PhD in linguistics. They were selected for interviews mainly because they were accessible to the lead author of this chapter.

The remaining 30 participants who responded to the questionnaire were working in language schools across the country at the time of this study. Specifically, they were working in eight different provinces of the country, indicating that the sample represented roughly one-third of the country's provinces. That said, two-thirds of the participants were from Khouzestan and Fars, two Southern provinces in Iran. Table 20.1 provides further details about the participants and the language institutes where they were working at the time of data collection. Of the 30 language agencies surveyed, 24 admit both male and female students, while six are either exclusively for male or for female students. The agencies surveyed ranged widely regarding their years of being active, from one year to more than 20 years.

Participants were of diverse levels of educational backgrounds from high school graduates to PhDs. However, the majority held either a bachelor's degree or a master's degree. Educational qualifications of 27 participants were in fields related to English language while three participants had studied in fields other than English.

Respondents were not required to provide information about the name or the size of their agencies. As such, we did not ask about the number of students enrolled or placed in each agency, which might limit the inferences that can be made of the findings, for such considerations do bear on placement decisions (Plakans and Burke 2013). Our rationale for not asking for such details was that these private language institutes are politically vulnerable in comparison with agencies affiliated with the

Table 20.1 Demographic information of the participants

Type of school		Frequency	%
	Female	3	10
	Male	3	10
	Male-female	24	80
Years of experience	1–2 years	1	3.3
	3–5 years	11	56.7
	6–10 years	4	13.2
	11–15 years	7	23.4
	16–20 years	5	19.9
	More than 20 years	1	3.3
Participants' level of education	Diploma	3	10
	Bachelor's degree	8	26.7
	Master's degree	14	46.7
	PhD	5	16.7
Participants' field of education	relevant to English language	27	90
	irrelevant to English language	3	10
Provinces where participants worked	Alborz	1	
	Bushehr	3	
	Fars	9	
	Khouzestan	11	
	Lorestan	1	
	Khorasan Jonoubi	2	
	Tehran	2	
	Zanjan	1	

government. Therefore, we made a conscious effort to provide the safety and security they needed to answer the questionnaire with no reservations.

20.4.2 Instrumentation and Data Collection

As the main instrument of data collection in the current study, a five-point Likert type questionnaire was developed by the authors. To do so, we drew on existing literature as well as on the interviews conducted with three people involved in making placement decisions in three different institutes. That said, Plakans and Burke's (2013) work was adopted as the theoretical model based on which the questionnaire was designed. In this model, four major factors inform placement decision making, namely test scores, student factors, program factors, and test user factors. Based on interviews, we added another factor to the model and named it the "power" factor. The questionnaire was

Table 20.2 Content structure of the placement questionnaire

	Factors involved in placement decisions	Number of items
1	Test factor	9
2	Student factor	7
3	Program factor	6
4	Test user factor	4
5	Power factor	8
6	Demographic information	5
Total		39

in Farsi, the official language of the country. Table 20.2 summarizes the content of the questionnaire.

Once the questionnaire was prepared, an online version of it was developed, and the link was sent out to 37 managers of private language institutes. Thirty participants returned the questionnaires.

As can be seen from Table 20.2, the final format of the questionnaire consisted of 39 items, 33 of which tapped into the five noted factors. The remaining five items were intended to gather demographic information about the participants. The congruence between items and the related factors was judged by another language testing expert. Given the rather small sample size, we did not examine the psychometric properties of the instrument, which must be considered in interpretation of the findings. To analyze the data, descriptive statistics such as frequency, percentage, mean, and standard deviation were used. Data analysis was conducted using SPSS software version 21.

20.5 Findings

20.5.1 Test Considerations in Making Placement Decisions

As noted earlier, the first cluster of items was about test issues that are involved in making placement decisions. The first item in this cluster was about language skills or sub-skills that are tested in making placement decisions (research question 1).

Table 20.3 indicates that for placement purposes, oral skills are twice as frequently tested as written skills of reading and writing. It also tells us that in the majority of language schools, the sub-skills of grammar and vocabulary are frequently tested. Tests of translation are also used in at least one-third of the surveyed language schools.

The remaining items in Table 20.4 were about the nature and content of placement tests used in each language school.

It appears that items from standardized language tests do not frequently feature in tests used for placement decisions, as only eight participants reportedly make use

Table 20.3 Nature and components of placement tests

		Frequency	%
Tests components	Speaking and listening	30	100
	Reading and writing	18	60
	Grammar and vocabulary	27	90
	Translation	11	26.7

Table 20.4 Test issues involved in placement decision making

		Always	Usually	Often	Rarely	Never
1	We administer our own in-house test for placing students at levels	4	8	4	7	7
2	We are sure that we make the right placement decisions	0	10	10	10	0
3	To make placement decisions, we borrow items from standardized tests such as TOEFL, IELTS, Konkour, etc	2	6	9	10	3
		Every semester	Annually	biannually	Every five years	Never
4	How often do you update your placement tests?	5	12	5	4	4
		Completely	To a large extent	To some extent	Slightly	Not at all
5	Is your placement test aligned with the content of textbooks taught in your language institute?	11	9	5	4	1

of items from such tests (see responses to item 7). On the other hand, in two-thirds of language schools, the tests used for placement decision making are aligned with the content of the textbooks used there. Regarding the updates they make of their placement testing practices, in five language schools the placement procedure is reportedly updated every semester. In twelve institutes, it is done annually and in five schools, updates are made biannually. Four schools do so every five years and in four institutes, the placement procedures are never updated.

20.5.2 Learners' Characteristics in Placement Decisions

The next group of items on the questionnaire was about characteristics of learners that might inform placement decisions (research question 3). These items were essentially about whether and the extent to which test takers' characteristics like language learning experiences, age, gender, and level of education bear on placement decisions made in language schools surveyed. Participants' responses to these items are given in Table 20.5.

With regard to whether they reexamine a language learner who has already attended the same language school for some time, the participants were divided. Almost half would go for assessing such learners for placement and the other half would not do so. Likewise, the learners' age seems to be an important consideration in making placement decisions, as only seven participants would reportedly rarely or never take age into account in deciding about placing learners at levels. Concerning learners' attitudes toward placement decisions, 25 managers agreed that younger language learners are more satisfied with the way they are placed. Finally, regarding gender, most managers agreed that female students more readily accept placement decisions.

Table 20.5 Test taker issues in placement decision making

	Item		Agree	Neutral	Disagree	
1	Those who have studied in our language institute before are subject to replacement		12	3	15	
2	Younger students are more satisfied with the outcome of placement decision making		14	11	5	
3	Compared to boys, female students are more satisfied with our placement decisions		11	11	8	
		Always	Usually	Often	Rarely	Never
5	Students' age is an important consideration in making placement decisions	7	6	10	5	2
6	Students' level of education is an important consideration in placement decision making	7	6	7	6	4
7	When placing students, we ask them about the number of years they have attended other language schools	13	9	4	3	1

20.5.3 Institutional Considerations in Placing Students

The next cluster of items was about institutional constraints and considerations bearing on placement decisions (research question 2).

The first couple of items in Table 20.6 asked the participants about the contingency of placement decisions upon enrollment rates and the number of course levels offered in their language schools. Surprisingly, most participants, nearly two-thirds, denied that such considerations affect their placement decision making.

The next four items in this cluster were about the extent institutional interests are involved in placement decisions. Generally speaking, responses to these items indicate that financial interests do play a role in making placement decisions. Twenty-two managers were of the idea that they would always, usually, or occasionally have courses for all proficiency levels. Similarly, two-thirds reported that they would never or rarely tell a learner "we do not have a course at your level of English proficiency." Likewise, nearly two-thirds of the managers surveyed agreed that they would place learners at least one level below their actual proficiency level, and an overwhelming majority were opposed to placing learners above their actual proficiency levels.

Table 20.6 Institutional considerations in making placement decisions

	Item		*Agree*	*Neutral*	*Disagree*	
1	Our placement decisions depend on enrollment rates		8	5	17	
2	Our placement decisions depend on the number of classes we have in our institute in a given semester		9	5	16	
		Always	*Usually*	*Occasionally*	*Rarely*	*Never*
3	We have classes for all proficiency levels in our language institute	4	7	11	7	1
4	We may tell a language learner "we do not have a course at your level of English proficiency"	1	6	3	17	3
5	To be on the side of caution, we place students one level below their actual proficiency level	2	1	15	9	3
6	To be on the side of caution, we place students one level above their actual proficiency level	0	1	2	17	10

20.5.4 Power Issues in Making Placement Decisions

Items in this category asked language school managers about the extent to which stakeholders, particularly learners, are involved in making placement decisions (research question 2) (see Table 20.7). Overall, responses in this cluster of items speak to a seemingly imbalanced power distribution between the institutions and the stakeholders. Nearly two-thirds of the participants were opposed to the involvement of learners in decision making about placement. Similarly, an overwhelming majority, 26 participants, did not approve of using self-assessment in placing students. Along the same lines, most participants would not allow participants who are unhappy with their placement decisions to choose the course level they deem fit to their language abilities. Further, 23 participants would rarely or never allow learners to choose their

Table 20.7 Power relations in placement decisions

	Item		*Agree*	*Neutral*	*Disagree*	
1	Language learners should have a role in placement decision making		10	3	17	
2	Language learners can object to our placement decisions		16	8	6	
3	We use students' self-assessments in making placement decisions		3	1	26	
4	Students' dissatisfaction with our initial placement decisions may lead to changing the placement decision		12	9	9	
		Always	*Usually*	*Occasionally*	*Rarely*	*Never*
5	If a student objects to the results of our placement procedure, we will reassess his/her proficiency	9	10	2	8	1
6	If a student is not happy with the placement decision, he/she can choose the course right for his/her level	3	2	2	13	10
7	Learners' parents play a role in making placement decisions	0	0	2	14	14
8	What percentage of initial placement decisions are altered and new decisions are made?	9	1	2	13	5

desired course levels, in case they feel that they have been wrongly placed. Likewise, almost all the participants believed that parents should have no role in assigning learners to course levels (item 7).

On the other hand, responses to the remaining items in this category were not as one-sided in favor of institutional power. For instance, in response to item 2, almost half of the participants agreed that learners are entitled to object to the placement decision made about them. Similarly, twelve respondents agreed that they would alter placement decisions if learners are not happy with the level they are assigned to. Likewise, two-thirds of the managers would be reportedly willing to reconsider the placement decisions made should the affected students demand so (item 5).

20.5.5 User Considerations in Placement Decision Making

The other important consideration in making placement decisions pertains to characteristics of users of placement instruments and processes (research questions 4 and 5).

More than half of the participants claimed that the person in charge of making placement decisions was an expert in language testing and assessment, and in the majority of the language schools surveyed, teachers were reportedly happy with the outcome of the placement process (Table 20.8). That said, when participants were asked about who made the placement decisions, 21 managers reported that the school managers themselves make the decisions. In 23 cases, there were reportedly certain English teachers who would make the placement decisions. Eight participants reported that placement decisions are made by any teacher available at the time when some placements should be made. Similar results were reported about analyzing placement test data. That is, in the majority of cases, placement decisions are made by the manager or by some teachers who are assigned to the role.

20.6 Insights Gained

This study investigated several aspects of placement decision making in English language schools in Iran. More specifically, it explored issues related to tests, test takers, test users, institutions, and power in connection to placement decision making.

In regard to content and skill areas tested, it was found that speaking and listening were more frequently the subject of assessment for placement purposes. Grammar and vocabulary were also tested with a relatively high frequency. This was consistent with previous research (Crusan 2002). Translation tests were also somehow common as placement procedures. Overall, these findings were expected, for in language schools, it is often not feasible to administer placement tests in a uniform, standardized manner because learners refer to language schools at their own convenience. In such circumstances, where placement decisions are distributed throughout

Table 20.8 User considerations in placement decision making

	Item		*Agree*	*Neutral*	*Disagree*	
1	The person in charge of making placement decisions in our agency is an expert in language testing and assessment		15	8	7	
		Always	*Usually*	*Occasionally*	*Rarely*	*Never*
2	Our teachers are satisfied with our placement decisions	4	14	9	3	0
		School manager	*Certain teachers*	*Any teacher available*	*The secretary*	
3	Who is in charge of making placement decisions?	21	23	7	0	
4	Who analyzes the results of placement testing?	23	23	8	1	

a semester, doing a short interview or giving a short list of vocabulary items to the learner seems more practical than administering a test of reading comprehension or asking learners to write a paragraph. Another expected finding was that in most cases, placement tests were aligned with the content of coursebooks used in the language institute.

In addition, we found that more than half of the language schools surveyed make use of items borrowed from standardized tests for placing students across levels (see Table 20.4). Whether placement decisions made based on such test scores are appropriate awaits further investigation. Yet, there is evidence that placement decisions made based on standardized test scores are of dubious validity (Kokhan 2013). Concerning the instruments used for placement, findings from this study diverge from those of similar studies like that of Kahn et al. (1994), in which most agencies reported using commercially available tests as at least part of their placement testing policies. Perhaps this underuse of commercial tests has to do with wider socio-economic issues, like the value of the country's currency compared with the dollar. For this reason, institutes cannot afford to purchase commercial language tests. In addition, there are no private national companies specializing in the design and provision of tests, which in turn might have to do with the fact that the socio-economic infrastructure for the flourishing of the testing industry is lacking in Iran and in the wider Middle East (Gebril 2016; Oakland 2009). Findings also suggest that the challenge of making placement decisions varies across proficiency levels. In addition, though school managers reported awareness as to the uncertainty inherent in placement testing, the majority appeared certain about the appropriateness of placement decisions made in their own institute.

Among language learners' characteristics, age and gender appeared to be more prominent variables in making placement decisions (see Table 20.5) as most participants agreed that young learners and female students were easier to place. Language schools varied in the importance they accorded to other learner factors such as level of education and their past language learning experiences. From a purely psychometric perspective, such learner variables introduce construct-irrelevant variance to the measurement process. Nevertheless, this observation can be explained by considering the wider socio-cultural considerations that further complicate placement decisions. In the high distance culture of Iran (Beeman 1986), many learners may prefer age homogeneity to proficiency homogeneity for the very uneasiness they may feel in a class where classmates are of different age groups. This demonstrates that validity of decisions made based on assessments hinges upon larger social values (Messick 1989). Another cultural aspect making placement decisions difficult is the compulsory gender segregation that has to be observed in most, if not all, language schools of the country. For this reason, language schools are not allowed to have mixed-gender classes. This cultural mandate adds to the complexity of placement decisions because it would cut back on the number of students who would have otherwise likely been placed at the same level.

Concerning the role of institutional considerations in making placement decisions, our findings seem to suggest that institutional interests affect how learners are assigned to course levels. This echoes the role of micro-political issues in the English language education industry (Alderson 2009). Most participants reportedly would accept learners across all proficiency levels, believing that they would always have courses appropriate for diverse language ability levels. Perhaps one reason for such thinking among the participants is that they equate an institute's capacity to address all proficiency levels with having and teaching textbooks labelled beginner, intermediate, advanced, etc. Whether language schools have qualified teachers and the right facilities to accommodate the needs of diverse proficiency levels remains unclear. The participants all reported that they would place learners below their real language competence levels. Though in the absence of further evidence, it is better to suspend judgment about their true motives for doing so, our own experience with language schools convinces us that in many cases such practice is for keeping learners in one's institute for a larger number of semesters to simply make more profit.

Regarding stakeholders' involvement in making placement decisions, it was found that though the majority of institute managers believed that learners were entitled to object to the placement decision, they did not believe that learners should be involved in the decision-making process. Nor did they make use of self-assessments in placements. Similarly, parents were almost never allowed to get involved in the process. Overall, such attitudes on the part of managers appear to be at odds with the ideals of democratic assessment and critical language testing (Shohamy 2001).

On the other hand, the use of self-reports and self-assessments has been found not to be proper data to build on for making placement decisions in tertiary education institutions admitting international students (Wall et al. 1994). Therefore, once again there seems to be a conflict between the psychometric and the ethical dimensions of language testing.

Finally, regarding the users of placement assessments, it appears that in the majority of cases, results from placement assessments are either analyzed by the language school managers or by some teachers chosen for doing so. Similarly, final placement decisions are also made either by managers themselves or by a select number of teachers. The majority of managers claimed that those who make the placement decisions are experts in language testing and assessment. This is in conflict with previous studies on the assessment literacy of teachers across the globe and in Iran (Oakland 2009; Popham 2009; Razavipour and Rezagah 2018; Riazi and Razavipour 2011). Possibly, participants' understanding of what constitutes expertise in language testing and assessment is different from what is considered language assessment literacy by the language testing community (Davies 2008; Fulcher 2012; Inbar-Lourie 2008).

20.7 Conclusion: Implications for Test Users

We end this chapter with a few recommendations for the improvement of placement decision making in what may wrongly be considered low-stakes situations. First, as a means to meritocracy (Fulcher 2015), sound testing and assessment can only take place under the right social, political, and cultural circumstances. One such requirement is that people should see themselves as agents of social change and justice, not subjects at the mercy of powerful states and institutions. In other words, citizens should make institutions accountable for their actions. In this regard, raising stakeholders' awareness about the hidden agendas and interests of global and local forces involved in English language teaching and testing can contribute to fairer language placement decisions.

Another social condition that must be present for sound language assessment practices is for the society to have the right infrastructure to allow for the flourishing of the testing industry (Oakland 2009). Such infrastructure demands that a sufficient number of people with the necessary expertise in testing and assessment be available (Oakland 2009). We believe that none of the noted two conditions are satisfied in the context where this study was conducted. In fact, in a culture where people do not make institutions accountable for transparency and justification for their actions, institutions do not feel the need to seek people with the right expertise in testing and assessment. While making wider social changes takes decades at least, in the short run enhancing the assessment literacy of language school managers, teachers, parents, and learners must be given priority by national and international language testing bodies. We believe that the current psychometric view of language testing that dominates the language testing programs at teacher training centers in Iran offers few insights when it comes to making placement decisions, which are, as we found, only partly about language proficiency. Instead, effect-driven testing (Fulcher 2010), which begins with a consideration of testing context and purpose, is more likely to cultivate the assessment literacy required in making placement decisions.

Finally, with regard to language learning and teaching in private language schools in Iran, there are no model standards, like those in other places such as in California (see Kahn et al. 1994). As a result, there is a chaotic situation in private agencies in regard to the number of levels, the language used in describing levels, content of each level, and how they are specified. Given the huge number of language institutes in the context of this study and elsewhere around the globe, defining and implementing model standards at national, state, or city level would help regulate the placement process. It would also help foster transparency and justice in making placement decisions, which would in turn improve the quality of language education.

References

Alderson, J. C. (2009). Setting the scene. In J. C. Alderson (Ed.), *The politics of language education individuals and institutions* (pp. 8–44). Bristol: Multilingual Matters.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Beeman, W. O. (1986). *Language, status, and power in Iran.* Bloomington, IN: Indiana University Press.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. Leighton (Ed.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–115). Cambridge: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessement*. New York: McGraw-Hill.

Brown, J. D. (2008). Testing-context analysis: Assessment is just another part of language curriculum development. *Language Assessment Quarterly, 5*(4), 275–312.

Clark, M. K., & Ishida, S. (2005). Vocabulary knowledge differences between placed and promoted EAP students. *Journal of English for Academic Purposes, 4*(3), 225–238.

Crusan, D. (2002). An assessment of ESL writing placement assessment. *Assessing Writing, 8*(1), 17–30.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327–347.

Ferris, D. R., Evans, K., & Kurzer, K. (2017). Placement of multilingual writers: Is there a role for student voices? *Assessing Writing, 32*, 1–11.

Fox, J. (2004). Test decisions over time: Tracking validity. *Language Testing, 21*(4), 437–465.

Fulcher, G. (2010). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. *Advances in Research on Language Acquisition and Teaching: Selected Papers*, 15–26.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. London: Routledge.

Gebril, A. (2016). Educational assessment in Muslim countries. In G. T. L. Brown & L. G. Harris (Eds.), *Handbook of human factors and social conditions of assessment* (pp. 420–435). New York: Routledge.

Green, A. (2013). *Exploring language assessment and testing: Language in action.* New York: Routledge.

Green, A., & Weir, C. J. (2004). Can placement tests inform instructional decisions? *Language Testing, 21*(4), 467–494.

Harrington, M. (2018). *Lexical facility: Size, recognition speed and consistency as dimensions of second language vocabulary knowledge.* London: Palgrave.

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System, 37*(4), 614–626.

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly, 12*(4), 333–362.

Hudson, T., & Clark, M. (2008). *Case studies in foreign language placement: Practices and possibilities.* Honolulu, Hawaii: National Foreign Language Resource Center.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402.

James, C., & Templeman, E. (2009). A case for faculty involvement in EAP placement testing. *TESL Canada Journal, 26*(2), 82–99.

Johnson, R. C., & Riazi, A. M. (2015). Accuplacer companion in a foreign language context: An argument-based validation of both test score meaning and impact. *Papers in Language Testing and Assessment, 4*(1), 31–58.

Johnson, R. C., & Riazi, A. M. (2017). Validation of a locally created and rated writing test used for placement in a higher education EFL program. *Assessing Writing, 32*, 85–104.

Kahn, A. B., Butler, F. A., Weigle, S. C., & Sato, E. Y. (1994). *Adult ESL placement procedures in California: A summary of survey results: Adult ESL Assessment Project*. Sacramento, CA: California State Department of Education.

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing, 29*(2), 291–308.

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a second language (ESL) courses. *Language Testing, 30*(4), 467–489.

Lam, Y. (2010). Lexical decision tests for foreign language placement at the post-secondary level. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquee, 13*(2), 54–72.

Larson, J., & Murray, M. R. (2000). R-CAPE. *IALLT Journal of Language Learning Technologies, 32*(1), 49–58.

Long, A. Y., Shin, S.-Y., Geeslin, K., & Willis, E. W. (2018). Does the test work? Evaluating a web-based language placement test. *Language Learning & Technology, 22*(1), 137–156.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*. New York: Macmillan.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage.

Oakland, T. (2009). How universal are test development and use. In E. L. Grigorenko (Ed.), *Assessment of abilities and competencies in an era of globalization* (pp. 1–40). New York: Springer.

Plakans, L., & Burke, M. (2013). The decision-making process in language program placement: Test and nontest factors interacting in context. *Language Assessment Quarterly, 10*(2), 115–134.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*(1), 4–11.

Razavipour, K., & Rezagah, K. (2018). Language assessment in the new English curriculum in Iran: Managerial, institutional, and professional barriers. *Language Testing in Asia, 8*(1), 1–18.

Riazi, A. M., & Razavipour, K. (2011). (In) Agency of EFL teachers under the negative backwash effect of centralized tests. *International Journal of Language Studies, 5*(2), 123–142.

Ruecker, T. (2011). Improving the placement of L2 writers: The students' perspective. *WPA: Writing Program Administration-Journal of the Council of Writing Program Administrators, 35*(1), 91–117.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391. https://doi.org/10.1177/026553220101800404.

Taylor, L., & Barker, F. (2008). Using corpora for language assessment. *Encyclopedia of Language and Education*, 2377–2390.

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11*(3), 321–344.

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C. A. Chapelle,*M.* K. Enright, & J. M. Jamieson (eds). *Building a validity argument for the Test of English as a Foreign Language* (pp. 259–318). New York: Routledge.

Chapter 21
Perceptions of (Un)Successful PET Results at a Private University in Mexico

Luis Alejandro Figueroa and Krisztina Zimányi

Abstract Set on the small-town campus of a private university in Mexico, this chapter presents the difficulties faced by high school students regarding institutional expectations when taking a selected exam in English. The problem arises when the students, often enrolled in B2-C2 classes at the PrepaTec in Zacatecas, do not obtain B2, or Pass with Distinction, on the Cambridge PET exam, the selected graduation requirement, which is also a prerequisite for the students to benefit from learning another foreign language. Working within the qualitative paradigm, using an exploratory case study design, and applying autobiographical notes and a focus group interview as data collection techniques, the current study analyzed 20 students' self-reported experience of the exam, with the aim of examining the factors that contribute to the students' achieving, or not, the institutional goal. The findings suggest that there seems to be little variation concerning the students' motivation, preparation strategies, or performance while taking the test. However, some differences were observed as only students with a B1 level certification recounted difficulties in the listening and speaking sections, something which may be rooted in in-class preparation practices. As every single point is crucial for the test takers and, under the current requirements, there is practically no margin for error, institutional policies should perhaps be reevaluated in selecting a test that considers not only reliability, content, construct, and face validity, but also *SEM* and consequential validity.

21.1 Introduction: Purpose and Testing Context

Language testing as an object of study has attracted the attention of researchers for over a century. However, most of the literature published over this significant time span has been dedicated to the analysis and assessment of test content and construct, with a specific interest in validity-related issues. Obviously, this perspective has called for a more numerical approach, with little consideration of the more

L. A. Figueroa (✉) · K. Zimányi
Universidad de Guanajuato, Guanajuato, Mexico
e-mail: alexluis0@yahoo.com

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_21

humanistic aspects surrounding testing. More recently, research on test takers' experiences has gained momentum, with the emergence of studies on exam-related anxiety (Alshahrani 2016; İpek 2016; Li 2015) or lexical choice in exams (Laufer and Goldstein 2004; Mizumoto and Takeuchi 2009; Rodríguez and Sadowki 2002). The study presented in this chapter follows in the footsteps of these contemporary contributions to the field and aims to examine to what extent the selected English language exam fulfills both the institution's and the test takers' expectations at a private university in Mexico.

The data selected for this chapter was collected as part of a larger study carried out at the Zacatecas Campus of the *Instituto Tecnológico y de Estudios Superiores de Monterrey* (ITESM or *Tecnológico de Monterrey* for short), specifically, the high school section, also known as *PrepaTec*, in the city of Zacatecas, Mexico. This private Mexican institution has 31 campuses around the country, and offers different educational opportunities depending on the campus, from middle school to postgraduate studies. At national level, the institutional goal is for the high school students to graduate with a Common European Framework of Reference (CEFR) B2 level in English. Students who achieve this early on in their high school years have the added incentive that they will be allowed to enroll in a third language (French or German) as their general language class. While obtaining B2 is the expectation across the different campuses, each can define local requirements and choose the exam. The PrepaTec at the Zacatecas Campus selected the *Pass with Distinction* result in the Cambridge English: Preliminary (PET) exam as the benchmark for their students. In the following section, the implications of this institutional decision are problematized.

21.2 Testing Problem Encountered

The current research arose from the concern that, contrary to expectations, less than a quarter of the students achieve the institutional and individual goal of attaining the B2 and reach a B1 level instead on their first attempt. Although some manage to obtain a B2 on the second or the third try, this may be too little, too late for them to be able to benefit from the additional languages offered by PrepaTec. This seems all the more disconcerting given the fact that, according to the school curriculum and the institution's classification, these students are supposedly enrolled in classes at B2 or even C2 level, which gives them a false impression of their competences. Such incongruence often results in their surprise when they are unable to achieve a B2 level on the standardized exam. In order to better understand the root of this apparent inconsistency and find possible solutions to the problems these students face, before moving on to the presentation and analysis of the results, an overview of the relevant concepts seems opportune.

21.3 Review of the Literature

Apart from a description of the PET, with special reference to the *Pass with Distinction* rating, the grade required by the PrepaTec in Zacatecas, the most salient test-related concepts include validity, reliability, and Standard Error Measurement (*SEM*). For the purposes of the argument presented here, they will be discussed in this particular order owing to their relevance in terms of the testing problems described in the previous section and the findings presented further below. The next section provides a detailed description of the PET in order to establish how the information it provides is essential to the analysis of the data.

21.3.1 Cambridge English: Preliminary

PET's official name is Cambridge English: Preliminary, while the acronym stands for Preliminary English Test. Part of the University of Cambridge General English and for Schools Cambridge Assessment English program, PET is the second of the five main Cambridge proficiency exams. As a proficiency test, it does not correlate to a course but, rather, prioritizes the candidate's mastery of the level (in this case CEFR B1).

The PET exam's various features mean it can be classified according to different testing categories. First, it is a direct test because it focuses on skills and does not indirectly assess subskills (Hughes 2003). Second, PET possesses both discrete-point and integrative properties since, on the one hand, it is designed to isolate aspects of language for the oral sections (speaking and listening), and, on the other hand, it treats the written sections (reading and writing) as interrelated (Gutiérrez 2017). Third, it is a criterion-reference test as it emphasizes the individual's score, and not the results in relation to the rest of the applicants (Hughes 2003). Fourth, it also has both objective and subjective test properties because the listening and reading sections have limited, specific answers (objective), while the speaking and writing sections are more subjective as the answers vary (Shaban 2014). Finally, an institution can use it as a placement test because it divides the results into five categories: below A2; A2; pass (B1-); Pass with Merit (B1 +); and Pass with Distinction (B2).

The difference between B1 (a Pass or Pass with Merit) and B2 (Distinction), a key indicator for the current study, is explained in the complementary material for teachers.

> Distinction: Cambridge English Scale scores of 160–170. Candidates sometimes show ability beyond Level B1. If a candidate achieves a Distinction in their exam, they will receive the Preliminary English Test certificate stating that they demonstrated ability at Level B2.
>
> Pass and Pass with Merit: Cambridge English Scale scores of 140–159. If a candidate achieves a Pass or Pass with Merit in their exam, they will receive the Preliminary English Test certificate at Level B1. (Cambridge English Language Assessment 2016, p. 5)

Thus, the question presented in the problematization section remains: Why can't students, enrolled in B2 and higher level classes at the PrepaTec, obtain a B2 (Pass with Distinction) level on the PET exam? The answer may be related to the concept of validity and its subcomponents, as the following subsections explain.

21.3.2 Validity

Bachman (1990) defines validity in assessment as the degree of how appropriate and meaningful the test results are. Moskal and Leydens (2000) consider it as "the process of accumulating evidence that supports the appropriateness of the inferences that are made of student responses for specified assessment uses" (p. 71). In addition, Choi (2008) explains that the validity of a language proficiency test is measured by predicting the test taker's performance in real-life situations. However, this practice tends not to be feasible, and validity is analyzed under different criteria.

In general, the Cambridge Assessment English (2018c) official website explains that the PET exam encompasses different types of validity, the most important of which are *content*, *construct*, *empirical*, *face*, and *consequential*.

- Hughes (2003) states that "a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc., with which it is meant to be concerned" (p. 26). A problem regarding content validity is that most real-life situations are not linked to language levels, while the CEFR provides a standard base of how a language learner is expected to react in certain situations. The PET exam is considered to be based on real scenarios that correspond to the levels defined in the CEFR.
- The term *construct* denotes "any underlying ability (or trait) that is hypothesized in theory language ability" (Hughes 2003, p. 31). In other words, it is the significance of connecting theory to the test and, currently, this validity is linked to communicative approaches, which is the third stage in the University of Cambridge assessment development process (Weir 2013). Similar to the other Cambridge exams, as it forms part of the Associations of Language Testers in Europe (ALTE), the PET includes construct validity by adapting to new theories and language shifts.
- Empirical validity refers to contrasting tests with similar or different criteria that have a comparable goal. This type of validity is commonly used in the creation of coursebook material and tests (Benmostefa 2014). This means that if some test takers take similar exams, they should be obtaining similar results. The PET exam has empirical validity because it is based on the CEFR.
- As Hughes (2003) mentions, face validity is not based on a scientific notion, but if a test lacks face validity, the test and results may not be accepted by candidates, employers, teachers, or other stakeholders. Gronlund (2006) complements this definition by adding that "students view the assessment as fair, relevant, and useful for improving learning" (p. 210). Mousavi (2012) proposes that face validity is

the degree to which a test looks and appears like a test. The PET exam has face validity, as Cambridge tests in general are often considered the standard.

- According to Cizek et al. (2008), consequential validity corresponds to the positive, negative, or neutral consequences arising as a result of testing. Thus, some ministries of education or institutions may decide to apply changes depending on either their own results or those of their peers. This type of validity corresponds to the effects of the results, and not to some characteristics within the exam.

The consequences of obtaining a particular result in the PET are far-reaching for students at the PrepaTec at Campus Zacatecas. It is for this reason that the reliability of the exam, which is discussed in the following section, is of such great importance.

21.3.3 Reliability

According to the Council of Europe (2001), reliability is “the extent to which the same rank order of candidates is replicated in two separate (real or simulated) administrations of the same assessment” (p. 177). For example, if a candidate takes the PET exam twice, the applicant’s results should be similar. In cases where the same result is obtained twice, the test is considered to have a reliability coefficient of 1. While expecting a candidate to obtain exactly the same results is perhaps too ideal a scenario, since the 1960s there have been estimations to measure test reliability. Reliability coefficients are not global, as exams may assess different skills individually or simultaneously. An important aspect for reliability is creating uniformity in grades, which depends greatly on examiners (Brown 2014).

Two types of reliability that influence test scoring are intra- and inter-rater reliability. Intra-rater reliability, also referred to as inter-trial and inter-replicate reliability, concerns “the degree to which each individual rater agrees with himself or herself over time when rating the same performance” (Fulcher and Davidson 2007, p. 132). Gwet (2014) describes this repetition as raters’ trials or replicates. On the other hand, inter-rater reliability is “the degree to which raters agree with each other when rating the same performances” (Fulcher and Davidson 2007, pp. 131-132). The raters may provide different results in either nominal or ordinal values depending on the requirements of the assessment. For the PET exam at PrepaTec, as in the case of all exams administered by Cambridge, the degree of inter-rater reliability should be expected to be considerable.

*21.3.4 Standard Error of Measurement (*SEM*)*

Reliability focuses on the test takers’ and evaluators’ performance on the day of the exam, although, given that no exam is flawless, these results are subject to a margin of error. Nevertheless, it is possible to estimate a candidate’s true score through some

Table 21.1 Cambridge English: Preliminary: Reliability and *SEM* Results

	Reliability	*SEM*
Reading/Writing	.88	2.25
Listening	.77	2.14
Speaking	.84	1.63
Total Score	.92	3.39

mathematical conversions (Hughes 2003). To make the estimations, one needs to obtain the standard deviation of the test (SD), which is "a sort of average of the differences of all scores from the mean" (Brown 1988, p. 69), and the reliability coefficient. Table 21.1, below, shows the *SEM* and reliability results as published by Cambridge Assessment English (2018a):

In addition, Brown (1999) indicates that "the standard error of measurement is related to test reliability in that it provides an indication of the dispersion of the measurement errors when you are trying to estimate students' true scores from their observed test scores" (p. 21). Now that the most pertinent concepts, validity, reliability, and *SEM* have been reviewed, the following section explains the methodological considerations taken during the completion of this project.

21.4 Methodology

The methodology used for this analysis was a qualitative instrumental case study where the data was collected through narrative autobiographies and a focus group interview. In contrast to the majority of research on testing, where a usually quantitative or mixed-method approach prevails, on this occasion a qualitative perspective was preferred. Based on the researchers' epistemological and ontological assumptions, it seemed appropriate to analyze students' perceptions through a qualitative lens, as it provides an opportunity to conduct a social inquiry that concentrates on how people interpret and make sense of their experiences and the world (Holloway 1997). While this may not allow for generalizability and representability, it is well posited in the interpretivist/constructivist knowledge claim where the goal is to create established knowledge of the participant's perspectives of a situation (Creswell 2003).

Regarding the case study method, this examines a temporally and spatially well delimited setting (Stake 1995) and explores descriptive or explanatory concerns (Yin 2003). In addition, case studies are commonly used as a methodology in EFL and ESL contexts on a variety of topics (see El Masry and Saad 2018; El Mortaji 2018; Saif 2018 for recent examples). A common feature of these projects is that they observe the same subject from different viewpoints. In regard to the current study, these perspectives include the students' experiences as well as the institution's and the testing body's official papers. For the purpose of this research, the instrumental case study, defined by Heigham and Croker (2009) to have "the goal of illuminating a particular issue, problem or theory" (p. 70), seemed to correlate best with the overall objective, as one of the aims was to understand why the PrepaTec's expectations were not being fulfilled. In the end, twenty Mexican students, between the ages of 14 and 18, participated, seven of whom achieved the institutional goal and obtained a B2 level, while the other thirteen did not.

Following a pilot study carried out to measure the feasibility of the research project (Lancaster et al. 2004) and probe the appropriateness of the research method or instruments (Van Teijilingen and Hundley 2002), the final research design comprised a guided autobiography designed to obtain narrative data and a focus group interview to gain further insights. The use of autobiographies in qualitative research provides a story or personal view of a case. Eakin (2008) considers that "narratives display the imprint of the culture and its institution on the individual's sense of identity" (p. 116). This instrument appeared particularly useful for three reasons: First, the participants could take the time they needed to write at their own pace; second, they were not influenced by the physical presence of a researcher who might be asking them questions; and, third, this instrument delivered unexpectedly valuable information, as will be seen in the following section. On the other hand, the focus group interview, described by Freitas et al. (1998) as "a type of in-depth interview accomplished in a group, whose meetings present characteristics defined concerning the proposal, size, composition, and interview procedures" (p. 2), beneficially contributed to the data collection process. Organized as a multi-contributor semi-structured interview, it comprised questions based on the doubts and interesting themes that had emerged from the autobiographies.

The data segments in the Findings below are coded according to these data sources ("A" for the autobiography and "FG" for the focus group interview, respectively), followed by the pseudonym chosen by each participant for him/herself. In addition, the results the students obtained in the exam are indicated, as this piece of information lies at the heart of the inquiry and also influences the research outcomes, which can be appreciated in the following section.

21.5 Findings

As will be revealed in the presentation and analysis of the data that follows, both the more privately shared autobiographies and the focus group interviews conducted in a true constructivist fashion provided fruitful information. The objective was twofold: first, to understand the problems the students face when taking the PET in the hope of obtaining a B2 level (Pass with Distinction grade) under the pressure of needing this in order to be able to graduate and/or to move on to learning another foreign language; second, to find possible solutions in the event of their failure. In order to comprehend their situation, in the initial stages, the differences between the students who obtained B1 and B2 in the PET exam were analyzed. In general, no significant deviation was found in their profiles in regard to the following aspects: their educational background, their motivation, their experience of international encounters, their use of English outside class, and even their preparation for the actual exam.

First, their educational background was very similar, as all of them had previously studied in private bilingual and trilingual schools. The autobiographies showed there was little to no observable difference between B1 and B2 students, for example, Luke and Bunny.

> *[In elementary school] I had English classes. If I remember well, every day I had an English class, and another class using this language.* (A-Luke, B1)
>
> *From elementary school, I studied classes that were in English as creativity or leadership.* (A-Bunny, B2)

Second, their motivation varied little, as most of the participants perceived English as an essential tool to pass their content English classes, obtain better opportunities in the future, and maintain communication with international friends. This can be noted in the contrasted segments from Wendy and Abigail's accounts, who attained B1 and B2 results, respectively.

> *Personally, I think that learning English really opens new opportunities and doors to the world. That guarantees me important jobs where this language is a requisite and it's used as a primary element. It can be in my Mexico or other countries.* (A-Wendy, B1)
>
> *The thing that motivated me the most about learning the language was that I can go to the United States of America quite often. I wanted to talk to strangers, have different conversations with random people and understand them. I have never been quiet nor antisocial. I wanted to talk to people as if I lived in USA.* (A-Abigail, B2)

Third, their experience of international encounters also showed a lot of similarities, as both B1 and B2 students, for example, Karen and Wendy, as compared with Albert, had travelled or studied in Australia, Canada, Denmark, Germany, Lebanon, the United States, among others, an indicator of their fairly privileged socio-economic status.

> *I went to Hamilton, Canada to study one semester of high school and improve the language in an English-French-speaking country, the school where I was studying called Columbia International College and it had students from more than 77 countries in the world.* (A-Karen, B1)
>
> *I went to a camp in Vancouver where I had classes of English and I get to know and talk with people there in English too, and last summer I went to Denmark where I had to communicate completely in English (there was no option).* (A-Wendy, B1)
>
> *My family used to make a lot of trips to English-speaking countries, which, I think, had a major impact in my development of English skills. These experiences made my knowledge of this language grow exponentially [...] And I'm my dad's translator [interpreter] every time we travel to an English-speaking country, so that makes it even more important.* (A-Albert, B2)

Fourth, there was little difference regarding their use of English outside class, as they were all exposed to the foreign language in similar scenarios, such as communicating with friends, playing videogames, and reading, among others. Here, a B1 student, Luke, and a B2 student, Athena, show commonalities regarding their linguistic habits.

> *Normally, I use English to watch videos and movies without subtitles, I like to play videogames in their original language (so I also play them in English), and most of my music is in English.* (A-Luke, B1)
>
> *I read, listen and write English. I play a lot of videogames and I listen to a lot of music in English. I write a lot because I have a Tumblr account and I run it in English. I must use English, so I can reach more people and share my content.* (A-Athena, B2)

Finally, the group was almost completely homogenous in terms of their preparation for the actual exam, as 19 out of 20 students relied only on their English classes in their learning experience leading up to testing day. As an example, there is little difference between the self-reported experiences between Karen and Nina, who scored differently on the test.

> *I prepared for this test in my last English class which was Advanced III.* (A-Karen, B1)
>
> *None of the times I took it I got a special training but the one in my regular class.* (A-Nina, B2)

Bearing in mind that the students were enrolled in courses that the institution classified at B2, C1 and even C2 level and used the corresponding material, they could be expected to pass the PET with Distinction.

However, despite their seemingly parallel experiences, the exam results showed considerable and quantifiable differences. The respondents who were already taking classes in another foreign language at the time of the data collection had obviously obtained a B2 level that allowed them to do so. They did not share the exact score they received on the PET exam, and neither did all of the B1 level participants. Nevertheless, the PET scores of those B1 participants who disclosed their results, as shown in Table 21.2, revealed intriguing correspondences regarding their perceptions of the exam, as will be seen further below.

From the seven applicants who obtained B1 and reported their grades, six attained a Pass with Merit and the other was a point behind. It is important to remember that if applicants attain between 153 and 159 points, they will be given a Pass with Merit certification, which means that any candidate with 159 points is exactly one point short of obtaining a Pass with Distinction certificate. As can be observed in the

Table 21.2 Participants' Cambridge and Standardized Test Results

Participant	Cambridge Scale Score	Standardized Score
Mabel	159	89
Wendy	159	89
Luke	157	88
Tony	156	87
Matt	154	86
Regina	154	86
Michelle	152	84

scores table, a difference of one single point in the Cambridge Scale Score does not necessarily correspond to a difference in the standardized score.

Due to the institutional policies explained earlier, whereby students with a Distinction can graduate from senior high school without further complications and enroll in other language classes if they still have at least a semester left of their studies, falling short by a single point has rather far-reaching consequences, as Mabel, a student with B1 results, mentioned:

> *Now that I am in high school, I want to study another language, but I can't because I don't have a level B2 PET certification. This is really frustrating for me. I know I can speak English. My grades are good in this subject and, in fact, I have been told by native speakers that my English is really good. I don't see why I can't pass to another language. Just because I missed one point on a stupid test [...] Thanks to a piece of paper that isn't true, a certification that I feel is nothing else than a way to make money and stress students out.* (A-Mabel, B1)

In this autobiographical excerpt, Mabel expresses her disapproval of the results of the test. She considers her English language skills to suffice for the level. Her frustration led her to comment on the subjectivity of the test and its monetary value. This is understandable, as her standardized score is 89, so she passed with merit, but nevertheless did not do well enough to reach the coveted B2, or Pass with Distinction level. Consequently, it is arguable whether this test was an appropriate medium through which to fulfill the goal established by the institution.

The Tecnológico de Monterrey Campus Zacatecas selected the PET exam with the expectation that their students would achieve a score of 160, corresponding to a Pass with Distinction. Cambridge offers a scale converter (Cambridge Assessment English 2018b), so that the test takers can compare their Cambridge Scale Score (CSS) to a 1–100 standardized testing score (STS). Hence, the minimum of 160 CSS points converts to a standardized result of 90. If we consider, on the one hand, that the exam has a *SEM* of 3.39, there is a chance that a student who may have obtained a 157 CSS might actually have achieved the 160 points. On the other hand, it means that all applicants should aim to score 93 STS to obtain 164 CSS to avoid the measurement error.

Most students are obviously unaware of this computation, yet they can easily compare the results they obtained on the PET with other previous exam scores. For example, some participants reported prior Cambridge experience with the lower level Flyers or KET exams, or, as in Tom's case on his Test of English as a Foreign Language (TOEFL) result:

> *I have also taken the TOEFL test and passed it with a C1 level.* (A-Tom, B1)

In contrast with this self-reported C1 level on another internationally recognized exam, Tom obtained a B1 score when he took the PET. Based on the principles of empirical validity, these types of conflicting outcomes should not frequently occur, as applicants are expected to obtain similar results. Given that Tom did not expand on the reasons for such a discrepancy, we can only surmise the possible causes, which may be due to greater familiarity with TOEFL than PET, either because of implicit cultural background or exam-specific preparation.

In contrast with Tom, who compared his own results obtained on two different tests, other students—unsurprisingly—assessed themselves against their peers, and sometimes found themselves to be wanting, as seen in the extracts below, taken from B1 students' contributions via the autobiographies and the focus group interview.

> *Some of our classmates don't know anything compared to us [the students in the focus group], and they obtained more points.* (FG-Michelle, B1)
>
> *I also have some friends who are B2 that I have asked to help me with my English homework and they can't because they don't know what to do and that I know my English is better.* (A-Mabel, B1)

Even keeping a critical perspective on self-reported comparative evaluation, it is noteworthy how the students' experiences with CEFR levels do not seem to be reflected in their exam scores. In this sense, their reservations about the PET results could be construed in terms of a perceived reduction in empirical validity for the purposes of the institutional context.

Furthermore, to some participants, the exam seemed to lack construct validity, that is, it bears little relevance to "real-life" language use, at least, in their experience or their particular context. Among others, Luke reported his frustration concerning the exam, as he considered that it mismatched his abilities and did not correspond to a natural linguistic setting:

> *About the slang expressions, or phenomena like that, they just see the formal use of English, something that we probably won't use if our plan is not to live in the English-speaking countries, so why do we need to prove that we can be formal when we just want to get along with people from other countries? The PET is fine. A certification can show your formality of a language, but if you don't get a C1 or B2 level it doesn't mean that you don't have it. It's subjective and everyone should evaluate why is he/she studying it. I have seen people who their certificates show a B2 or higher level, but after a year, they forgot how to write or speak. You need to continue using the language in order to preserve it.* (A-Luke, B1)

Even though Luke understood that a test is based on academic patterns, he perceived the language as a tool to communicate rather than a stilted set of rules, as his remark on the importance of colloquial language would suggest. A study carried out by Zhan and Wan (2016) found that "the participants perceived that the communicative tasks […] were inauthentic, which might cause them to rely on their native language and overuse the compensatory test taking strategies as they took the test" (p. 12). This is similar to the unnaturalness perceived by Luke, where he intends to use the language in both formal and colloquial scenarios.

This chapter has already discussed the similarities between the B1 and B2 students. However, there were also some differences. The B1 students in particular recounted their struggles while taking the exam, especially with the oral section, including both the listening and the interview. Regarding their performance on the listening section, the B1 test takers cited problems with the acoustics and the content. As far as the former is concerned, they seemed to have had difficulties with the speakers and the seating location.

> *The last time I took it I couldn't listen well enough because I was in the back of the room, so the sound was quite distorted.* (A-Wendy, B1)

> *I was in front of the speakers, so I didn't have problems with the audio.* (FG-Matt, B1)

Thus, it appears, that the infrastructure was not entirely satisfactory for those who happened to be seated at a greater distance from the speakers, where distortion and low volume may have affected the listening experience.

In terms of the content, the participants, who are used to North American accents in the Mexican context, voiced their disappointment with the quality of the audios, and their unfamiliarity with British accents and shared the following in the focus group interview.

> *It was weird because the audio had a noticeable British accent. We are not used to that accent… but it's not only an accent, they use a different vocabulary.* (FG-Michelle, B1)
>
> *Everything was fine except the British accent; they said words that I have never heard before in movies or songs.* (FG-Regina, B1)
>
> *In classes our teachers speak American English. We [the students] also speak American English. In television, we watch American series, but when we present a test, they use the British accent*. (FG-Tony, B1)

In these fragments the test takers expressed that they were not properly prepared for the listening section. To illustrate, Tony, who had lived in Texas for almost two years, conveyed that he felt an incongruence between the exposure they received in Mexico and the variety of English selected for the exam. Therefore, context seems to be a key issue that surely needs to be addressed by the institution during exam preparation.

In addition, some of the participants also seemed to have experienced complications in the speaking section, especially in terms of the vocabulary, their interview partner, and, on a couple of occasions, the rapport with the evaluators. Michelle, who also obtained a B1 level, commented on the subject during the focus group interview.

> *The vocabulary is important, too. There were unknown words in the listening and the speaking, it made things complicated.* (FG-Michelle, B1)

Another aspect of the speaking section is that, over the course of a series of interviews that lasted more than five hours, the evaluators may well have experienced boredom, fatigue, and lack of concentration. Some of the participants reported that the interlocutors acted with a certain apathy, which could create low coefficients of intra-rater reliability. Among the students who addressed the subject, Renata mentioned:

> *On the speaking section, I got a little bit nervous because the woman who was in charge of the interviews with me and my classmate was intimidating. That woman was old, and she has a kind of villain face. In addition, she had a very bad sense of humor because when my classmate and I were talking about something that she told us to speak about. My classmate said something funny and we both laughed for a second and she was completely serious, but not good serious I mean serious like angry. So that fact made me nervous and I think that is why I did not get very good results in that section.* (A-Renata, B1)

As explained above, in addition to validity, two important aspects that influence test scores are inter- and intra-rater reliability. From the students' excerpts, it may be inferred that they perceived that intra-rater reliability could have had an impact on the scores. However, while there is the possibility of issues in intra-rater reliability, this excerpt perhaps better illustrates that the candidates were not completely prepared for certain tasks.

Similarly, it is interesting to note that two other students with B1 outcomes attributed their weaker performance in the speaking section to human factors.

> *Part of the reason [why I obtained the lowest score in the speaking section] is that we were in a very small place… They make you feel intimidated. The interviewers spoke way too fast. I had to pay a lot of attention… They were not transmitting to us any confidence.* (FG-Michelle, B1)
>
> *In my PET results I did excellent in all the areas except for speaking, but that doesn't mean I don't know how to speak English. I was nervous in my interview and I had a mean evaluator and not a great partner.* (A-Mabel, B1)

Even though they both mentioned their own issues, such as anxiety and the difficulty of maintaining attention, neither of them had any compunction about, at least partially, blaming their partner and/or the evaluators, whom they viewed as less sympathetic than they had been accustomed to during their preparation.

21.6 Insights Gained

Some limitations were identified during the analysis. These included the expected bias in the participants' self-reported experiences—something characteristic of qualitative studies that inquire into people's perceptions. In addition, it is possible that the timing of the data collection *vis-à-vis* the respondents' receipt of their results may have affected their attitude. However, the emerging patterns discussed throughout this chapter seem to validate research of this kind, as it provides valuable insight into institutional practices from a hitherto unexamined perspective.

If we approach this from the opposite direction, for reasons outlined in the previous paragraph, it can first be noted that the students who did not obtain the B2—that is, the Pass with Distinction—certificate on their PET exam reported more negative experiences of the exam. They appeared much more critical of the circumstances, questioning the exam's reliability and empirical validity. While some acknowledged their own weaknesses, some redirected the responsibility toward the infrastructure, such as their location during the listening section, and even human factors, including their interview partners and the evaluators and so also mistrusted the intra-rater reliability of the exam. Considering human nature, this is hardly surprising, although there are further extenuating circumstances.

First and foremost, there is much at stake for the candidates. This said, however, they are not completely mindful of the benchmarks. Some of them know that they need 160 CSS, but are unaware of what this actually entails. It is simply an abstract number and they have little awareness of how it is broken down or how it might be achieved. In fact, given that the *SEM* is 3.39 points, these students should be aiming to obtain 164 CSS or an STS of 94 points (provided the PET continues to be the evaluation instrument). Meanwhile, the institutional and self-imposed pressure will also continue.

Secondly, most applicants take classes that correspond to a C1 or C2 CEFR level at the PrepaTec, and due to this placement in supposedly more advanced courses, they seem to be under the impression that their competences in English are higher than they actually are. This leads to false expectations of passing exams, something that should also be taken into account when considering the institution's language learning policies. Additionally, as regards access to third language classes, another possible option is to disassociate these test results from institutional requirements in terms of the students' opportunity to start learning another foreign language.

Furthermore, it is clear from the data that the students would benefit from a more exam-oriented preparation process, both in terms of the target test's form and its content. In regard to the former, especially the oral sections, perhaps the students could be exposed to less favorable circumstances when preparing for the listening exam and trained to face a variety of scenarios relating to the speaking section. In addition, tailored exam preparation activities could be included in the syllabus, without, of course, falling into the trap of a negative washback scenario.

With respect to the content, and despite the geographical proximity to North American varieties of English, if the students continue to be required to take a Cambridge exam, they should be prepared with test-relevant material, including exposure to different British accents. This would help them both in the speaking element, and even more so in the listening sections, where they reported being confused by the accents, words, expressions, and the pacing. In general, the students seemed to be used to friendly scenarios and to struggle with the unknown, for which reason they could benefit from being pushed a little more out of their comfort zone.

However, having considered all the above, the most significant issue seems to be related to consequential validity. In other words, there is much in the balance for the students when taking the test. If exam-related anxiety in itself were not enough, the stakes are raised even higher by the fact that, due to contextual restraints, it is not enough for the test takers to "pass" the exam, they have to go further and obtain a distinction. Unless they do so, they will not comply with institutional requirements for their graduation, or, in less critical circumstances, they cannot enroll in classes to start another foreign language. This prospect seems to weigh on them heavily, but could be remedied by choosing a different exam, perhaps one aimed at achieving a B2 level without having to obtain the highest scores, or even by designing an institutional one, which, of course, may carry its own risks.

21.7 Conclusion: Implications for Test Users

Having discussed the implications of this study carried out at the high school section of the Instituto de Monterrey at their Zacatecas Campus with respect to the test takers' experience of their high-stakes exam, it can be concluded that a qualitative approach can yield results that complement more numerical research and thus contribute to our understanding of exam-taking practices in a particular setting. This may be fruitful for other institutions who sometimes use national and international certifications for a variety of reasons, including saving money, time, and research investment to develop their own context-driven tests. However, these institutions should consider what additional stress they associate with these high-stakes exams, or what consequences the exams could have for the test takers. In this study, the student participants experienced stress while taking these certification exams. It is advisable to avoid negative scenarios similar to the one under study, where the test scores served as a basis for further institutional requirements. In terms of future research into comparable contexts, a number of areas for exploration can be identified. First, a parallel project could be carried out at other Tecnológico de Monterrey high schools around the country in order to compare and contrast students' experiences. Second, and perhaps with a view to different settings, other stakeholders, including the administration, teachers or even parents could be included as participants to gain a more rounded image of the practices at hand. Finally, it would be useful to analyze various exams applied in either, say, the same context or elsewhere for consequential validity. The opportunity in Zacatecas may arise sooner rather than later, as it appears that the institution has also realized that the current arrangements may not be best suited to their purposes and is contemplating a change in test provider. How these new measures pan out would certainly supply fertile ground for future investigations.

References

Alshahrani, M. A. (2016). The level of anxiety on the achievement of the Saudi EFL learners. *Arab World English Journal, 7*(3), 65–76. https://doi.org/10.24093/awej/vol7no3.5.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Benmostefa, N. (2014). *Reflections upon the baccalaureate EFL tests as a source of and a means for innovation and change in ELT in Algeria (Unpublished doctoral thesis)*. Chetouane, Algeria: University of Tlemcen.

Brown, H. D. (2014). *Principles of language learning and teaching* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Brown, J. D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5,* 19–31. https://doi.org/10.1177/026553228800500102.

Brown, J. D. (1999). Standard error vs. standard error of measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter, 3*(1), 20–25.

Cambridge Assessment English. (2018a). *Quality and accountability: Reporting reliability figures*. https://www.cambridgeenglish.org/research-and-validation/quality-and-accountability/. Accessed 15 December 2018.

Cambridge Assessment English. (2018b). *The Cambridge English scale*. https://www.cambridgeenglish.org/exams-and-tests/cambridge-english-scale/. Accessed 15 December 2018.

Cambridge Assessment English. (2018c). *Validity and validation*. https://www.cambridgeenglish.org/research-and-validation/validity-and-validation/. Accessed 15 December 2018.

Cambridge English Language Assessment. (2016). *Cambridge English preliminary for schools: Handbook for teachers*. Cambridge, UK: Cambridge University Press.

Choi, I.-C. (2008). Test fairness and validity of the TEPS. *Language Research, 35*(4), 571–603. https://doi.org/10.1177/0265532207083744.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397–412. https://doi.org/10.1177/0013164407310130.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Eakin, P. J. (2008). *Living autobiographically: How we create identity in narrative*. New York, NY: Cornell University Press.

El Masry, T., & Saad, M. R. M. (2018). On the cultivation of their community of practice: A case study of EFL Malaysian pre-service teachers. *The Qualitative Report, 23*(4), 952–977.

El Mortaji, L. (2018). Effects of sustained impromptu speaking and goal setting on public speaking competence development: A case study of EFL college students in Morocco. *English Language Teaching, 11*(2), 82–98. https://doi.org/10.5539/elt.v11n2p82.

Freitas, H., Oliveira, M., Jenkins, M., & Popjoy, O. (1998). The focus group, a qualitative research method: Reviewing the theory, and providing guidelines to its planning. *Journal of Education, 1,* 1–22.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. New York, NY: Routledge.

Gronlund, N. E. (2006). *Assessment of student achievement* (8th ed.). Boston, MA: Allyn & Bacon.

Gutiérrez, X. (2017). Explicit knowledge of the Spanish subjunctive and accurate use in discrete-point, oral production, and written production measures. *Canadian Journal of Applied Linguistics, 20*(1), 1–30.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics LLC.

Heigham, J., & Croker, R. A. (2009). *Qualitative research in applied linguistics: A practical introduction.* Basingstoke, UK: MacMillan.

Holloway, I. (1997). *Basic concepts for qualitative research.* Oxford, UK: Blackwell Science.

Hughes, A. (2003). *Testing for language teachers* (5th ed.). Cambridge, UK: Cambridge University Press.

İpek, H. (2016). A qualitative study on foreign language teaching anxiety. *Journal of Qualitative Research in Education, 4*(3), 92–105. https://doi.org/10.14689/issn.2148-2624.1.4c3s5m.

Lancaster, G. A., Dodd, S., & Williamson, P. R. (2004). Design and analysis of pilot studies: Recommendations for good practice. *Journal of Evaluation in Clinical Practice, 10*(2), 307–312. https://doi.org/10.1111/j.2002.384.doc.x.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x.

Li, H. (2015). A study of EFL listening anxiety in a test setting. *International Journal of English Linguistics*, *5*(2), 106–114. https://doi.org/10.5539/ijel.v5n2p106.

Mizumoto, A., & Takeuchi, O. (2009). Examining the effectiveness of explicit instruction of vocabulary learning strategies with Japanese EFL university students. *Language Teaching Research, 13*(4), 425–449. https://doi.org/10.1177/1362168809341511.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10), 71–81.

Mousavi, S. A. (2012). *An encyclopedic dictionary of language testing* (5th ed.). Tehran, Iran: Rahnama Publications.

Rodríguez, M., & Sadowki, M. (2002). Effects of rote, content, keyword, and context/keyword methods on retention of vocabulary in EFL classrooms. *Language Learning, 50*(2), 385–412. https://doi.org/10.1111/0023-8333.00121.

Saif, M. S. (2018). Self-assessment in EFL grammar classroom: A study of EFL learners at the Centre for Languages and Translation, IBB University. *International Journal for Research in Education, 42*(2), 289–324.

Shaban, A. S. (2014). A comparison between objective and subjective tests. *Journal of the College of Languages, 30,* 44–52.

Stake, R. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

Van Teijilingen, E., & Hundley, V. (2002). The importance of pilot studies. *Nurs Stand, 16*(40), 33–36. https://doi.org/10.7748/ns2002.06.16.40.33.c3214.

Weir, C. J. (2013). *Measured constructs: A history of Cambridge English language examinations 1913–2012.* Cambridge, UK: Cambridge University Press.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

Zhan, Y., & Wan, Z. H. (2016). Test-takers' beliefs and experiences of a high-stakes computer-based English listening and speaking test. *RELC Journal, 47*(3), 363–376. https://doi.org/10.1177/0033688216631174.

Part IV
Learning from Tests of Language Skills

Chapters in this part discuss tests of reading, writing, speaking, and/or listening in local high-stakes testing placement tests, a placement test, and classroom assessment, as well as general issues about testing reading and speaking. Chapters 22–26 are experience-based papers and Chapters 27–28 data-based. The chapters in this part are as follows:

- Shin (Chapter 22) argues that including prosody as part of reading fluency gives a more accurate assessment of this skill, which can be measured using reading fluency scales, thus avoiding underrepresentation of reading fluency.
- Chapter 23 by Ngo explains how using authentic, real-world audio texts from university lectures for a high-stakes listening comprehension test was problematic – because they did not provide suitable material for development of listening items according to test specifications.
- In Chapter 24 Boraie and Shabara describe how a paired speaking task meant to assess interactive competence (a debate with different written information input given to each test taker) was problematic in measuring the speaking proficiency of test taker pairs with different English proficiency levels.
- In Chapter 25 Khabbazbashi et al. document how automated speaking evaluation narrows assessment of speaking by focusing on what is easy to measure (elicited speech) vs. what is integral to spoken interaction (speaking freely).
- Dursun et al. (Chapter 26) highlight a different reading skill, which is reading comprehension for academic research, and describe how a test using a translation task did not assess this skill which is required of graduate students researching in additional languages.
- Sabieh's Chapter 27 is action research investigating the problem she identified of her students using analytic rubrics as checklists, resulting in limited learning.
- Zabala-Delgado's Chapter 28 investigates another aspect of the rating process, which is rater training, in a small institution where the raters are experienced teachers who work together.

With the focus on specific language skills, three areas of challenge emerge as themes in the chapters in this section: more effectively measuring constructs (aspects of language ability being assessed), improving the rating process, and collaboration with stakeholders. Five chapters share the theme of more effectively measuring aspects of reading, speaking, and listening in language tests (Chapters 22–26). Challenges with assessing constructs of listening and speaking are highlighted in three experience-based chapters (Chapters 23–25). The theme of challenges with the rating process is seen in two data-based chapters, Chapters 27 and 28. Challenges and opportunities in collaborating with people involved with test development and use are particularly seen in Chapter 26, and also in Chapters 24 and 28.

Chapter 22
Completing the Triangle of Reading Fluency Assessment: Accuracy, Speed, and Prosody

Jihye Shin

Abstract In response to a mismatch between the definition and assessment practices of reading fluency, this chapter aims to draw L2 educators' and researchers' attention to a comprehensive approach to reading fluency assessment derived from a complete conceptualization of the construct. The traditional view of reading fluency focuses on accurate decoding of a text at a sufficient rate. In line with this view, reading fluency is commonly assessed by having students read aloud a given text for one minute and recording their words correct per minute (WCPM), which measures accuracy and speed. However, reading fluency has evolved into a more multifaceted concept which now includes reading speed, accuracy, and prosody as its components. Despite the recognition of reading prosody as another critical aspect of reading fluency, prosody is often neglected in assessment, resulting in a discrepancy between the depiction of reading fluency and its assessment. This discrepancy in turn leads to a series of issues such as potentially reducing the construct of reading fluency and placing an overemphasis on reading accuracy and speed, hindering informed assessment and instruction. It also becomes particularly problematic in L2 contexts where students may decode without comprehending what is read. In support of incorporating prosody into reading fluency assessment, the chapter draws upon the ongoing discussions surrounding reading fluency to address the negative consequences of using WCPM alone as a reading fluency measure and presents existing rating scales for measuring prosody.

22.1 Introduction: Purpose and Testing Context

It has long been perceived that fast and accurate reading is a hallmark component of reading fluency, supported from the automaticity theory (Kuhn and Stahl 2003). However, scholars have recognized that, in addition to fast and accurate reading, using appropriate prosody such as suitable pitch, stress, and phrasing is another defining feature of reading fluency (Dowhower 1991; Grabe 2009; Rasinski 2012). Despite

J. Shin (✉)
Northern Arizona University, Flagstaff, AZ, USA
e-mail: js3654@nau.edu

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_22

the lack of a unified view of reading fluency, the consensus is that reading fluency should be (re)defined as the ability to read with speed, accuracy, and appropriate prosody (Kuhn and Stahl 2003; Rasinski 2004a). Yet, reading fluency assessment is still playing catch-up and tends to rely solely on the automaticity aspect of reading fluency, that is, reading speed and accuracy (Kuhn et al. 2010).

Assessment of reading fluency has commonly been done by having students read aloud a given text for one minute and recording their words correct per minute (WCPM). However, recognizing that the operational definition of WCPM no longer parallels the construct of reading fluency, efforts from L1 researchers and educators have been underway to align assessment practices with the current definition that incorporates prosody (Schrauben 2010). Developing reading fluency has become an important issue in L2 settings as well (Grabe and Stoller 2011). Yet, such research-based practices are still underutilized in L2 research and classrooms, and the acknowledgment of the (re)conceptualization of reading fluency has not been as robust in L2 contexts.

Drawing on the ongoing theory and research-based discussions of reading fluency, this chapter aims to inform L2 educators and researchers that the conceptualization of reading fluency has evolved, which now includes speed, accuracy, and prosody as its components. More importantly, the chapter aims to bring awareness to a more comprehensive approach to assessment that incorporates prosody. To this end, this chapter discusses (1) a mismatch between the definition and current assessment of reading fluency, (2) potential issues rising from assessment that focuses heavily on the automaticity aspects (i.e., accuracy and speed), and (3) rating scales that can be used for measuring reading prosody.

22.2 Testing Problem Encountered

22.2.1 Mismatch Between the Definition and Assessment of Reading Fluency

First and foremost, a mismatch between the definition and assessment must be addressed in discussion surrounding current issues of reading fluency assessment. The traditional conceptualization of reading fluency is rooted in the classic automaticity theory of reading (LaBerge and Samuels 1974). The underlying assumption of this theory is that efficient execution of low-level processes, such as decoding and word recognition, results in freeing up attentional capacity for higher level processes and meaning construction, which then leads to more successful reading performance (Samuels 2006). The automaticity view clearly accounts for two components of reading fluency: reading accuracy and speed.

While these automaticity components reflect decoding skills and are essential prerequisites for building fluency, they are not sufficient conditions for fluency (Kuhn and Stahl 2003). A fluent rendering of a text represents more than simply accurate and quick reading of words. Fluent readers replicate the author's intended phrasal structure to some degree by employing appropriate variations in pitch, stress patterns, and pauses, all the while building meaning (Dowhower 1991; Hudson et al. 2005). Thus, readers' use of prosody reflects their ability to segment text according to syntactic and semantic elements, as well as their comprehension, which otherwise would be an invisible process (Grabe 2009, 2010; Jiang 2016). Conversely, disfluent readers read with a monotone and in a word-by-word manner while ignoring punctuation and any other syntactic boundaries, often without comprehension (Rasinski et al. 2009). The (re)conceptualization of reading fluency is well encapsulated by the definition proposed by Kuhn et al. (2010):

> Fluency combines accuracy, automaticity, and oral reading prosody, which, taken together, facilitate the reader's construction of meaning. It is demonstrated during oral reading through ease of word recognition, appropriate pacing, phrasing, and intonation. It is a factor in both oral and silent reading that can limit or support comprehension. (p. 242)

As illustrated in this definition, reading fluency is manifested in accuracy, speed, and prosody, working in concert to facilitate reading comprehension (Klauda and Guthrie 2008). However, assessment of reading fluency still heavily relies on reading accuracy and speed by using a metric of WCPM, which creates a discrepancy between the definition and assessment (Kuhn et al. 2010).

22.2.2 Negative Consequences of Using WCPM as an Isolated Measure of Reading Fluency

There is no denying that accuracy and speed are critical components of reading fluency; yet, they are by no means the only ones. Relying solely on a metric of WCPM places a heavy focus on reading accuracy and speed, resulting in an inadequate representation of reading fluency. This leads to a series of subsequent issues that are interdependent.

Given the current definition, using WCPM as an isolated reading fluency measure raises an issue of construct validity—the degree to which a test measures the construct as it claims to be. As Deeney (2010) puts it, "widespread use of specific assessments can ultimately define the construct being assessed" (p. 442). Although WCPM is generally understood and widely used as a *fluency* measure, it is rather an *automaticity* measure that can indicate readers' ability to identify words accurately and recognize them instantly (Deeney and Shim 2016). Researchers have therefore warned against using a measure that assesses only accuracy and speed, because it can potentially lead to reducing the construct of reading fluency (Kuhn et al. 2010; Samuels 2007). Construct validity is a critical issue especially at this juncture where promoting the current definition of reading fluency is still an ongoing process particularly in L2 settings.

A related issue is that a metric of WCPM alone may not provide a full picture of ability to read fluently because it does not necessarily show the reader's comprehension. This issue raises concerns particularly for L2 readers who may decode what is read without actual comprehension (Lems 2003). This is alarming given that (1) reading fluency should reflect reading comprehension as suggested in the current definition of reading fluency (Grabe 2009; Hudson et al. 2005; Schrauben 2010) and (2) the ultimate goal of fluency interventions and assessments is to help learners achieve better reading comprehension. Fluency without comprehension, as Pikulski and Chard (2005) stated, is of limited value.

Negative consequences of reading fluency assessment confined to WCPM can also affect reading instruction by overemphasizing reading speed and accuracy, possibly at the expense of comprehension (Kuhn et al. 2010). Due to the focus on WCPM, it becomes inevitable to grant a corresponding privilege to reading accuracy and speed in instruction rather than to reading with appropriate phrasing and pacing. Samuels (2007) cautioned that fast, staccato reading can result from the excessive focus on fast decoding and that such reading behavior may interfere with rather than facilitate reading comprehension, which goes against the end-goal of reading fluency. In fact, Ardoin et al. (2013) reported that when fluency instruction was provided with a focus on reading rates, students improved their rate of reading demonstrated in WCPM scores but tended not to use pauses, and ignored sentence and paragraph structures. Appropriate phrasing and pausing embedded in oral expression, however, are precisely what signify the reader's active interpretation and construction of meaning from the text (Rasinski 2004b). These findings also raise concerns about the outcome of reading fluency assessment becoming "not fluency in its broad sense but increasing WCPM scores" (Deeney and Shim 2016, p. 110).

Furthermore, using WCPM scores for assessment of reading fluency and/or student progress monitoring may not adequately serve its purpose, calling into question consequential validity—the degree to which consequences of assessment match the expected consequences. Disfluent reading behavior includes inappropriate use of prosody or a lack thereof, such as reading in a monotone and word-by-word manner while blowing through syntactic boundaries. The current assessment approach reflects reading accuracy and speed but is not capable of capturing the prosodic aspect of reading (dis)fluency. Readers who are struggling with the latter can still score high enough by attending to simply reading fast (without appropriate prosody and comprehension). As a result, they will not be identified as disfluent readers and thus will not receive instructional support that they need and deserve (Valencia et al. 2010). Taken together, data gained from WCPM scores may be too limiting to be used for formulating instructional decisions and diagnostic performance profiles.

An associated impact of relying solely on WCPM also includes educators' and students' misperception of reading fluency. In their study on consequences associated with WCPM, Deeney and Shim (2016) found that 25% of their teacher sample (n = 77) defined fluency as accuracy and speed specifically in terms of WCPM. Students often adjust their views and learning behavior according to what and how they are assessed and could very well be affected by assessment practices as well.

In fact, critics have feared that the WCPM measurement can unintentionally portray an oversimplified picture of reading fluency: That is, fluent readers simply read fast (Kuhn et al. 2010; Valencia et al. 2010). This narrowed view of reading fluency, influenced by the current assessment practices, in turn contributes back to the issue of the underrepresentation of the construct of reading fluency.

22.3 Solution/Resolution of the Problem

22.3.1 Aligning Assessment with the Definition of Reading Fluency

With the goal of establishing assessment practices that can foster reading fluency in its complete form, this chapter advocates for assessment that addresses all three components of reading fluency—reading accuracy, speed, and prosody—within the context of reading comprehension. Incorporating prosody into assessment not only counterbalances the overemphasis on accuracy and speed (Kuhn et al. 2010) but also allows readers' understanding of the text to be reflected in their oral expression, which is where the current assessment approach falls short (Klauda and Guthrie 2008; Veenendaal et al. 2015).

Procedures for measuring reading fluency in terms of accuracy, speed, and prosody within the context of reading comprehension are described as follows. First, researchers and educators suggest that level-appropriate passages should be used as the reading material either directly from the curriculum (Jenkins et al. 2003; Lems 2003) or from outside sources (Klauda and Guthrie 2008; Young et al. 2015), in which case readability of the passages can be checked and modified as needed to approximate that of materials used in the curriculum. To score WCPM for accuracy and speed, students typically read the given text aloud for one minute during which the number of words accurately read are recorded (see Jenkins et al. 2003; Rasinski 2004b; Schwanenflugel et al. 2015). Errors include mispronunciations or substitutions, omissions, and hesitations of more than three seconds (Price et al. 2016).

While accuracy and speed are relatively objectively measured, prosody is more complex and subjective in nature and therefore is not as straightforward to describe (Jeon 2012; Pinnell et al. 1995). Nonetheless, rating scales have been developed to guide the assessment process and have been considered valid and useful for classroom use and research (Miller and Schwanenflugel 2008; Rasinski et al. 2009; Valencia et al. 2010; Young et al. 2015). Though acoustic or spectrographic analysis can provide more objective information of prosodic reading, this chapter presents two well accepted rating scales which can be used in L2 classrooms and research without any technology required.

First, the National Assessment of Educational Progress (NAEP) Oral Reading Fluency Scale, which was created by Pinnell et al. (1995) for their large-scale study

Table 22.1 NAEP oral reading fluency scale

Level	Description
4	Reads primarily in larger, meaningful phrase groups. Although some regressions, repetitions, and deviations from text may be present, these do not appear to detract from the overall structure of the story. Preservation of the author's syntax is consistent. Some of most of the story is read with expressive interpretation
3	Reads primarily in three-or four-word phrase groups. Some smaller groupings may be present. However, the majority of phrasing seems appropriate and preserves the syntax of the author. Little or no expressive interpretation is present
2	Reads primarily in two-word phrases with some three- or four-word groupings. Some word-by-word reading may be present. Word groupings may seem awkward and unrelated to larger context of sentence or passage
1	Reads primarily word-by-word. Occasional two-word or three-word phrases may occur—but these are infrequent and/or they do not preserve meaningful syntax

(Pinnell et al. 1995)

with 4th graders in the USA, is a four-point scale. It centers around three key elements of reading prosody: phrasing, adherence to syntactic structure, and expressiveness (see Table 22.1). The scale was originally intended for English-L1-speaking children, but it has been used in an L2 study (Jiang 2016) with acceptable inter-rater agreement (79% exact match, 100% adjacent match, Cohen's kappa = .70). Although the NAEP scale allows for quick assessments (Rasinski 2004b), weaknesses of this scale include a small variation of points (maximum three) among students and the possibility of some students receiving the same mark even if they read with different pause structures (Ardoin et al. 2013).

Recognizing the restriction of score range and precision issues of the NAEP scale, the Multidimensional Fluency Scale was developed and has been used for instructional and evaluative purposes. Adapted from the original rubric developed by Zutell and Rasinski (1991), the revised Multidimensional Fluency Scale (Rasinski 2004b) can be used to assess students' reading prosody on a four-point scale in four categories: expression and volume, phrasing, smoothness, and pace in reading (see Table 22.2). The summed overall ratings can range from 4 to 16. This scale has been shown to be a valid and reliable measure in a number of L1 studies (e.g., Veenendaal et al. 2015; Young et al. 2015) and one L2 study (Khor et al. 2014). Moreover, Moser et al. (2014) substantiated the reliability of the Multidimensional Fluency Scale by simultaneously estimating the effects of multiple sources of error variability, such as the passage, raters, students, and rating occasions. While subjective judgments are required to assess reading prosody when using a rating scale, which has its shortcomings (Kuhn et al. 2010), Moser and collaborators (2014) showed that it can still be done reliably.

Table 22.2 Multidimensional fluency scale

Dimension	1	2	3	4
Expression and Volume	Reads with little expression or enthusiasm in voice. Reads words as if simply to get them out. Little sense of trying to make text sound like natural language. Tends to read in a quiet voice	Some expression. Begins to use voice to make text sound like natural language in some areas of the text, but not others. Focus remains largely on saying the words. Still reads in a quiet voice	Sounds like natural language throughout the better part of the passage. Occasionally slips into expressionless reading. Voice volume is generally appropriate throughout the text	Reads with good expression and enthusiasm throughout the text. Sounds like natural language. Reader is able to vary expression and volume to match his/her interpretation of the passage
Phrasing	Monotonic with little sense of phrase boundaries, frequent word-by-word reading	Frequent two- and three-word phrases giving the impression of choppy reading; improper stress and intonation that fail to mark ends of sentences and clauses	Mixture of run-ons, mid-sentence pauses for breath, and possibly some choppiness; reasonable stress/intonation	Generally well phrased, mostly in clause and sentence units, with adequate attention to expression
Smoothness	Frequent extended pauses, hesitations, false starts, sound-outs, repetitions, and/or multiple attempts	Several "rough spots" in text where extended pauses, hesitations, etc., are more frequent and disruptive	Occasional breaks in smoothness caused by difficulties with specific words and/or structures	Generally smooth reading with some breaks, but word and structure difficulties are resolved quickly, usually through self-correction
Pace	Slow and laborious	Moderately slow	Uneven mixture of fast and slow reading	Consistently conversational

(Rasinski 2004b)

22.4 Insights Gained

Incorporating prosody into assessment brings richer insight into reading fluency from instruction as well as research perspectives. Through the use of the prosody rubrics in conjunction with WCPM, the concept of reading fluency is made more transparent for both students and teachers. Such assessment practice can guard against unbalanced reading fluency instruction (Valencia et al. 2010). In fact, it allows teachers to identify students' strengths and weaknesses in not only quantitative aspects (accuracy and speed) but also qualitative aspects of fluency. That is, effective instruction and interventions can be designed and guided based on the diagnostic information specific to each of the components of reading fluency. Because specific behavioral indicators of prosodic reading are included in the descriptors in the rubrics, students themselves can also learn to make sense of the descriptors so that they can evaluate and develop awareness of their own reading fluency (Rasinski 2004b).

With the attention to reading prosody on the rise, more and more research has shown its relevance and importance in reading fluency as well as comprehension. A growing body of research has shown that reading with appropriate prosody can also predict and assist comprehension (e.g., Benjamin and Schwanenflugel 2010; Jiang 2016). Some studies have found that accuracy and speed are more important aspects of reading fluency in younger, beginning readers, whereas reading prosody becomes more prominent in older, more advanced readers (Valencia et al. 2010; Veenendaal et al. 2015). Though more research with L2 readers is needed to confirm the contribution of prosody to L2 reading comprehension, such findings offer useful implications and future directions for L2 research to better inform L2 reading fluency instruction and assessment.

22.5 Conclusion: Implications for Test Users

In support of a comprehensive assessment approach derived from the (re)conceptualization of the construct, this chapter addressed issues caused by using WCPM alone as a reading fluency measure. Potential issues included reducing the construct of reading fluency and implementing unbalanced instruction, which can lead to distorted views of reading fluency. Though mostly in L1 contexts, research has shown that expanding reading fluency assessment to include all three components (accuracy, speed, and prosody) is not only informative for both students and assessors but also more appropriate. Despite the qualitative judgment involved, reading prosody assessment can be done reliably, as shown in multiple studies.

To foster reading fluency in its complete form that can guide effective instruction and avoid misconception of reading fluency, L2 researchers and educators are encouraged to include the prosody aspect in their assessment. As reading fluency has been gaining more attention in L2 contexts, it is critical to establish assessment that can shape and drive instruction to best serve L2 students. Instead of confining reading fluency assessment to WCPM, which can lead to an overemphasis on accuracy and speed, assessment practices that endorse reading prosody in conjunction with WCPM can keep construction of meaning central to reading and provide additional insight into L2 students' reading fluency development. Furthermore, use of the prosody assessment tools such as the rubrics presented in this chapter could lead to refining and modifying them to meet the specific needs of L2 reading fluency assessment.

Acknowledgements I wish to thank Dr. Fredricka Stoller for her support and feedback on this chapter and Jesse Conway for his helpful comments.

References

Ardoin, S. P., Morena, L. S., Binder, K. S., & Foster, T. E. (2013). Examining the impact of feedback and repeated readings on oral reading fluency: Let's not forget prosody. *School Psychology Quarterly, 28,* 391–404.

Benjamin, R. G., & Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly, 45,* 388–404.

Deeney, T. A. (2010). One-minute fluency measures: Mixed messages in assessment and instruction. *The Reading Teacher, 63,* 440–450.

Deeney, T. A., & Shim, M. K. (2016). Teachers' and students' views of reading fluency: Issues of consequential validity in adopting one-minute reading fluency assessments. *Assessment for Effective Intervention, 41,* 109–126.

Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory into Practice, 30,* 165–175.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* New York, NY: Cambridge University Press.

Grabe, W. (2010). Fluency in reading—Thirty-five years later. *Reading in a Foreign Language, 22,* 71–83.

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd ed.). New York, NY: Routledge.

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58,* 702–714.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95,* 719–729.

Jeon, E. H. (2012). Oral reading fluency in second language reading. *Reading in a Foreign Language, 24,* 186–208.

Jiang, X. (2016). The role of oral reading fluency in ESL reading comprehension among learners of different first language backgrounds. *The Reading Matrix, 16,* 227–242.

Khor, C. P., Low, H. M., & Lee, L. W. (2014). Relationship between oral reading fluency and reading comprehension among ESL students. *GEMA Online® Journal of Language Studies, 14*, 19–32. https://doi.org/10.17576/gema-2014-1403-02.

Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology, 100,* 310–321.

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45,* 230–251.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95,* 3–21.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293–323.

Lems, K. (2003). *A study of adult ESL oral reading fluency and silent reading comprehension.* Unpublished doctoral dissertation, National Louis University, Chicago, IL.

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly, 43,* 336–354.

Moser, G. P., Sudweeks, R. R., Morrison, T. G., & Wilcox, B. (2014). Reliability of children's expressive reading. *Reading Psychology, 35,* 58–79.

Pikulski, J. J., & Chard, D. J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher, 58,* 510–519.

Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud: Data from NAEP's integrated reading performance record (IRPR) at grade 4 (NCES 95–726).* Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Price, K. W., Meisinger, E. B., Louwerse, M. M., & D'Mello, S. (2016). The contributions of oral and silent reading fluency to reading comprehension. *Reading Psychology, 37,* 167–201.

Rasinski, T. V. (2004a). Creating fluent readers. *Educational Leadership, 61,* 46–51.

Rasinski, T. V. (2004b). *Assessing reading fluency*. Honolulu, HI: Pacific Resources for Education and Learning.

Rasinski, T. V. (2012). Why reading fluency should be hot! *The Reading Teacher, 65*, 516–522.

Rasinski, T. V., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades? *Literacy Research and Instruction, 48*, 350–361.

Samuels, S. J. (2006). Toward a model of reading fluency. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 24–46). Newark, NJ: International Reading Association.

Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly, 42*, 563–566.

Schrauben, J. E. (2010). Prosody's contribution to fluency: An examination of the theory of automatic information processing. *Reading Psychology, 31,* 82–92.

Schwanenflugel, P. J., Westmoreland, M. R., & Benjamin, R. G. (2015). Reading fluency skill and the prosodic marking of linguistic focus. *Reading and Writing, 28,* 9–30.

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45,* 270–291.

Veenendaal, N. J., Groen, M. A., & Verhoeven, L. (2015). What oral text reading fluency can reveal about reading comprehension. *Journal of Research in Reading, 38,* 213–225.

Young, C., Mohr, K. A., & Rasinski, T. V. (2015). Reading together: A successful reading fluency intervention. *Literacy Research and Instruction, 54,* 67–81.

Zutell, J., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory into Practice, 30,* 211–217.

Chapter 23
(Re)Creating Listening Source Texts for a High-Stakes Standardized English Test at a Vietnamese University: Abandoning the Search in Vain

Xuan Minh Ngo

Abstract This chapter argues for viewing listening source texts (LSTs) as a hybrid genre and shares a practical approach to developing LSTs for a high-stakes standardized English test at a Vietnamese university. Presented as a series of autoethnographic vignettes, the paper details the enormous challenges facing the author in his quest for "perfect" authentic texts that had to fit both text and item specifications. These problems were not surmounted until the author, inspired by empirical item writing research, had taken a more liberal approach which involves substantial editing or even creating LSTs from scratch to match test items. The author's success in adopting this liberal approach has led him to support the view of LSTs as a hybrid genre with features derived from both the exams they are situated in and real-world texts in the target language use domain. The chapter concludes with implications for item writer training, and simultaneously calls for caution in using the special genre of LSTs.

23.1 Introduction: Purpose and Testing Context

There is little dispute about the challenges of designing listening tests (Buck 2001; Green 2014; Green and Hawkey 2012) which involves both writing items and adapting or creating the accompanying source texts. Although item writing guidelines do exist in both educational assessment (Haladyna and Rodriguez 2013) and language assessment (Spaan 2007), these offer little information about how to craft suitable listening source texts (henceforth LSTs). Although some pioneering studies have examined the practice of writing items for reading and listening tests, these either were conducted in an English-as-a-second-language (ESL) environment (Kim et al. 2010) or involved writers working for a well-resourced international exam board (Green and Hawkey 2012; Salisbury 2005). In other words, there has been a lack of research investigating the development of LSTs in the English-as-a-foreign-language (EFL) context, where the vast majority of listening tests are administered.

X. M. Ngo (✉)
School of Languages and Cultures, The University of Queensland, Brisbane QLD 4072, Australia
e-mail: x.ngo@uq.edu.au

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_23

To fill this gap, this chapter shares a practical approach to designing LSTs for a high-stakes standardized English test at Babel (a pseudonym), a public university specialized in foreign language education in Vietnam, a Southeast Asian developing nation with a booming EFL teaching and learning scene (Ngo 2018a, b). As the chapter draws on my own experiences, parts of it are a series of autoethnographic vignettes, i.e., short stories of self. Moreover, this increasingly popular qualitative approach is a legitimate choice since it enables researchers to "speak into … literature from a place of experience" (Stahlke Wall 2018, p. 1) and in this case contribute to the limited literature on LSTs. In the first vignette below, more information about the testing context is provided.

Vignette 1: How It All Started My venture into item writing started one day after I returned to Vietnam, my home country, following my Master's course in Australia. When I emailed my direct supervisor at my Vietnamese institution—Babel—she promptly got back with the following message.

> Welcome back, Minh!
>
> …Our university has been working on a CEFR-aligned test development project. The test will be administered first on our own students, and then nationwide. I'm Head of the Listening group and we've finished developing the test specifications. I'm looking for some well-qualified item writers! It's not an easy job, so I need people with MA degrees like you :P.

Enthusiastic about the chance to apply the assessment knowledge and skills freshly gained from Australia, I immediately jumped at this offer. After numerous team meetings, I gradually grasped the big picture behind the Babel Test of English (a pseudonym—henceforth BTE) development project. It turned out that BTE was commissioned by the management of Project 2020, "the most significant and ambitious foreign language reform in modern Vietnam" (Ngo 2018a, p. 48; see also Ngo 2017; Ngo 2018b). Since Project 2020 plans to conduct large-scale assessment of Vietnamese foreign language learners (Vietnamese Government 2008), the project management considers the development of standardized tests such as BTE a top priority. This explained why Babel's top university leaders were directly involved in managing the BTE development project, whose goal was to develop a battery of four skill-based subtests (Listening, Reading, Writing, Speaking) aligned with Levels B1 to C1 in the Common European Framework of Reference (CEFR) (Council of Europe 2001). This test would serve as the mandatory graduation exam for students initially at Babel and then nationwide when BTE was officially approved by the Ministry of Education and Training. See Ngo (2018a) for a detailed account of mandatory exit testing for Vietnamese university students.

Fast forward to today, I have worked as a BTE listening item writer for over four years, during which time I have produced over 50 listening tests. When I started this job, I would never have dreamed of becoming such a prolific writer due to the problem described in the next section.

23.2 Testing Problem Encountered

Vignette 2: A Search in Vain "Believe me, I know exactly where to find the materials for this part of the test …" I reassured other team members and volunteered to write what was considered the hardest set: a lecture at Level C1 followed by five multiple-choice items. Following Bachman and Palmer's (1996) framework which emphasizes the need to sample real-world texts from the target language use (TLU) domain, I immediately thought of Massive Open Online Courses (MOOCs) platforms. These sources were particularly attractive for me because they offered hundreds of introductory courses on numerous disciplines, and most importantly, each lecture was separated into short clips, each lasting at most 10 min. However, this apparently logical decision proved problematic because I had to closely adhere to a detailed set of test specifications or "generative blueprints for test design" (Davidson and Lynch 2002, p. 1) as follows (see Table 23.1).

For starters, I had great difficulty locating short clips that could be adapted into lecture extracts lasting 3–3.5 min, given that most MOOC videos lasted 7–10 min. Additionally, the materials had to be abstract, yet non-technical, and of interest to the general audience. Hence, after three days of browsing numerous videos, I managed to shortlist five extracts in the first lessons of introductory-level courses in philosophy, history, finance, cinema, and media. However, the real dilemma started when I began crafting items. As seen in Table 23.1, I had to create two questions focusing on important details, each with one keyed (correct) option and three plausible distractors following guidelines such as those of Haladyna and Rodriguez (2013) and Spaan (2007). The BTE specifications (Babel University 2016, p. 25) also included the following requirements:

The distractors should

- Repeat some key words but convey a different message
- Use words that sound similar to key words

Table 23.1 BTE Part 3 Text Specifications

Nature of information	Abstract
Domain (CEFR)	Educational
Interaction	Monologue
Topics (examples)	A lecture in an introductory college-level course
Text length	350-400 words
Lexical level	EVP (English Vocabulary Profile) level: $\leq$ C1
Grammatical level	high (most sentences should be complex)
Speech rate	110-170 words per minute
Attached items	Level 4 * 1: 1 item focusing on main/topical ideas Level 5 * 4: 2 items on understanding details; 1 item focusing on understanding inferences and 1 item focusing on understanding idiomatic or colloquial vocabulary

- Are [sic] incorrect or inaccurate according to the recording
- Are [sic] irrelevant (not mentioned in the recording).

> The correct answer should contain the exact detail but be significantly restructured. (For example, there is a change in word formation or voice).

Unsurprisingly, none of the shortlisted videos contained sufficient material for writing the required number of items, considering that both the correct option and two other distractors had to occur in the video. To make matters worse, only two videos contained idiomatic expressions, and instead of implying their main points, all lecturers tried to express these key ideas as clearly as possible, and even occasionally repeated and rephrased them. These facts meant that there was little relevant material for me to produce the two items testing colloquial vocabulary and understanding inferences. As expected, I eventually had to discard all selected videos and restart the text searching phase from scratch.

23.3 Resolution of the Problem

Vignette 3—No Text is "Sacred" "Oh, I've got it all wrong …" That's what flashed through my mind when I perused Salisbury's (2005) doctoral dissertation on Cambridge listening item writers' expertise development. I was particularly struck by Chapters 6 and 7, which list the "ruses" or tricks such as text-item barter, script padding, and trimming used by expert item writers to modify base texts, i.e., authentic texts, into the final listening source texts (LSTs). This extract nicely summarizes my awakening moment:

> This list shows just how 'cavalier' the writers are in their attitude towards text/script: by employing these ruses to alter the text in this fashion they are clearly not regarding base text as sacred, or items as immanent within it. The use of ruses indicates that text is a construct which can and indeed, must, be altered to make the test workable. (Salisbury 2005, p. 221)

Following Salisbury's (2005) expert writers, I began seeing base texts as catalysts rather than "sacred" entities and adopting the item-first approach which prioritized the creation of items followed by the extensive rewriting of base texts. This proved a huge time and effort saver and enabled me to both submit my commissions on time and produce LSTs with comparable length and levels of lexical and grammatical complexity. However, my practice over time gradually evolved into a recursive rather than linear process, and instead of the item-first approach, I tended to move back and forth between composing items and texts. To reflect the recursive nature of and the priority given to items over texts in my item writing workflow, I call it an item-centered approach (see Fig. 23.1).

In this approach, I normally started with the test specifications and then searched for promising base texts. However, instead of finding the "perfect" one, I simply chose a base text whose topic was at the appropriate level of abstraction as stipulated in the specifications, and which contained sufficient material for half of the required number

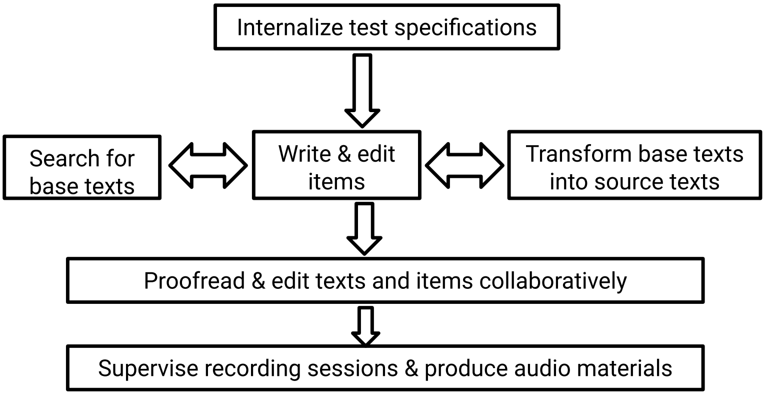

Internalize test specifications
Search for base texts
Write & edit items
Transform base texts into source texts
Proofread & edit texts and items collaboratively
Supervise recording sessions & produce audio materials

Fig. 23.1 An item-centered approach to writing listening test items and source texts

of items. Where necessary, I also integrated information from two or three sources into the final LST. Afterwards, I would write items, starting with the stem, then the key, and three distractors. I attempted to observe item writing rules (Haladyna and Rodriguez 2013; Spaan 2007), making sure all elements were as concisely expressed as possible, and the options were parallel both grammatically and lexically. Most importantly, I meticulously ensured all options' plausibility so that no candidate could choose the correct answer simply by using background knowledge. It should be noted that while the stem and the key were closely based on the base text, the distractors were often created from my background knowledge and my recollection of common mistakes made by my students. Subsequently, I rewrote the base text extensively, transforming it into the LST which had to include both the key and distractors. This step frequently involved synthesizing information from various sources and inserting features of oral speech such as fillers, false starts, pauses, and self-corrections (Buck 2001). As shown in Fig. 23.1, while I was writing an LST, I frequently reverted to editing the accompanying items to ensure a logical text-item relationship in the same set. Although limited resources and the confidential nature of my work did not allow pre-editing and editing meetings similar to those in Salisbury (2005) or Green and Hawkey (2012), the exam board I worked for also recruited native English speakers to act as proofreaders and voice actors. I found collaboration with them very fruitful as their linguistic background enabled them to identify and rectify unnatural segments in my LSTs. During the recording session, I also encouraged voice actors to pause and discuss whenever they spotted portions of LSTs that needed editing for better authenticity.

23.4 Insight Gained: Listening Source Texts as a Hybrid Genre

As shown in Sect. 23.2, my initial approach of using real-world texts seems theoretically sound and is still advocated in contemporary literature (Green 2014; Liao et al. 2018) but did not work well in reality. This conundrum, interestingly, is a common problem facing standardized test development teams (Buck 2001; Green 2014) that have to strictly comply with test specifications and produce multiple test forms of comparable characteristics, leading to most, if not all high-stakes exam boards using scripted rather than authentic texts (Green 2014). This absence of authentic texts in language tests is hardly surprising because authentic texts do not possess desirable features of source texts required for standardized exams. As item writers in Green and Hawkey (2012, p. 126) put it:

> From the item writer's perspective, informed by test guidelines and experience, IELTS texts need to be propositionally dense and avoid repetition. But, unlike their sources, they do not need to attract the reader's attention, to cross-refer to other texts, to locate themselves within an academic discipline or to underline essential information. The texts need to be accessible, but not controversial, and above all to present enough new information (distributed evenly through the text) to support large numbers of test items and to distract the less proficient test taker.

Considering the substantial differences between the ideal exam source text and the corresponding text in the TLU domain (e.g., a lecture in a listening test versus a real college lecture), I would argue in line with Peirce (1992) and Norton (2006) that LSTs constitute a hybrid genre with features derived from both the exams they are situated in and real-world texts in the TLU domain. The term "genre" here is not synonymous with the layman understanding of genre as a text type but should be construed as "constituted within and by a particular social occasion that has a conventionalized structure, and which functions within the context of larger institutional and social processes" (Norton 2006, p. 94).

The view of LSTs as "a socially constituted genre" (Norton 2006, p. 94), I believe, offers some major benefits. First, this perspective releases item writers from the burden of having to search in futility for authentic texts that strictly meet test specifications, which also means increased productivity as writers can concentrate on their core tasks, i.e., writing items and (re)creating LSTs. Second, lifting LSTs to genre status may lead to more research directed toward uncovering the characteristics of this under-studied genre, which will potentially inform item writing practice and test design. A promising direction appears to be the use of natural language processing (NLP) software to specify the lexical and syntactic features of LSTs as opposed to authentic texts (Green 2014; Green et al. 2010). However, since most NLP applications focus on measuring written texts' readability, substantial research will be needed to develop programs for analyzing spoken texts for listening tests. Third, viewing LSTs as a hybrid genre does not mean accepting these texts as devoid of oral speech features. On the contrary, the acknowledgment that LSTs are partly derived from spoken texts in the TLU domain will compel item writers to integrate more oral features into LSTs or to "authenticate" LSTs (Liao et al. 2018, p. 7).

Finally, on a more critical note, considering LTSs as related to but distinct from authentic texts would serve to highlight the limits of standardized testing (Jenkins and Leung 2019; Shohamy 2017) in assessing candidates' ability to communicate in real-life contexts, hence discouraging test users from relying exclusively on standardized testing when making high-stakes decisions about test takers' listening ability.

23.5 Conclusion: Implications for Test Users

While LSTs should ideally replicate real-life situations in the TLU domain, this can hardly be achieved at least with our current state of knowledge (Green 2014). Hence, it would be more productive to regard LSTs in standardized tests as a hybrid genre with characteristics originating from the exam they are situated in and their original TLU domain. Given this view, it seems more sensible for item writers to (re)create LSTs to fit the items instead of searching in futility for perfect authentic sources that match both the text and item specifications. Admittedly, this is not an ideal solution, but seems a necessary compromise required in language testing practice (Bachman and Palmer 1996).

As LSTs can be viewed as a hybrid genre, I am convinced that a promising model for training item writers is the genre pedagogy cycle (Hyland 2018). Accordingly, in the first stage, the genre purposes and settings of use should be clearly articulated. The best source for such information would arguably be the test specifications, which trainees should be encouraged to refer to during the item writing process. Next, in the modeling phase, trainees should be afforded the chance to analyze samples of the target genre and learn about its features. In this case, the view of LSTs as a hybrid genre means that prospective item writers need to do extensive analysis of both representative LSTs and authentic texts to appreciate their final products' desirable features. To assist trainees' intuition, NLP packages (Green 2014; Green et al. 2010) may be utilized; nevertheless, as previously indicated, care must be taken since most NLP programs are written for analyzing written rather than spoken language. In the third stage, trainees are to complete guided practice tasks whose design should be informed by empirical findings about item writing and standard guidelines. In other words, trainees should be familiarized with ruses or tricks used by expert item writers (Green and Hawkey 2012; Salisbury 2005) and standard rules for writing quality items (Haladyna and Rodriguez 2013; Spaan 2007). As a hybrid genre, LSTs are meant to carry features typical of oral speech, so trainees should be introduced to techniques to authenticate scripted texts (Liao et al. 2018) to bridge the gap between real-life communication and tests (Bachman and Palmer 1996). Finally, they should be required to write items independently and receive both qualitative and quantitative feedback on their work. The qualitative feedback may be given in pre-editing or editing meetings, as in Salisbury (2005) and Green and Hawkey (2012), and ideally, where pre-testing is possible, quantitative feedback should be provided in the form of item difficulty and discrimination indices (Ingham 2008).

On the other hand, as LSTs are distinct from authentic texts, candidates' performance on a standardized listening test should not be considered a true reflection of their ability to understand real-life spoken speech (Jenkins and Leung 2019). Hence, test users including but not limited to language program directors, institutions, and policy makers are advised to consider multiple sources of information especially when making high-stakes decisions. In practical terms, this implication means that various assessment methods such as self-assessment, teacher observation, and portfolio assessment (Ngo 2015; Shohamy 2017) need to be employed to complement standardized tests.

On balance, while I do appreciate the need for alternatives to standardized testing (Jenkins and Leung 2019), I believe that there remains a place for large-scale standardized tests of listening, especially in low-resourced countries where millions of students' English proficiency must be assessed in a cost-effective manner. As an example, in my home country of Vietnam, where the grammar-translation method remains the norm (Ngo 2018a, b), a test of listening in itself, especially a locally developed one like BTE, is already a major innovation. Hence, I hope that the insight and implications in this chapter will embolden exam development teams in similar contexts to create their own listening tests, thus motivating their students to move beyond learning grammar.

References

Babel University. (2016). *Babel Test of English (BTE) specifications for item writers*. Hanoi, Vietnam: Babel University.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Council of Europe. (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, MA: Yale University Press.

Green, A. (2014). Adapting or developing source material for listening and reading tests. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed.). Hoboken, NJ: John Wiley & Sons Inc.

Green, A., & Hawkey, R. (2012). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing, 29*(1), 109–129. https://doi.org/10.1177/0265532211413445.

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing, 27*(2), 191–211. https://doi.org/10.1177/0265532209349471.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hyland, K. (2018). Genre and second language writing. *The TESOL encyclopedia of English language teaching*. https://doi.org/10.1002/9781118784235.eelt0535.

Ingham, K. (2008). The Cambridge ESOL approach to item writer training: The case of ICFE listening. *Research Notes, 32,* 5–9.

Jenkins, J., & Leung, C. (2019). From mythical 'standard' to standard reality: The need for alternatives to standardized English language tests. *Language Teaching, 52*(1), 86–110. https://doi.org/10.1017/S0261444818000307.

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., & Roullion, V. (2010). A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual item writers' characteristics. *Language Assessment Quarterly, 7*(2), 160–174. https://doi.org/10.1080/15434300903473989.

Liao, Y.-F., Wagner, E., & Wagner, S. (2018). Test-takers' attitudes and beliefs about the spoken texts used in EFL listening tests. *English Teaching & Learning, 42*(3), 227–246. https://doi.org/10.1007/s42321-018-0013-5.

Ngo, X.M. (2015). *Ethics in language testing—International practices and implications for Vietnam.* Paper presented at the Second VietTESOL Conference, Hanoi, Vietnam.

Ngo, X. M. (2017). Diffusion of the CEFR among Vietnamese teachers: A mixed methods investigation. *Asian EFL Journal, 19*(1), 7–32.

Ngo, X. M. (2018a). Sociopolitical contexts of EFL writing assessment in Vietnam: Impact of a national project. In T. Ruecker & D. Crusan (Eds.), *The politics of English second language writing assessment in global contexts*. New York, NY: Routledge.

Ngo, X. M. (2018b). A sociocultural perspective on second language writing teacher cognition: A Vietnamese teacher's narrative. *System, 78,* 79–90. https://doi.org/10.1016/j.system.2018.08.002.

Norton, B. (2006). Not an afterthought: Authoring a text on adult ESOL. *Linguistics and Education, 17*(1), 91–96. https://doi.org/10.1016/j.linged.2006.08.005.

Peirce, B. N. (1992). Demystifying the TOEFL Reading Test. *TESOL Quarterly, 26*(4), 665–691. https://doi.org/10.2307/3586868.

Salisbury, K. (2005). *The edge of expertise? Towards an understanding of listening test item writing as professional practice (Unpublished doctoral dissertation).* London: King's College.

Shohamy, E. (2017). Critical language testing. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 1–15). Cham: Springer.

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly, 4*(3), 279–293. https://doi.org/10.1080/15434300701462937.

Stahlke Wall, S. (2018). Reflection/commentary on a past article: "Easier said than done: Writing an autoethnography." *International Journal of Qualitative Methods, 17*(1), 1–2. https://doi.org/10.1177/1609406918788249.

Vietnamese Government. (2008). *Teaching and learning foreign languages in the national education system 2008– 2020 (Decision 1400/QĐ-TTg).* Hanoi, Vietnam.

Chapter 24
The Oral Standardized English Proficiency Test: Opportunities Provided and Challenges Overcome in an Egyptian Context

Deena Boraie and Ramy Shabara

Abstract The School of Continuing Education (SCE) of the American University in Cairo (AUC), Egypt produced an oral standardized test called the OSEPT (Oral Standardized English Proficiency Test) aligned with the Common European Framework of Reference (CEFR) and used to assess students' proficiency as well as place students in SCE's English oral communication programs. The OSEPT is based on a direct format where two test takers are assessed by two trained speaking examiners. It comprises five easy-to-challenging tasks covering five CEFR levels, e.g., A1, A2, B1, B2 and C1, and requires 15–20 min to complete. The focus of this chapter is to describe a testing problem encountered in the design of the fifth oral task and how it was resolved. The insight gained from the experience and implications for test users are also discussed.

24.1 Introduction: Purpose and Testing Context

The School of Continuing Education (SCE) of the American University in Cairo, Egypt has a large English language program where approximately 15,000 students are enrolled per year in a variety of English programs: general English, oral communication, English for young learners, English for specific purposes, and international test preparation courses. SCE designed and produced a standardized test called the SEPT (Standardized English Proficiency Test) used to assess students' proficiency as well as to place students in SCE's English programs. The SEPT assesses listening, reading, and writing skills and is aligned with the Common European Framework of Reference (CEFR).

The need emerged to design an oral test, and accordingly, the testing department designed the OSEPT (Oral Standardized English Proficiency Test) also aligned with

D. Boraie (✉) · R. Shabara
The American University, Cairo, Egypt
e-mail: dboraie@aucegypt.edu

R. Shabara
e-mail: ramy.shabara@aucegypt.edu

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_24

the CEFR, to measure the speaking proficiency of SCE students for whom English is a second or foreign language. The scores obtained from the OSEPT are used for placement decisions of SCE learners into the different levels of the SCE CEFR-aligned English Language Conversation Program or other oral communication programs. The OSEPT is based on a direct format where two test takers are assessed by two trained speaking examiners. One examiner acts as the interlocutor and the second is the evaluator. The oral test comprises five easy-to-challenging tasks covering five CEFR levels; e.g., A1, A2, B1, B2 and C1, and requires 15–20 minutes to complete.

Three types of stimuli are manipulated in this test. The first stimulus is verbal. Though it is used throughout the test, it is used particularly in Task 1 and Task 2. The second one is visual. It is used in Task 3. The third one is written and used in Task 4 and Task 5. "In the test as a whole, the construct consists of three main interactional modes: question and answer (examiner to candidate/s), goal-oriented conversation (candidate to candidate) and long turn (candidate to examiner and other candidate)," explain Macqueen and Harding (2009, p. 468).

The first two tasks (A1–A2) require four to six minutes to complete (each needs two–three minutes). They tackle factual/personal and experience-based questions, respectively. They adopt a one-on-one format where the interlocutor, using scripted questions, asks each test taker individually questions such as "What is your favorite food? How did you spend your last holiday?" The third task (A2–B1) requires the two test takers to speak together to describe a picture. In this collaborative task, the two test takers are allowed to ask each other questions and/or state their opinions about the picture in two to three minutes. The fourth task (B1–B2) is a monologue-based one where each test taker is given a written prompt about which they are asked to speak for 1–1:30 minutes after a minute of preparation. The prompt is guided by key words to help test takers elaborate on the idea given. The fifth task (B2–C1) is a debate where the two test takers are given two written opposite ideas and asked to debate about them.

The design of these tasks using the paired mode has several advantages associated with this approach. This format allows test takers to feel relaxed while interacting and provides them with opportunities of varied patterns of interaction (Saville and Hargreaves 1999). It also allows them to demonstrate their language proficiency by providing them with "opportunities to employ a wider range of speech events…" (Foot 1999, p. 39) and thus rich and authentic speech samples and functions can be easily elicited (Brooks 2009; Taylor 2003). Moreover, it has positive washback in the classroom by giving the chance to students to practice interaction while preparing for the test (Saville and Hargreaves 1999; Van Moere 2006). This format has also been reported to be more authentic than one-on-one test formats, where it resembles natural conversation and what happens in the classroom of various peer interactions (Kasper 2013; Sandlund et al. 2016). Testing using the paired mode has been found to "[promote] and [improve] students' interactional competence, creating students' co-constructed discourse, and providing insights for better scale development and rater training" (Prasetyo 2018, p. 105). Other merits of such a format include cost effectiveness and time efficiency in administration (Ducasse and Brown 2009). As reported by Ffrench (2003), Van Moere (2006), and East (2015), paired formats are much more preferred by test takers and teachers than one-on-one formats.

24.2 Testing Problem Encountered

Dyadic speaking assessments reflect several complicated interactions between candidate, task, and interlocutor (Davis 2009; O'Sullivan 2002). Accordingly, despite numerous merits, this type of test format has been subject to dissection because of the different possible factors resulting from such patterns of interaction that may, consequently, affect validity (Macqueen and Harding 2009). These factors include "the potential for breakdowns in communication between candidates [with different proficiency levels]" (Davis 2009; Macqueen and Harding 2009, p. 470), the possible merits/demerits of being paired with a familiar/unfamiliar test taker (Foot 1999) and interlocutor (Davis 2009; Son 2016), and the impacts of such variables as task topic and difficulty, personality, extraversion, age, gender, and test preparation (Berry 2007; Fulcher and Reiter 2003; Leaper and Riazi 2013; Lumley and O'Sullivan 2005; Macqueen and Harding 2009; Nakatsuhara 2011; Norton 2005).

The testing challenge encountered was that a large number of students were misplaced into SCE conversation programs and teachers had to send them back to the testing department for reassessment. On analyzing the scores, it was found that in all these cases of misplacement, a breakdown in communication had occurred between test takers of different proficiency levels, particularly in the fifth task of the OSEPT. The fifth task (at the B2–C1 CEFR level) is a debate where the two test takers are given two cards (each test taker is given a card) with opposite positions on a topic and are required to debate them. The first test taker is asked to provide an argument on a topic in one minute, and the second test taker is asked to listen carefully then make a rebuttal for 30 seconds. The second test taker is then asked to argue the opposition position while the first test taker listens carefully and then is required to refute what was said by the first test taker. The problem of communication breakdown between test takers of different proficiency levels was a serious validity threat because test takers were unable to appropriately complete the task and test scores may have been significantly affected. Several instances of communication breakdown were observed when two test takers were of different proficiency levels. The first situation occurred when the high-proficient test taker started arguing his/her topic. In this case, the low-proficient test taker was unable to understand the argument and consequently could not produce the rebuttal as required. Another case was when a low-proficient test taker found his/her topic difficult and could not initiate talking about it and consequently the high-proficient test taker found nothing to refute. A third situation observed was when the low-proficient test taker started to argue his/her topic and found it too difficult. This communication breakdown resulted in ending the test before the due time, leading to inaccurate scores for one or both test takers.

24.3 Solution of the Problem

To solve the testing challenge encountered in the design of Task 5 of the OSEPT, the test developers decided to change the design of the task. Accordingly, Task 5 was redesigned and changed to a new format to minimize communication breakdown. It was decided to move away from a written prompt and use a verbal prompt like Task 1 and Task 2. The verbal prompt consists of a set of five controversial, open-ended questions designed to be completed in about five to six minutes where test takers could agree, disagree, and/or negotiate toward a goal. Each question takes approximately one minute to answer, and the questions range from easy to challenging in terms of the degree of abstraction. The number of questions covered is determined by test takers' proficiency within the five–six minutes allotted to the task. The first test taker is asked a question and then provides a response. The second test taker is asked to comment and express an opinion based on his/her understanding of the response of the first test taker. In case there is a breakdown in answering the question, the examiner redirects the same question to the second test taker and then based on the answer, goes back to the first test taker and asks for a comment or an opinion. If there is a breakdown in communication with the second test taker, the examiner uses a series of prompts to elicit a response. If test takers are unable to make comments or express an opinion, the examiners move to the next question and so on. To assess initiation, developing ideas and goals-oriented negotiations, the two test takers may be asked one question at the same time. This throw-out technique also allows for assessing various types of interaction and scaffolding.

The following is an example of Task 5 between two test takers, X and Y:

1	The Examiner to Test Taker X	:	"Do you think community service is important? (Why? /Why not?)"
	The Examiner to Test Taker Y	:	"What do you think, Y? (Why? /Why not?)"
2	The Examiner to Test Taker Y	:	"Some people think that it is better to volunteer your time than donate your money. What do you think? (Why? /Why not?)"
	The Examiner to Test Taker X	:	"What about you, X? (Why? /Why not?)"
3	The Examiner to Test Takers X & Y	:	"It is said that community service can be a legal punishment for certain crimes instead of jail? Do you agree?"

24.4 Insights Gained

As a result of redesigning Task 5, misplacements into SCE's conversation classes were reduced and helped us deal with the challenge of the different proficiency levels of test takers. This experience offers some insights into test validity. The solution we came up with did not change the proficiency levels of test takers but enabled us to assess their proficiency more accurately. Thus, test scores were affected by various factors other than test takers' proficiency levels. The type of tasks and the type of interaction/performance required affect test scores rather than the proficiency levels of test takers. In the new task, test takers were assessed in pairs, but the performance of one test taker had less impact on the performance of the second test taker. Thus, the difference in proficiency level was minimized in the effect on their scores. This task also adds a new dimension to test validity, fairness, and usefulness, which is interactiveness. Interactiveness shifts the common emphasis from the cognitive and individualist view of communicative competence suggested by Bachman and Palmer (1996) to the concept of interactional competence that focuses more on such "resources [as] turn-taking, appropriate use of linguistic register, and the ability to recognize and signal boundaries of communicative events" (Hellermann 2006, p. 378). The redesigned task (Task 5) enabled us to assess test takers in pairs at different proficiency levels more accurately while at the same time assessing interactional competence. Consequently, it contributes to the authenticity aspect of the test and adds to the validity evidence of the test in measuring what it is supposed to measure.

24.5 Conclusion: Implications for Test Users

In light of this experience, it is clear that the balance between the practicalities of testing large numbers of candidates and ensuring that the oral test tasks are measuring the intended construct of oral language use is not easy and requires continuous monitoring and evaluation. Several implications for future research are suggested for test users. First, further validation work on the impact of different kinds of pairing is recommended. Second, procedures on pairing test takers should be investigated and followed up to minimize the effect of the discrepancy of test takers' proficiency levels on their test scores (Macqueen and Harding 2009). Third, more investigations on the influence of test takers' gender and relationships on their performance are needed. Fourth, in-depth examinations of the impact of oral assessment tasks in terms of types, structures, and instructions on oral production are also required.

References

Bachman, L. F., & Palmer, A. (1996): *Language testing in practice.* Oxford: Oxford University Press.

Berry, V. (2007). *Personality differences and oral test performance.* Frankfurt: Peter Lang.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 20,* 89–110. https://doi.org/10.1177/0265532209104666.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367–396. https://doi.org/10.1177/0265532209104667.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing, 26,* 423–443.

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing, 32,* 101–120. https://doi.org/10.1177/0265532209104669.

Ffrench, A. (2003). The change process at the paper level. Paper 5, Speaking. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge proficiency in English examination 1913–2002* (pp. 367–471). Cambridge: Cambridge University Press.

Foot, M. (1999). Relaxing in pairs. *ELT Journal, 35*(1), 36–41. https://doi.org/10.1093/elt.53.1.36.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing, 20,* 321–344. https://doi.org/10.1191/0265532203lt259oa.

Hellermann, J. (2006). Classroom interactive practices for developing L2 literacy: A microethnographic study of two beginning adult learners of English. *Applied Linguistics, 27,* 377–404. https://doi.org/10.1093/applin/ami052.

Kasper, G. (2013). Managing task uptake in oral proficiency interviews. In S. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 258–287). New York, NY: Palgrave Macmillan.

Leaper, D. A., & Riazi, M. (2013). The influence of prompt on group oral tests. *Language Testing, 31*(2), 177–204. https://doi.org/10.1177/0265532213498237.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*, 415–436. https://doi.org/10.1191/0265532205lt303oa.

Macqueen, S., & Harding, L. (2009). Review of the certificate of proficiency in English (CPE) speaking test. *Language Testing, 26,* 467–475. https://doi.org/10.1177/0265532209104671.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28,* 483–508. https://doi.org/10.1177/0265532211398110.

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal, 59*(4), 287–297. https://doi.org/10.1093/elt/cci057.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277–295. https://doi.org/10.1191/0265532202lt205oa.

Prasetyo, A. H. (2018). Paired oral tests: A literature review. *LLT Journal, 21,* 105–110. https://doi.org/10.24071/llt.2018.Suppl2110.

Sandlund, E., Sundqvist, P., & Nyroos, L. (2016). Testing L2 talk: A review of empirical studies on second-language oral proficiency testing. *Language and Linguistics Compass, 10*(1), 14–29. https://doi.org/10.1111/lnc3.12174.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal, 53,* 42–51. https://doi.org/10.1093/elt/53.1.42.

Son, Y. (2016). Interaction in a paired oral assessment: Revisiting the effect of proficiency. *Papers in Language Testing and Assessment, 5*(2), 43–68.

Taylor, L. (2003, August). *The Cambridge approach to speaking assessment*. University of Cambridge Local Examinations Syndicate Research Notes, pp. 2–4.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23,* 411–440. https://doi.org/10.1191/0265532206lt336oa.

Chapter 25
Opening the Black Box: Exploring Automated Speaking Evaluation

Nahal Khabbazbashi, Jing Xu, and Evelina D. Galaczi

Abstract The rapid advances in speech processing and machine learning technologies have attracted language testers' strong interest in developing automated speaking assessment in which candidate responses are scored by computer algorithms rather than trained human examiners. Despite its increasing popularity, automatic evaluation of spoken language is still shrouded in mystery and technical jargon, often resembling an opaque "black box" that transforms candidate speech to scores in a matter of minutes. Our chapter explicitly problematizes this lack of transparency around test score interpretation and use and asks the following questions: What do automatically derived scores actually mean? What are the speaking constructs underlying them? What are some common problems encountered in automated assessment of speaking? And how can test users evaluate the suitability of automated speaking assessment for their proposed test uses? In addressing these questions, the purpose of our chapter is to explore the benefits, problems, and caveats associated with automated speaking assessment touching on key theoretical discussions on construct representation and score interpretation as well as practical issues such as the infrastructure necessary for capturing high-quality audio and the difficulties associated with acquiring training data. We hope to promote assessment literacy by providing the necessary guidance for users to critically engage with automated speaking assessment, pose the right questions to test developers, and ultimately make informed decisions regarding the fitness for purpose of automated assessment solutions for their specific learning and assessment contexts.

N. Khabbazbashi (✉)
University of Bedfordshire, Luton, UK
e-mail: nahal.khabbazbashi@beds.ac.uk

J. Xu · E. D. Galaczi
Cambridge Assessment English, Cambridge, UK

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_25

25.1 Introduction: Purpose and Testing Context

The assessment of L2 speaking has traditionally involved face-to-face proficiency interviews which are both delivered and scored by trained examiners (e.g., the IELTS Speaking test) or semi-direct speaking tests which are tape or computer-mediated but scored by humans (e.g., the *TOEFL iBT*® Speaking test). Rapid advances in speech processing and machine learning technologies, however, are transforming how speaking is tested. In speaking tests that deploy automated speech evaluation (ASE) technologies, e.g., the Versant English Test and *TOEFL Go!*®, human examiners are removed from the assessment context, as delivery of prompts and scoring of responses are done by computer algorithms (Bernstein et al. 2010). Such tests may not fully capture the complexities of co-constructed interaction, but nevertheless provide several advantages over traditional speaking assessment, such as practicality in terms of delivery and scoring, quick turnaround of results, and reduced influence of human-related bias and thus increased scoring consistency. Further, ASE has the potential for generating automated individualized feedback for learning, which is a defining feature of the next generation of language assessment tools (Xu 2015). As such, automated speaking tests serve as an attractive alternative to more traditional speaking tests for test developers and users alike.

Given the prevalence of ASE technologies in speaking assessment, the purpose of this chapter is to critically engage with ASE and discuss key issues regarding its use.

25.2 Testing Problem Encountered

Despite its increasing popularity and continuous evolution, ASE is still shrouded in mystery and technical jargon, often resembling an opaque "black box" that transforms candidate speech to scores, without disclosing much about its internal workings or the issues surrounding the scores it generates. Such limited transparency presents a concern in the language testing community where transparency is seen as integral to professional standards. It also hinders informed decision-making in various contexts, as test users may have limited understanding of ASE and the meaning of scores generated by such systems.

While not offering a solution to this problem in this chapter—given the array of technological offerings from various test providers and the transient and continuously evolving nature of such technologies—we believe that a constructive step forward in addressing aspects of the problem lies in enhanced understanding and closer engagement with key issues vis-à-vis ASE.

25.3 Solution/Resolution of the Problem

In this section, we will attempt to address the limited transparency around test score interpretation and use in ASE by focusing on how ASE scores are derived and by discussing the main issues and challenges that ASE encounters.

Automated Speech Evaluation: How Does It Work?

The core technology used in ASE is an *auto-marker* that scores the audio of human speech nearly instantaneously. Once a candidate finishes recording speech on a user interface, the audio files are sent to the auto-marker via an Application Programming Interface (API), i.e., a set of data transfer protocols between the two programs. On finishing scoring, the auto-marker pushes the results back to the user interface via the API. A measure commonly used to indicate the auto-marker time efficiency is the "real time factor" or RTF. An RTF value of one is comparable to the speed of human scoring. It indicates that the computer processing time is equal to the length of an audio file, or, in other words, an automated score is generated as soon as the speech ends. The RTF of the Cambridge English Speak & Improve auto-marker, for example, is 0.84, suggesting that the auto-marker is slightly faster than a human marker (Cambridge Assessment English 2016).

A speech auto-marker consists of three major components (Wang et al. 2018): a speech recognizer, a feature extraction module, and a scoring model/grader (see Fig. 25.1 for the architecture of a speech auto-marker).

The *speech recognizer* conducts automatic speech-to-text transcription, identifying words and phrases in the spoken language, and converting them into text. The two main components of the speech recognizer are the Acoustic Model that maps sounds to phonemes/words, and the Language Model that estimates the probability of a hypothesized word sequence based on training corpora (Yu and Deng 2016). Lieberman et al. (2005, p. 1) illustrate the functioning of the Acoustic Model with two possible outputs as the best recognition results for a string of speech: "wreck a nice beach you sing calm incense" or "recognize speech using common sense." Based on prior knowledge gained from the corpora, the recognizer will select the latter output as the most probable sequence. The Acoustic Model must be trained on a set of accurately transcribed spoken data. The training process involves pairing the audio speech with the transcriptions of that speech, so that the model learns the association between sounds and their orthographic representations (Yu and Deng 2016). The performance of a speech recognizer is usually measured by word error rate (WER) or the rate of misrecognition in the machine transcription compared to a gold-standard human transcription.

The *feature extraction module* contains a set of programs that can automatically extract construct-relevant features from both the audio signal and the transcription generated by the speech recognizer. The features extracted directly from the audio signal—known as acoustic features—are used as proxies for measuring fluency and pronunciation and have nothing to do with the *content* of speech. Typical fluency proxies include features related to rate of speech and pauses/hesitations, such as average duration of speech chunks, articulation rate, and pauses per utterance; typical

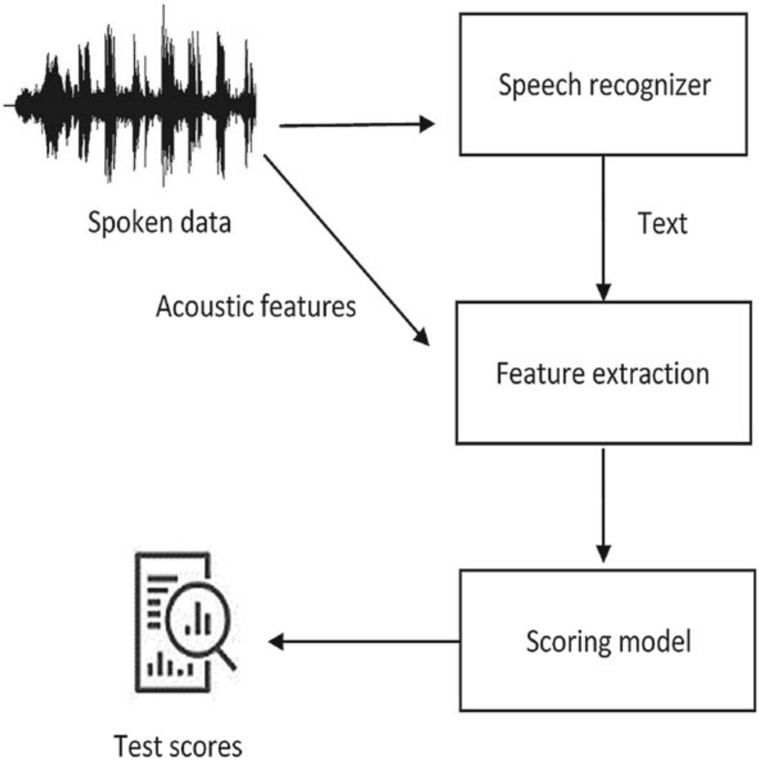


Fig. 25.1 The architecture of a speech auto-marker

pronunciation proxies include confidence in mapping audio to text, percentage of phonemes not superfluously added, and pitch range. The features extracted from the transcribed speech—known as language features—tap into constructs such as grammar, vocabulary use, coherence, and content relevance. Grammatical analysis is conducted by a parser which identifies the part of speech of each word and the syntactic structures of the speech product. Based on this analysis, grammatical accuracy, and complexity are estimated. Vocabulary features mainly focus on lexical diversity and complexity. For example, lexical diversity is measured by normalized frequency of unique unigrams (single words), bigrams (two-word sequences), and trigrams (three-word sequences). Lexical complexity is measured by the distribution of words at various frequency levels according to reference corpora. Coherence and content relevance features apply Latent Semantic Analysis, which estimates the semantic relationship or distance between words, sentences, and passages. Coherence thus is measured by the semantic relationships among words/phrases/sentences in the speech, and content relevance is measured by the semantic distance between the test prompt and response (Van Moere and Downey 2016).

Finally, the *scoring model* (also known as grader) applies the automated features mentioned above to generate test scores. Different scoring models have been used to predict human scores, such as multiple regression (Xi et al. 2012), classification tree (Xi et al. 2012), and non-linear models (Van Moere and Downey 2016). The development of these models usually requires large amounts of training data—thousands of L2 spoken responses with associated gold-standard human scores.

25.4 Issues and Challenges

Despite swift advancements in ASE technologies, challenges remain in the reliable and valid assessment of L2 spoken performance. Here we focus on several key ASE issues.

25.4.1 Performance of the Speech Recognizer

As mentioned earlier, speech recognition is the first and arguably the most important step for ASE. Speech recognition errors may reduce the accuracy of the linguistic analysis module in an ASE system, leading to inaccurate automated scores and feedback (Knill et al. 2018). The training of a speech recognizer relies on a large amount of human-annotated spoken data (Wang et al. 2018). There is, however, limited availability of large L2 spoken corpora that cover different L1s and proficiency levels for training purposes. A speech recognizer's transcription accuracy tends to vary across accents and oral proficiency levels, and usually performs better on proficient speakers or accents that are well represented in the training data. That variability—the so-called training data effect—can result in high WER. Chen et al. (2018), for example, report a WER range of 28.5%–38.5% for the latest version of *SpeechRater^{SM}*—an automated speech system for the scoring of nonnative spontaneous speech. Slightly better figures are reported by Wang et al. (2018) for their state-of-the-art speech recognizer, with WER between 20% and 30% on free L2 English speech and between 10% and 20% on read-aloud tasks. In addition to training data, factors such as audio quality (e.g., background noise, a quiet speaking voice, or breathing on the microphone) may further reduce the accuracy of the speech recognizer (Yu and Deng 2016).

25.4.2 Task Types and Scoring Features

The most commonly used task types in ASE systems are constrained task types such as *read aloud* and *sentence repetition.* Their use is closely related to the functioning of the speech recognizer, since predictable speech can be recognized and scored with greater accuracy compared to spontaneous unpredictable speech (Xi et al. 2012).

In such task types, the ASE system knows what to expect in the learner output and therefore the speech recognizer can be trained to maintain high accuracy even with heavily accented speech. The use of these task types—also known as elicited imitations tasks—however, has been subject to much debate in terms of construct representation, authenticity, and face validity (see Galaczi 2010 for a more thorough discussion). Here we focus on the issue of construct representation.

Drawing on psycholinguistic theories of language processing that posit an important role for automaticity and speed of processing in second language acquisition (SLA), Van Moere (2012) advocates the use of elicited imitation tasks as an indication of whether core linguistic skills, which are the "building block of any conversation," have been mastered (Bernstein et al. 2010, p. 359). While these task types may be useful in determining "a minimum basic level of competence in core proficiency" (Van Moere 2012, p. 340), they have nevertheless been criticized for their narrow construct representation. Consider a speaking test in which a learner reads aloud 10 decontextualized sentences and listens to and repeats another 10 sentences. While the speed and accuracy of performing these tasks can give an indication of the extent to which some underlying cognitive processes are automatized, the tasks, in themselves, do not require the learner to think of ideas nor to draw on L2 lexical resources and syntactic knowledge to translate those ideas into speech for any extended length of time (Field 2011). In other words, the cognitive processing demands on the learner are limited. Moreover, the scoring features underlying such tasks are predominantly related to sub-features of fluency and pronunciation. As such, the speaking construct underlying such tasks is narrow and limited.

More recently, advancements in automated speech recognition and deep learning technologies have paved the way for broadening of the speaking construct through the use of free-speech tasks. ASE systems can now go beyond constrained tasks to elicit spontaneous speech which can be assessed for additional features related to L2 speaking ability, such as vocabulary use, grammatical complexity, and topical coherence, as explained in Section 25.3. Scoring systems have therefore improved their capacity to "use rich information, as human raters do" (Chen et al. 2018, p. 24). Challenges nevertheless persist.

While an auto-marker can be trained to detect hundreds of features in speech, it is important for scoring algorithms to incorporate those that are shown to be valid, fine-grained measures of the construct of interest—based on empirical research—and not solely those easiest to extract automatically. For example, measures of speech rate and pausing are widely used in ASE, and yet several studies in SLA have emphasized the importance of the *location* of pauses rather than their *frequency* in influencing perceptions of fluency (de Jong 2018). Isaacs (2018) also challenges the over-reliance of automated systems on pronunciation accuracy and the extent to which L2 speakers' utterances match native-speaker norms and instead argues for the focus on pronunciation errors to be based on their role in comprehension.

Another challenge is the limited capacity of current technologies to capture high-level features of speech, e.g., content appropriateness, topic development, and discourse organization. In addition to such features, which cannot be *easily measured*, there are aspects of speaking which are currently *not elicited* at all in automated tests, e.g., interactional speech features.

In face-to-face speaking tests, tasks are *dialogic* and involve interaction between two or more interlocutors. Automated speaking tests, on the other hand, predominantly include *monologic* tasks, which do not require a candidate to co-construct interaction with another interlocutor, respond to the interaction as it evolves, negotiate meaning, take turns, and adapt their speech to the context. This results in a narrowed language construct underlying these tests that also has SLA implications: both face-to-face and automated test formats require a candidate to produce *output*; but only one can elicit *interaction*, which has been shown to be essential for the development of language competence (Gass and Mackey 2014).

In face-to-face speaking tests, examiners are trained to manage interaction and adapt questions based on candidates' performance. In contrast, automated speaking tests are generally linear tests and not adaptive. That is, candidates are tested with a preassembled set of questions selected from a wider item pool rather than questions geared toward their proficiency levels.

At this point in time, automated systems have not been trained successfully to simulate interaction and co-construct conversations, for example by giving backchannel feedback (e.g., "yeah"), confirming comprehension (e.g., "Exactly!"), or asking follow-up questions, which are features shown to be part of the construct of interaction (Galaczi and Taylor 2018). These limitations have led to the ongoing debates on construct representation and the validity of automated speaking assessment for various testing purposes (e.g., Galaczi 2010; Xi 2010; Xu 2015). While research in spoken dialogue systems has great potential in addressing these gaps in the future (see Litman et al. 2018), challenges remain.

25.4.3 Test Impact

Test impact refers to the effects or consequences of tests on teaching and learning as well as educational systems and the broader society. Some concerns have been raised about the potential negative impact of ASE, which we now turn to.

As mentioned in Section 25.4.2, ASE systems rely heavily on constrained and monologic task types. A possible negative impact is therefore an excessive preoccupation with monologic speech in classrooms at the expense of interactive tasks and co-constructed dialogues. The use of such task types is also seen as a threat to the perceived authenticity of automated tests. In a study focusing on candidate attitudes toward the automated Versant English Test (Fan 2014), respondents showed a stronger preference for more open-ended tasks such as story retelling compared to read aloud and sentence repetition tasks. Qualitative respondent feedback suggested that the latter were seen as lacking authenticity, with one participant commenting that "in real life we are never required to use language that way" (Fan 2014, p. 14).

Another critical issue is the increased likelihood of candidates displaying abnormal test behaviors in an attempt to cheat automated systems. That is, if the scoring algorithms “fail to assign credit to qualities of a response that are relevant to the construct that the test is intended to measure” (Chapelle and Douglas 2006, p. 41), candidates may choose to ignore such qualities in their language production. In a survey conducted by Xi et al. (2016) on TOEFL Practice Online users in China, results showed that 20.6% of respondents consciously changed their speaking behaviors when knowing this low-stakes speaking practice test was scored by a computer; specifically, they tried to pronounce words very carefully, kept on speaking even when they made little sense, spoke as quickly as they could, and paid less attention to logic and content, two aspects they felt the automated system was not good at scoring. Further, when asked about a hypothetical scenario in which only ASE is used to score a high-stakes speaking test, 57.3% indicated they would be likely to apply strategies to fool the computer. It is unfortunate that ASE developers seldom choose to publish research on malpractice—likely due to an effort to conceal the weaknesses of scoring algorithms—despite it being a critical piece of validity evidence for the trustworthiness of automated scores (Xi 2010; Xu 2015). Bernstein et al. (2010), for example, acknowledged the lack of research on ASE’s robustness against “off-construct coaching” (p. 374) and warned against using “automated scores alone” (p. 372) for high-stakes decision-making.

25.5 Insights Gained

As discussed so far, ASE tests offer many possibilities and yet are not without limitations. These limitations often mean that, in comparison to more communicatively oriented face-to-face speaking tests, the construct underlying ASE does not necessarily capture the many complexities of spoken performance and interaction in the real world and the scoring features of ASE usually do not fully cover the range of evaluation criteria used by human raters. Other issues to consider are the potential negative impact of ASE systems on language learning in classrooms and the threat of cheating. An important insight from these discussions is that despite its many advantages, automated assessment should not be used as the sole basis for high-stakes decision-making, as this is a complex issue which should be informed by a range of considerations. Another key insight is the need for increased assessment literacy so that test users can become better informed of the various debates surrounding ASE. To facilitate this, Table 25.1 is a compilation of a list of key questions for ASE users to consider.

25.6 Conclusion: Implications for Test Users

The need for transparency about the facets contributing to a test’s validity is a fundamental principle in language assessment, and in this chapter, we have aimed to contribute to the transparency of ASE tests. An important implication from our

Table 25.1 Key questions for ASE users

Key question	Why is this important?
What data has the ASE system been trained on?	The breadth of speech recognizer training data, especially in terms of L1s and proficiency levels, determines how accurately it performs with learners of different backgrounds/language levels
How is the ASE test administered in practice?	The accuracy of ASE scores relies on appropriate test administration such as minimal background noise, correct microphone setup, high internet speed, and clear test instructions
What tasks are used in the test?	An ASE test should include a range of tasks types which go beyond highly constrained tasks (e.g., reading aloud, sentence repetition) to unrestricted tasks (e.g., free monologue speech), and potentially tasks which simulate dialogue
What scoring features are extracted to inform a score?	A range of speech features that contribute to successful communication should be captured by an ASE system. They should extend beyond pronunciation and fluency features to grammatical and lexical features and ideally those related to organization of speech, relevance of content, and topic development
What is the potential for cheating on the test?	The range of construct-relevant features included in ASE's scoring model determines its robustness against cheating. If content relevance and topic development, for example, are covered by the scoring model, then the potential for cheating is likely to be greatly reduced
What is the impact of the test on language learning?	A test should have a positive impact on learning. The broader the range of tasks and extracted features and the more relevant they are to the target language use domain, the higher the potential for positive impact
Is there a good fit between the purpose and stakes of the test and the ASE system used?	No test is valid in itself. It is valid *for* a specific purpose. The questions here therefore need to be considered in their own right and also in the context of the intended test purpose. For example, is there a good match between a low-stakes practice test and the ASE system used? Or between a high-stakes university entry test and the construct underlying ASE?

discussions is that the *same* test result—often reported as a Common European Framework of Reference for Languages (CEFR) level—on an automated speaking test and a face-to-face speaking test can have *very different meanings* in terms of what is actually assessed on the test. The ultimate implication for test users is that a deeper understanding of ASE and its challenges can help them become informed users who can critically engage with such systems, pose the right questions to test developers, and better understand the meaning of ASE scores. Such critical awareness will support them in selecting tests that are right for their needs and contexts and judging if ASE tests are fit for purpose.

References

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377.

Cambridge Assessment English. (2016). *Cambridge English automatic scoring of English-speaking tests* (Internal Cambridge Assessment English research report).

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology.* Cambridge: Cambridge University Press.

Chen, L., Zechner, K., Yoon, S., Evanini, K., Wang, X., Loukina, A., et al. (2018). Automated scoring of nonnative speech using the *SpeechRater*^{SM} v. 5.0 engine. *ETS Research Report Series, 1,* 1–31.

de Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*(3), 237–254.

Fan, J. (2014). Chinese test takers' attitudes towards the Versant English test: A mixed-methods approach. *Language Testing in Asia, 4*(6), 1–17.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 65–111). Studies in Language Testing volume 30. Cambridge: UCLES/Cambridge University Press.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Computer-based assessment of foreign language speaking skills* (pp. 29–51). Luxemburg: European Union.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236.

Gass, S., & Mackey, A. (2014). Input, interaction, and output in second language learning. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 108–206). New York, NY: Routledge.

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly, 15*(3), 273–293.

Knill, K. M., Gales, M., Kyriakopoulos, K., Malinin, A., Ragni, A., Wang, Y., & Caines, A. (2018). *Impact of ASR performance on free speaking language assessment*. Paper presented at Interspeech, Hyderabad, India.

Lieberman, H., Faaborg, A., Daher, W., & Espinosa, J. (2005). *How to wreck a nice beach you sing calm incense.* Paper presented at the 10th International Conference on Intelligent User Interfaces, San Diego, CA.

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly, 15*(3), 294–309.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325–344.

Van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 342–357). Berlin: De Gruyer Mouton.

Wang, Y., Wong, J., Gales, M., Knill, K. M., & Ragni, A. (2018). *Sequence teacher-student training of acoustic models for automatic free speaking language assessment*. Paper presented at the 2018 IEEE Spoken Language Technology Workshop.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291–300.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing, 29*(3), 371–394.

Xi, X., Schmidgall, J., & Wang, Y. (2016). Chinese users' perceptions of the use of automated scoring for a speaking practice test. In G. Yu & Y. Jin (Eds.), *Assessing Chinese learners of English* (pp. 150–175). Basingstoke: Palgrave Macmillan.

Xu, J. (2015). *Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech.* Unpublished doctoral dissertation, Iowa State University, Ames, IA.

Yu, D., & Deng, L. (2016). *Automatic speech recognition.* New York, NY: Springer.

Chapter 26
Developing a Meaningful Measure of L2 Reading Comprehension for Graduate Programs at a USA Research University: The Role of Primary Stakeholders' Understanding of the Construct

Ahmet Dursun, Nicholas Swinehart, James McCormick, and Catherine Baumann

Abstract University of Chicago graduate students are required to demonstrate their ability to read in a foreign language in order to conduct research and participate in an international community of scholars. Previously, like many institutions, the University asked students to show evidence of this ability through a translation exam. A number of concerns arose from various stakeholders regarding the use of this exam, primarily that a translation exam failed to measure the skills required in the research domain. To address these concerns, the University of Chicago Language Center initiated the process of developing an alternate assessment measure—the Academic Reading Comprehension Assessment (ARCA$^{\mathrm{TM}}$). This chapter details the steps taken to enact this change, beginning with transferring administration of the exam to language assessment specialists and then conducting meetings with each department to discuss the construct of reading for research purposes, introduce the format of the exam, and convince faculty and deans of its validity. Next, we discuss the results from follow-up focus-group interviews with these key stakeholders to explore their understanding of the theoretical and pedagogical rationale underpinning the construct of academic reading comprehension and task types in the ARCA and its effect on a re-evaluation of their language requirements. We then present a model that can be replicated to gain buy-in among key stakeholders in similar contexts, and we summarize our findings into insights that can be utilized to bring similar innovations or improvements to language assessment practices by increasing assessment literacy.

A. Dursun (✉) · N. Swinehart · J. McCormick · C. Baumann
University of Chicago, Chicago, IL, USA
e-mail: adursun@uchicago.edu

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_26

26.1 Introduction: Purpose and Testing Context

The University of Chicago (UChicago) is a private, R1 university with approximately 6,000 undergraduate and 12,000 graduate and professional school students. The breadth of languages offered at UChicago has long been a hallmark of the institution—more than 55 different languages in each academic year, from modern and classical languages to rarely taught languages of the Near East or South Asia. The University of Chicago Language Center (CLC) supports all language instructors and learners on campus. Strong language abilities are considered integral to many academic disciplines on campus, principally in Humanities and Social Sciences fields. Graduate students specializing in the study of a particular area are expected to develop expertise in the language of their primary research objects. But programs also expect that students will be able to access scholarly writing about their area of study (i.e., secondary literature) which may be written in a number of languages. Twenty graduate programs on campus—in the Humanities, Social Sciences, and Divinity School—have some form of foreign language reading requirement in addition to primary course requirements.

In order to be conversant in the full range of scholarship necessary for comprehensive research expertise, a scholar in training cannot be limited to English-language publications but must also incorporate research being produced and published in other languages, or scholarship from an earlier era that has not been translated into English. Particular research language needs differ from field to field, but students at UChicago often need to be able to read scholarly writing in French, German, and Spanish and less frequently in other languages such as Chinese, Hebrew, and Japanese. The ability to address scholarly developments beyond the English-speaking world allows a scholar in training to situate her own contributions within a network of researchers that stretches across the world and thus deepens her engagement with her field of study.

This study outlines a detailed and rigorous process for training stakeholders on the UChicago campus in language assessment literacy—specifically the concept of test construct—in order to introduce an assessment tool that is tailored to measure the target abilities which justify the foreign language reading requirements described above. This literacy training was essential in our efforts to introduce a new assessment format, the Academic Reading Comprehension Assessment (ARCA). In contrast to previous assessments, the ARCA focuses on reading comprehension rather than translation. The goal of this study is to provide an overview of the approach and steps taken in this training so that other groups can adapt the process and introduce new assessments at their own home institutions.

26.2 Testing Problem Encountered

Prior to the 2014–2015 academic year, the University of Chicago Registrar's Office administered a translation exam in a variety of languages, and graduate departments used the results of this exam to determine fulfillment of their foreign language reading requirements. Students translated two passages drawn from secondary literature in two different fields in a given target language. The exam was administered once per quarter for any student who sought to fulfill a departmental requirement. All students received the same two texts, without regard for the texts' relevance to the students' areas of study.

Exams were typically created and graded by a graduate student or lecturer with expertise in the particular target language, and the test creators received minimal guidance on text selection and scoring method. Test taker translations were graded word-by-word, with each word weighted equally and emphasis given to literal accuracy. No rubric was created for any exam. The translation test format called for students to apply grammatical knowledge and vocabulary knowledge—including effective dictionary usage—in order to produce an English version of the original. A high degree of attunement to the nuances of English usage in technical contexts was thus an implicitly required skill.

The translation format was a poor measure of the desired abilities—reading secondary scholarship in the target language for use in research—in a number of respects: (1) mismatch to target domain; (2) unfair advantage to students from specific disciplines; (3) substantial disadvantage to non-native English speakers; (4) simultaneous under- and over-estimation of test takers' reading abilities. Each of these respects is discussed below.

First and foremost, the translation task itself is not representative of the target domain of academic reading. A scholar with strong reading skills in a target language does not produce a word-for-word translation of a book or journal article in order to read and incorporate it into her research. Rather, she simply reads and processes the text, perhaps taking notes or paraphrasing for use in her own writing. Second, any student who happens to receive a text from his or her discipline on the test would have a distinct advantage over a test taker unfamiliar with the general topic of the text, and the latter would, conversely, be at a disadvantage. This discrepancy could have been addressed through tailoring exam texts to individual fields, as has been done with the ARCA, but this was not standard practice with the translation format of the exam. Third, non-native English speakers face a significant barrier to success that does not affect native English speakers, namely, the ability to choose the particular words or idioms in English which capture the intended meaning when producing a translation. It is thus unclear in such cases whether results of the exam reflect the L2 reading skills or English translation/writing abilities and inabilities.

Finally, the translation task frequently misrepresents the skills of the test taker, as grading a translation can be a subjective practice and therefore highly susceptible to raters' varied standards of a quality translation. Translation is, in different respects, more and less difficult than reading. A fine translation, accurately capturing the meaning of the original text and expressing it in precise English, requires very advanced language abilities in both languages: perfect L2 comprehension and creative composition abilities in English. Depending on a grader's insistence on the importance of nuance or of a test taker's imperfect skill in English, some test results might fail to achieve a passing grade even if the underlying comprehension might have been adequate.

For the purposes of assessing a graduate student who is not yet expected to be an expert, however, the standard for judging a translation cannot be so high; some level of inaccuracy and imperfect expression must be tolerated. This leaves literal fidelity to the original text as the dominant element being assessed. Test takers, who often recognized that they were unlikely to be able to translate perfectly, thus fell back on literal word-for-word translating, as it conferred some element of confidence that the meaning was accurate. This practice, in turn, made the test into a complicated dictionary exercise. A test taker could come "close enough" to the original meaning to gain a substantial amount of credit without necessarily having a high level of comprehension of the function of individual words and clauses within the context of the whole. In this way, the test often over-estimated the reading abilities of some test takers.

This last factor led, in the case of UChicago's translation test, to situations in which a student with a strong grammatical foundation and vocabulary could pass the test but still have only limited reading abilities. Thus, there was a general perception among students and faculty on campus that the test was a relatively meaningless hurdle that had to be cleared simply because it was "the university reading exam" rather than because anyone believed it provided valuable information about reading ability. Reports were legion of students frustrated by spending time preparing for something they viewed as useless or stalled in their progress because they were unable to pass the test.

The new comprehension format of the ARCA was developed to address the shortcomings of the translation format. It achieves this by: (1) using the actual target domain as the model for the format: reading, synthesizing, and reproducing, rather than translation; (2) offering field-specific texts to test takers from different programs with different background knowledge; (3) minimizing the effects of construct-irrelevant factors (e.g., non-standard or unidiomatic English usage by non-native speakers when grading); (4) standardizing grading and fitting it to the desired skill set through rigorous oversight of the rubric developed for each test. These improvements led to an exam format that closely matches the ways students actually use L2 reading comprehension in a research context and emphasizes the skills and strategies of reading comprehension and activation of background knowledge over the more mechanical skills of literal translation.

26.3 Solution/Resolution of the Problem

26.3.1 ARCA Design and Implementation

After a rigorous domain analysis of the task of conducting academic research in a secondary research language, the construct for the ARCA was defined as comprising the following abilities: (1) to read scholarly texts, (2) to comprehend arguments and evidence, and (3) to reproduce those arguments in one's own words in the primary research language. Thus, it intends to specifically measure students' reading comprehension, not their interpretation, analysis, or evaluation of the text. The ARCA construct also employs the use of discipline-specific texts as input: each student receives a text drawn from his/her area of study.

The ARCA consists of three components. In part 1, students receive a discipline-specific academic text. They read, annotate, and take notes with the help of a print dictionary. The text is taken away, and students write a summary protocol, reproducing the central arguments in their own words in the primary research language (i.e., English). Removal of the text while writing the protocol ensures that they do not attempt to translate it. Instead they focus on articulating the details of the author's arguments in their own words. In part 3, students have access to the text again and must respond to short-answer questions before translating a short excerpt.

The summary protocol measures students' ability to synthesize the whole text, isolating what they determine are the central arguments and leaving out what is less important. Short-answer questions measure students' ability to connect isolated concepts or phrases to the larger argumentative structure of the text, often requiring close examination of key concepts in the text. The translation task is designed to evaluate the ability to render finely detailed information accurately. This paragraph would be the type of passage that students might cite in a footnote or quote in their own paper, thesis, or dissertation. The focus of this task, however, remains overall comprehension and not on the ability to reproduce the L2 sentence structure literally.

Each of these tasks is graded on the basis of a corresponding analytic rubric developed for each unique text. The point distributions in each task are weighted according to the value they represent in the response, and each rubric and its content must be agreed upon by the test developer and an independent anonymous reviewer. Moreover, ARCA rubrics are proficiency-oriented, keeping raters focused on the material that test takers comprehended rather than penalizing them for what they could not.

26.3.2 Training Stakeholders in Test Construct

After the ARCA was designed, developed, and piloted, the CLC sought to gain support from key stakeholders on campus before introducing the new assessment. Stakeholders included graduate faculty, students, deans of students, and department administrators. It was necessary for each of these stakeholders to understand the construct of the ARCA and what it measures so the use of this new assessment would be accepted as an improvement over the previous translation format. Indeed, accurate understanding of the construct and interpretation and use of the ARCA results required the CLC to educate stakeholders with relevant language assessment literacy, knowledge clearly absent prior to this initiative.

Pill and Harding (2013) highlight the danger of this lack of assessment literacy: those who make decisions on the basis of test scores without the necessary assessment literacy may inaccurately interpret the scores and thus use the test beyond its intended purposes. They add, however, that "there has been little research to date on the level of 'language assessment literacy' displayed by non-practitioners in their conceptualizations of language assessment, making it difficult to establish how best to raise awareness of assessment practice and processes within these stakeholder groups" (Pill and Harding 2013, p. 382).[1] The *Standards for Educational and Psychological Testing* recommend: "when test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, and how scores are intended to be used" (American Educational Research Association 2014, p. 119). Accordingly, we now outline a methodology and process that built assessment literacy among key stakeholders and involved them in the introduction of this innovative reading comprehension assessment.

26.3.3 Methodology and Process

The CLC has autonomy in test design and administration but plays only an advisory role for the University's various graduate departments, with no authority to make direct changes to requirements or policies. Simply announcing the change from one assessment model to another would likely have been met with resistance and confusion. Therefore, it was important to use a bottom-up approach to gain support at each level, from students to faculty to department chairs. The CLC thus organized a series of meetings to increase assessment literacy among each group of stakeholders.

Throughout the process, we sought to recognize and maintain each department's sense of ownership over their requirements and how those requirements were met, then to delineate the Language Center's role as advisers in the process.

In individual departmental meetings, faculty and department chairs were invited to articulate the reading skills they wanted graduate students to possess. When stakeholders said they wanted their students to be able to "read secondary literature," we probed the responses in order to initiate discussions about how students use texts in research and what skills and abilities were needed in that domain, in contrast to the

[1] Taylor (2013) also underlines the dearth of research-based practices and outcomes in helping to "inspire and shape new and innovative initiatives for disseminating core knowledge and expertise in language assessment to a growing range of test stakeholders" (p. 405).

translation of texts or of primary literature. We attempted to highlight these misalignments in ways that were sensitive to the delicate power balance among faculty and academic staff. We simultaneously attempted to create awareness of the construct definition for the ARCA exam—the ability to read scholarly texts in order to reproduce the central arguments and evidence in one's own words in the primary research language—and provided information about the domain of reading for academic or research purposes.

We used our expertise to explain why the comprehension model was not just more aligned but more pedagogically sound, and that changing the format of the test would change the way students prepared for it, bringing their practice activities and study methods more in line with the desired real-world functions. We presented the new comprehension format to faculty and department chairs, demonstrating how it measures the key knowledge areas, abilities, and skills of the construct, as opposed to the separate skills and abilities of translation. We also proposed realigning graduate-level reading-and-research courses to prepare students for the comprehension exam, streamlining completion of the requirement for students and relieving departments of this responsibility.

Throughout the process we addressed misunderstandings, particularly in terms of task types. Some faculty assumed that a "reading comprehension test" meant multiple choice items. We strove to illustrate the difference between multiple choice (selected-response) and summary protocol (constructed-response) tasks and the extent to which each aligns (or does not align) with the intended domain and construct. Leaning on the work of Bernhardt (2011), we provided a theoretical rationale for the summary protocol and short-response questions used in the ARCA, and described how these task types provide data which reflect the nature of the reading process in terms of encoding, restructuring, and analyzing information and are a more valid measure of reading comprehension.

We also presented detailed information about the format, scoring, and procedures of the ARCA exam. We provided faculty with samples of actual exams, again highlighting the assessment tasks used in the ARCA in contrast to multiple choice and other discrete measures. The scoring of the ARCA also represented a shift from the translation exam, which lacked a consistent scoring and rating model, resulting in questions about test reliability, including inconsistent or incorrect interpretation of students' performance. We outlined all this to show the stakeholders how the ARCA scoring and rubric models could increase the accuracy, consistency, and therefore reliability of this high-stakes test.

After these initial meetings, we maintained contact with stakeholders involved in the process. Each department was approached multiple times over multiple academic years to follow up on internal discussions with regard to the use of the ARCA exam. Ideally, the ARCA requires coordination from multiple levels within a department: chairs were intentionally kept involved in the process, coordinating text selection for the test by requesting from faculty texts and journal articles that their graduate students could be expected to read. This required regular contact with multiple groups of stakeholders within each department, in what can best be described as a "multi-front campaign." We continued to maintain contact after the development and implementation of the ARCA was underway by holding further meetings and eliciting feedback, questions, or concerns regarding the ARCA, its implementation, and its washback effects.

26.4 Insights Gained

Following these steps and the implementation of the ARCA, we sought to evaluate the overall effectiveness of the transition by interviewing stakeholders to investigate if the process of building assessment literacy and understanding of the ARCA construct was successful and led to the desired changes. We conducted semi-structured individual interviews, some face-to-face and some over the phone. We invited six participants, ultimately conducting three interviews from three separate departments. Two participants were faculty members and one was a dean of students. These interviews were transcribed and analyzed using thematic coding by two independent coders based on the following themes: factors that affected stakeholder buy-in, opinions on the clarity of the training, stakeholders' recollections of departmental discussions, and remaining issues.

26.4.1 Factors that Affected Stakeholder Buy-in

A key goal in implementing this transition was ensuring that stakeholders viewed the change as legitimate and beneficial. Success was apparent in the positive views stakeholders held of the new exam and its effects. In follow-up interviews, stakeholders viewed the assessment of reading comprehension as more relevant to the target domain and more beneficial to students than translation, leading to a more comprehensive retention of language skills. They also had received more positive feedback from students, commenting that students were able to see the value of this exam and that anxiety about the test had gone down considerably. The transition led to improved washback in the form of less focus on teaching students how to translate and more focus on teaching them how to read in the second language. Stakeholders also commented on the improved format and delivery of the ARCA exam, particularly how it presented students with texts related to their fields. Finally, they described the benefits the ARCA brought about within their respective departments by making it easier to explain the reading requirement to students in the context of a long-term, retainable skill that would benefit their future careers.

26.4.2 Opinions on the Clarity of the Training

Stakeholders held positive views of the training (e.g., conversations in meetings with CLC experts) they received about the new exam, its procedures and scoring, and its intended use. They viewed the explanations from language assessment specialists as clear and logical, and said the ability to use articles from specific fields was important to them. They also stated that the explanation of the underlying pedagogical principles was helpful in explaining the new format to students and "pitching" the change to faculty.

26.4.3 Stakeholders' Recollections of Departmental Discussions

Since the use of a reading comprehension model as fulfillment of the reading requirement was a voluntary decision to be made by each department, stakeholders were asked to reflect on internal departmental discussions. The adoption of the ARCA exam appears to have met little resistance. Faculty recognized the comprehension exam as more relevant to departmental goals and students' future careers. They also saw the benefit of passing the development and administration of these exams to specialists in the university's Office of Language Assessment, thus removing this responsibility from their own faculty and adding impartiality to the process. It should be noted, however, that not all departments have adopted the ARCA exam; some are still relying on the translation model.

26.4.4 Remaining Issues

Some departments are still trying to decide what prerequisites should be in place for reading-and-research courses and how to fit those prerequisites into the timeline of a graduate degree. Some students enroll in reading-and-research courses and take the subsequent examination with no previous experience in that language, leading to failed attempts, negative perceptions, and delayed progress. Another remaining challenge is a lack of equal buy-in and participation among faculty members within a department. The ARCA protocol requires that texts are selected from a range of faculty and specialists within a department, leading to a variety of texts that are reflective of the discipline. When only a few faculty members select most or all of the texts within a department, however, there is a strong chance of bias in text selection. In addition, some faculty members were not following the predefined specifications for text selection. As a result, we have begun requesting the names of academic journals rather than specific articles, increasing participation from faculty and the scope of texts.

26.5 Conclusion: Implications for Test Users

The bottom-up approach in implementing the ARCA was far more arduous and time-consuming than simply announcing a new exam format to departments, but the conversations and support this approach generated were vital to the ARCA's long-term success. Below we present an outline of steps other groups can take when planning to implement similar changes to language assessment practices at their institutions.

26.5.1 Gaining Buy-in from Stakeholders When Implementing Changes in Testing Practices

- **Acknowledge the role of stakeholders in decision-making.** Stakeholders must feel they are active agents in the change taking place. In high-stakes situations such as a graduation requirement, it is important that stakeholders feel ownership over the requirement and how it is fulfilled.
- **Identify problems with existing assessment practices.** Raising stakeholders' assessment literacy is effective, though care is needed when working across differing levels of an institution's hierarchy. In our context, this meant illustrating the misalignment between the target language use domain and the existing test construct to faculty and department administrators.
- **Present the test construct, format, scoring, and procedures.** Illustrate the face validity of the construct and explain how the scoring allows differentiation in performance.
- **Highlight pedagogical benefits of proposed changes.** Highlight the positive impact of the proposed solution in assessment and its washback in teaching practices.
- **Address misunderstandings.** Be prepared for misunderstandings to arise as a necessary component in the process of raising assessment literacy.
- **Follow up and track progress.** Maintain contact with stakeholders and decision-makers to ensure that they feel involved in the process and that progress is being made.

Research has revealed that diffusion of innovation is a long-term process requiring the active participation of agents who are affected by its implementation. It is equally important to establish the necessary knowledge-base to increase these agents' awareness and understanding of how innovation functions and the gains it offers (Rogers 2003). In this work, we characterized steps for involving key stakeholders and developing relevant assessment literacy in order to adjust and improve on a long-standing testing practice at the University of Chicago. Through a process of cultivating primary stakeholders' understanding of test construct, language testers can build valuable partners for implementing pedagogical innovation.

References

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bernhardt, E. B. (2011). *Understanding advanced second-language reading*. New York, NY: Routledge.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30,* 381–402.

Rogers, E. M. (2003). *Diffusion of innovations*. New York, NY: Free Press.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30,* 403–412.

Chapter 27
Challenging the Role of Rubrics: Perspectives from a Private University in Lebanon

Christine Sabieh

Abstract One problematic aspect of classroom assessment is in how rubric instruction, specifications, and marking scales are planned and used. Using illustrations of student work and rubrics, I consider the use of rubrics and allude to behavioral, constructivist, and social learning theories as I discuss critical thinking, score inflation, and real-life learning. Through action research I investigate how using rubrics differently in the learning process maximized student learning.

27.1 Introduction: Purpose and Testing Context

Rubrics are an integral part of the teaching-learning process today. Mertler (2004) noted that assessing students' performance is "one of the most critical aspects of the job of a classroom teacher" (p. 49). Action research enabled me to engage in a systematic inquiry about the role of rubrics on learning. I believe educators plan to create optimal learning and use rubrics to encourage critical thinking rather than use rubrics to benchmark acquired knowledge. I believe rubrics should be used as holistic all-inclusive learning tools and not as outcomes tools to measure the degree of task completion. Rubrics, descriptive scoring schemata, are defined by criteria, indicators and scales to assist in the analysis of student work.

The purpose of this chapter is to discuss the impact of rubrics on real-world learning and to describe how rubrics should not dictate mechanical responses but promote holistic thinking and long-term learning. To gain insight into this dilemma, first, I will use descriptive activities to show that rubrics, as alternative assessment tools, mirror standardized assessments but may measure short-term learning and limited problem solving due to eliciting mechanical responses. Then, I will share examples from my teaching-learning spaces to support my critical but practical inquiry to systematically reflect and suggest insights to develop the use of rubrics for improved learning, specifically for long-term learning.

C. Sabieh (✉)
Notre Dame University, Beirut, Lebanon
e-mail: csabieh@ndu.edu.Ib

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_27

For many years, working at five accredited English-speaking universities in Lebanon, I have engaged in self-reflection on the role of the rubric in promoting real-world learning and appeal to educators to consider the impact of rubric use in their learning spaces. My action research focused on my students and how rubrics were used to impact their learning process.

27.2 Testing Problem Encountered

Good assessment mirrors good teaching (Coombe et al. 2007). Over the years, many academics have utilized the rubric as an alternative assessment tool to support their more learner-centered teaching approaches. Rubrics are transparent guides for learners to carry out their tasks and reflect on content achievement. In diversifying learning tasks, many educators have used rubrics to offer assessment options. Was the use of rubrics a blind move on the part of the teachers to follow other educators, or was it a studied change to introduce varied assessment opportunities into the teaching/learning space? I have perceived how my students have used rubrics in completing their assigned learning tasks. My observation was that irrespective of course subject matter, student rubric-engagement may have been directly related to completing task expectation that was based on the criteria indicators and descriptors outlined in the rubric provided with the assigned learning task.

Thus, the solution to be investigated in this action research is the use of rubrics with pre-defined criteria, indicators and scales that ascertain what learners do not know to measure the journey toward real-world learning. A less detailed rubric, known as a single-point rubric, would measure whether learners are assimilating content since it is a rubric that identifies general guidelines the learners have to address to complete the task. According to Fluckiger (2010), the single-point rubrics is an "ethical tool" to guide the learners in identifying details needed to meet the general category guidelines and self-assess their own work. The students are provided with main categories, but the pre-defined criteria, indicators and scales are not present to guide them to accomplish the successful completion of a task. The single-point rubric generally describes the expected outcomes but does not guide the students with details; the students are expected to recall the assimilated content to complete the task. Students create the outcome and self-assess the work before submitting the assignment to the educator for final assessment. However, the creation of a single-point rubric is not sufficient as a stand-alone since it gives students much responsibility, which may challenge low-achieving students.

This investigation intends to explore—with the intention to advocate for—the building of self-generated rubrics through brainstorming and exploring as the best option to measure engaged learning. Here, the learners are forced to use assorted cues in their context to construct knowledge. The learners are to reflect on the content that has been assimilated based on meaningful and diverse construction. This action research will investigate the common and limiting view of criterion-based rubrics to show that self-generated rubrics are guides to measure the learning of real-world language use.

To investigate further, I engaged in the following framing questions to guide my action research: Were students acquiring the skills to efficiently emulate long-term learning? Or did rubric-guiding principles reflect what conventional traditional testing did to regulate answer production? Could one problematic aspect of the assessment method be in how the rubric was planned and used by the learner and/or by the educator? Did this use of rubrics measure real-world learning for simulated contexts of the future? I investigated this by questioning the role the rubric played in the teaching/learning setting.

The following research questions were addressed:

Research question 1. How did students engage with the detailed, multiple-level rubrics to measure learning for real-world language use?

a. Did students use rubrics descriptors as the guide to supply their activity answers?
b. Did students' use of rubric indicators and outcome expectations inflate their grades when compared to traditional test scores?

Research question 2. How did students use rubrics holistically to measure learning for real-world language use?

a. Did students use rubrics holistically to measure learning development when they were focusing on rubrics with task-specific category pointers?
b. Did students use rubrics to assist them in monitoring their learning development?
c. Did students use rubrics with single-point descriptors to facilitate tangible holistic real-life learning?
d. Did students' use of self-generated rubric criteria or indicators facilitate identifying their missing knowledge and increase their learning efficacy?
e. Did students' use of rubrics that provide criteria but describe indicators posited at opposite ends of a scale assist in developing real-life learning?

My action research focused on my students and how rubrics were used to impact their learning.

27.3 Review of Literature

To facilitate the understanding of the impact of rubrics on real-world learning, Information Processing, Behavioral, Constructivist, and Social Learning theories as approaches to learning were considered, as was the notion of rubric styles and assembly to demonstrate long-term learning growth through a spiraling learning action research initiative.

Students respond to external stimuli that educators provide. Teachers plan the learning and provide students with the rubric that will be used to assess the completion of the task as a measure that learning has happened. Students follow the rubric and complete the expected assignment. Within the current teaching/learning context, it may no longer suffice to assume that the completion of the task meant that learning happens or learning for real-life usage is accomplished. Critical thinking may remain minimal if task guidance is merely completed by following rubric descriptors. According to the Information Processing Theory (Baker 2017), students transform the external information received from educators, or from following the rubrics, into encoded knowledge that is stored for retrieval use which may occur in the immediate future or sometime later.

So, students use their senses (e.g., sight, touch) to register stimuli and analyze content to understand it. Depending on course content teaching/learning objectives, educators may ask students to store content in four ways: (1) To store material as is; (2) to store information by assimilating and accommodating through meaning; (3) to store given content for immediate retrieval to use in a task; and/or (4) to store it for later use (Baker 2017). When information is stored based on understanding, the retrieval knowledge is to emanate from long-term memory and not direct, time-defined, space-limited, short-term memory that would be used for immediate tasks. Interpreting the theory endorses that cognitive growth is a process of making meaningful connections that include new material from the context, old material from storage, and experience of past content use. Thus, in the optimal teaching/learning context, students freely make connections and increase their understanding of the content being taught.

Moreover, educators may assist their students in accomplishing the learning through scaffolding exercises. Constructivists are keen advocates of scaffolding as the framework to optimize students becoming independent and responsible real-world learners (Wright 2018). The Information Processing theory also provides support for learning through a more passive behavioristic approach.

When senses are registered as stimuli and are encoded as such, the perception requires minimal or no thinking during the encoding process. The retrieval process is based on a mechanical stimulus-response (S-R) arrangement in which content is recovered immediately for use or, when needed, is recovered to be used in the same manner as it was encoded (Skinner 1989). According to the Behavioral approach to learning (Zhou and Brown 2015), the retrieval response is automatically determined by the expected outcome; if it is correct the response is reinforced; if not, the response results in a punishment. The consequences elicit or emit the future responses and involve no cognitive thinking process. The consequences of eliciting or emitting future responses to stimuli are the grounds the behaviorists use to advocate that habitual associations of S-R may lead to change in learning outcomes by shaping the responses based on associations and/or reinforcement consequences (Skinner 1969). As outcome responses are satisfied through desired reinforcements, the degree of the response re-occurring automatically increases. The process does not involve a thought process or rational problem solving; it involves rehearsal to make the response a motorized action—the stimulus dictates the action. The outcome response is contingent upon the stimulus content and the association is controlled and perfunctory (Skinner 1969, 1989). Critical thinking, problem solving, as well as mastering diverse connectivity through meaning, are limited if the behavioral approach to learning is adopted.

The teaching that takes place in today's classrooms needs to account for additional facets of how content information is processed. Taking on a more active role in how information is processed, students incorporate meaning, mnemonics, and context cues to construct new ways of thinking about content and formulate an understanding of different associations; thus, diversifying the learning process (Zhou and Brown 2015). According to Constructivism, learners should be encouraged to become autonomous, engaging, elaborating, collaborating, in their growth as they acquire an understanding of the knowledge and the strategies they are exposed to (Brooks and Brooks 1993). The students construct as learning is facilitated, enabling cognitive maturation. Students develop their abilities to recognize patterns and create associations, categorizing content, and rehearsing actively to transform, connect, and retrieve more efficiently (Baker 2017). The teaching atmosphere endorses multimodal, contextual, and creative thinking. The students develop a sense of ownership as they re-create and construct meaning (Brooks and Brooks 1993). The students assess their own learning in and across contexts for real-life learning.

Whether facilitated or guided, learning may occur individually or collaboratively. Educators may scaffold students individually or in groups, aiding them to understand the role of the rubric as a means-to-an-end and as an end-in-itself tool (Vygotsky 1978). The journey is to help the students acquire holistic long-term learning through the use of the rubric as a means-to-the-end instrument. In learning spaces, educators need to bridge students' zones of proximal development to help students become independent learners (Wright 2018; Vygotsky 1978). Through explanations and exploratory opportunities, the learners create associations, accommodating new meaning, problem solving, and diversifying their content knowledge within the different contexts as they discover the content (Wood et al. 1976). Discovery learning and social learning initiatives help learners increase their efficacy (Bandura 1977). Through exposure to the educators as the students' model, the scaffold experience becomes the means for the students to learn how to self-regulate their thinking and assimilation. They learn to monitor and develop a self-belief system that promotes understanding and growth with the knowledge for real-life language use (Zhou and Brown 2015). When the learners respond to a stimulus, they act with intent and self-reflectiveness. Social Learning theory postulates that the observing of a model is the basis on which the learners build mental representations that enable them, through the scaffold experience, to observe, monitor, judge, and reflect. Reflective critical thinking nourishes contextual use of knowledge in real-life settings. The efficacy is maintained when active engagement is part of the learning challenge since progress is monitored (Bandura 1977, 1986).

Accordingly, educators and students are to use the rubrics to reflect their real-life learning journey. The rubric is able to aid in facilitating language development as a criterion benchmark to quantify the strength and weakness of the learning by accurately gauging for long-term survival. Moskal (2000) noted that pre-defined criteria, indicators, and scaled options eliminate educators' subjective viewing of product quality. Like students, the educators view the rubrics as criteria indicators to provide the assessment score. Such checklists undermine the power of the rubric to be a learning tool to initiate self-reflection and development growth. The rubrics may be used as checklists to define the outcome-product based on indicators (Popham 2005).

However, along with that, rubrics should provide a holistic view of learning as they need to be perceived as tools to help facilitate the overall learning journey. To self-assess, students and teachers need to judge (Zhou and Brown 2015). Facts need to be perceived, questioned, synthesized, and evaluated. Educators need to mentor students to critically gauge their learning. Likewise, educators need to do the same to judge their students' growth through the tasks. When educators discuss the rubrics with students, two things should happen: The learners are expected to follow the model and appreciate what guidance is given to them to master the information and facilitate the learning journey, and they are expected to have a clear understanding of the descriptors to complete the task.

Andrade (2000) noted that rubric pointers are informative and constructive. The rubrics become tools to be used as means-to-an-end and as ends-in-themselves, thus giving the rubric all-inclusive power. Holistically, these two options provide fertile ground to allow both the educators and the students to monitor and self-reflect on the progress and journey development. As the students observe, they redirect their learning to complete what they appraise as missing (Bandura 1977). The rubrics are used to initiate real-life content learning and context application. Popham (2006) noted that the "real payoff" (p. 248) of rubrics is for students to be able to self-evaluate their work and appraise their growth. Educators need to mentor students to use rubrics to do this. Social Learning theory advocates self-perception to monitor progress (Bandura 1977).

In contrast, the traditional test as an assessment tool does not provide educators or learners with the same information. To plan for a test, the educator decides on the content, based on the intended learning outcome, decides on the instructional task and the prompts, questions and answers. The measurement is aligned with the preplanned endeavor (Coombe et al. 2012; Greenberg 2012). When students take a test, the educator is able to measure each student's alignment to and mastery of the learning objectives.

The two assessment tools differ, but they both provide the students with instruction cues to direct their thinking. The traditional test instruction cues and assigned task rubric provided by educators served as "similar" indicators and outcome expectations. The rubric may or may not be used as a means-to-an-end tool, but it will be used as the traditional assessment is—as an end-in-itself. The rubric might not guide or facilitate task completion; in this way, the students are not utilizing the power of the rubrics for holistic content growth for life-long learning.

Behaviorists advocate that stimulus-response associations change learning outcomes (Zhou and Brown 2015). Satisfying achievement scores reinforces responses, shaping desired behaviors. Zhou and Brown (2015) noted that such responses "are broken down into discrete, concrete units, or positive movements, each of which is reinforced as it progresses toward the overall behavioral goal" (p. 10). This is what rubric criteria indicators do; they view responses categorically and measure performance as concrete units satisfying the accomplishment of each criterion.

The critical evaluations come into play when descriptors, criteria, and/or scale details are incomplete or do not direct the consequences and reinforcements modes, making learning appraisal challenging.

To construct real-life learning, students need to add meaning to their S-R action. Simplifying the criteria, indicators and scales may force students to consciously store content knowledge based on purpose. Using single-point rubrics, learners weigh their actions based on reason (Fluckiger 2010). They need to respond intentionally to complete the task as described in the single-scale categories. The scale as provided in the single-point rubrics may be perceived as incomplete, needing more information. So, the students draw knowledge from content stored, enabling the students and the educator to appraise real-life learning. According to Constructivism, the focus on single-point rubrics forces students to be engaged in creating relevant, tangible outcomes (Hanfstingl et al. 2019; Ewing et al. 2011). Similarly, drawing from memory, self-generated rubric criterion or indicators leave room for learners to discover missing knowledge and increase learning efficacy (Bandura 1986). Through the promotion of self-generating from memory, the descriptors or detailed lists that students include complete the construct; these rubrics do not enable learners to measure knowledge that they do not possess. What they are able to do is to appraise content produced by assimilation and construction versus content cues for S-R automatic association.

As part of the initiative to help master and construct rubrics to align with the learning checklist process, educators should mentor students (Popham 2006; Wood et al. 1976). It may be true that providing students with rubrics that are more than a single-point rubric is less stressful since a skeleton descriptor provides the basis to illustrate the purpose of a needed outcome measure; it is the rest of this single-point rubric model that is minimal and requires the learners to show proof of scholarship (Hanfstingl et al. 2019). To facilitate this process of thinking, educators should provide one or more indicators to exemplify what posits the details or groupings within the scale. For example, the educators may show a scale that may represent good and bad indicators or good and very good indicators. It remains an initiative on the part of the learners to decide and complete the rubric to measure achievement. The more concrete the students make the construction, the less challenging the rubric becomes since the guidance provided and the construction of knowledge unfolded collaboratively. Students supply a rubric as an end product and reflect on a model-based approach to demonstrate a constructed long-term learning process and not merely a rubric produced on a recall exercise (Zhou and Brown 2015). Constructed rubrics reflect assimilated learning and need descriptors to identify criteria details and not indicators to mirror recall initiatives.

27.4 Methodology

27.4.1 Design

An Action Research design, specifically the Dialectic Action Research Spiral, was used to investigate the impact of rubrics on learning (Gay et al. 2006). The design is widely used in education and emphasizes action and change. The four step process includes identifying the area of focus, data collecting, data analysis, and action planning before it spirals again with the collected data and plans for the next planning step.

The data collected were the various outcome measures collected from students described across the years. The data served as the insights to spiral the research model upwards addressing the two research topics: student engagement with the rubrics and student use of rubrics holistically to measure learning for real-world language use.

27.4.2 Participants

Participants were graduate and undergraduate students across several years of instruction who part took in my courses. The participants were taking courses in English language, English in the Workplace, and Education and Psychology at major universities in Lebanon.

27.4.3 Instruments and Procedures

Different planned activities were designed for learning. The activities were distributed in accordance with the course curricula. As part of course assignments, participants received classroom or homework activities to complete and submit to benchmark their learning. In line with the course learning outcomes, assignments included letter writing, essays, critical reflections, projects, research papers, essays, interviews, charts, illustrations, film making assignments, and/or achievement tests. In Sect. 27.5, I share examples to support the discussion points.

27.5 Findings

Research question 1. How did students engage with the rubrics to measure learning for real-world language use?

Findings suggest that students did not engage with the rubrics to measure real-world language use. To shed light on this observation, two sub-questions were posited to discuss the impact of the rubric as a guide to task completion and as a facilitator to accomplishing the task. Findings addressing these two sub-questions are discussed together.

a. Did students use rubrics descriptors as the guide to supply their activity answers?
b. Did students' use of rubric indicators and outcome expectations inflate their grades when compared to traditional testing scores? In 2017, the rubric and traditional task achievements of 11 students were compared and showed significantly different results (see Table 27.1). The scores of the achievement test represented the complete content of the course. The same course content was also assessed through the alternative assessment tasks; the assimilation of content was done through the course unit assignments: charts, presentations, illustrations, and projects, as well as reflected through a course content film and a course content portfolio. Each alternative assignment was scored, and the total of the alternative assessment scores were averaged to represent the learners' overall course content learning. Table 27.1 shows the alternative assessments score and the traditional assessment score for each participant. Although the difference in assessment format could have affected some students' performance, my observations of the students' work indicated that they used rubric descriptors as the guide to supply their activity answers, as was seen from the near-perfect outcomes alternative assessment scores. The alternative assessment rubrics were the checklists the students followed to make sure they earned the high marks.

The observation based on the scores was that the students had followed the rubric criteria or focused on following the indicator. They did not "freely" produce their answers. Explicit rubric criteria and defined indicators had not prompted their use of implicit understanding or critical thinking strategies. The rubric had dictated the

Table 27.1 Participants' traditional and alternative assessment scores of assigned tasks

Participant	Traditional Assessment: Achievement test score (10 points)	Alternative Assessments: Average score for Charts, Projects, Illustrations, Presentations, Film, Portfolio (10 points)
Participant 1	6.2	9.5
Participant 2	7.2	9.0
Participant 3	6.0	9.5
Participant 4	5.8	9.9
Participant 5	8.4	8.0
Participant 6	5.8	8.9
Participant 7	6.4	8.0
Participant 8	6.4	10.0
Participant 9	5.2	7.8
Participant 10	6.6	9.0
Participant 11	5.4	9.1

expectation. The stimulus produced the response that I received and scored as near-perfect.

The students' use of rubric indicators and outcome expectations inflated their grades when compared to traditional testing scores (see Table 27.1). The near-perfect rubric-based scores indicated that learners had efficaciously acquired theoretical knowledge when they used criteria indicators to create Charts, Projects, Illustrations, Presentations, Portfolio, and Film as tasks. That was not the case with the traditional assessment scores (Sabieh 2017). The 11 students' test scores ranged from 5.2 to 8.4 out of 10 points; specifically, four students failed the test, five scored 6, and two scored 7 and 8, respectively. The content knowledge transferred from doing the alternative assessment tasks was not acted on in a different assessment context.

By comparing activity scores resulting from the use of explicit rubric criteria and traditional test taking, assessment scores showed that the two tools differed. Traditional standardized testing involved using defined prompts with demarcated answer choices or pathways (Coombe et al. 2012; Genesee and Upshur 1996). The students' responses were measured by the rubric to indicate that the standard had been achieved (Greenberg 2012). My observation was that both traditional and rubric assessment tools provided students with instruction cues to direct their thinking. Traditional test instruction cues and assigned task rubrics provided by educators serve as "similar" indicators and outcome expectations. My observation was that the rubric did not guide learners as they completed their tasks; the rubric was used as the end-in-itself to quantify learning, as the students did in the traditional testing setting. The two tools—rubric and test—produced similar outcomes to benchmark learning; however, the rubric guidance produced higher scaled responses. As observed, the rubric-based grades were inflated; they were observed as considerably higher. My observation was that the rubric enabled many of the students to produce near-perfect assignments. The students followed the rubric focus to ensure they received full marks. However, as seen in the test scores in Table 27.1, the traditional testing condition proved to be a stimulating but more taxing task for the students to complete. Thus, their performance was challenged to a greater extent since the instructions in the test may have assisted in content retrieval but did not guide students' responses to the test items as directly as the rubric descriptors had.

Drawing on the data analyzed to answer Research Question 1, the discussion points have supported the idea that students engaged with the rubrics to measure learning in two ways. First, they used the rubrics to guide task completion, and second, they used the rubric indicators and outcome expectations to facilitate accomplishing the task and the assessment showed student achievement was measured due to criterion alignment but the scores did not reflect a real measure of learning; more simply, they did not measure achievement for real-world learning.

The second purpose of my action research was to further investigate how a rubric could be utilized to measure long-term learning. Research Question 2 addressed the development of real-life strategies through rubric scaffolding.

Research question 2. How did students use rubrics holistically to measure learning for real-world language use?

Students did use rubrics holistically to measure their learning when the rubric did not facilitate task completion. Rubrics with single-point descriptors, self-generated rubric criteria or indicators, and rubrics that facilitated identifying missing knowledge and increased learning efficacy measure learning for real-world purposes.

Rubrics that provide task-specific category pointers to guide task completion and assessment remain popular for task completion. Students' task completion was facilitated by the details. The learners' rubrics were created to meet required holistic thinking and understanding of material to enable task completion. To shed light on the observation, four sub-questions were addressed to discuss the impact of the use of the rubrics holistically for real-world language use.

a. Did students use rubrics holistically to measure learning development when they were focusing on rubrics with task-specific category pointers?

To determine how insignificant holistic rubric use was in the presence of specific category points, I observed the results of the 26 students in my Fall 2016 Argumentative Rhetoric course writing a one-page Response essay that was due after two class sessions. Without informing the students, I sent 13 students the Response Essay Rubric I would be using to grade their essays. During the next class session, five students, unaware that only 13 had been sent the rubrics, informed me that they would need to extend their Response essay due date since they had not received the Response Writing Rubric the others in our class had received. With no rubric, they informed me that they had not started working on the Response task because they had expected to receive the rubric for guidance or to facilitate the task of writing the response essay. However, seven out of the 13 students informed me that they had taken the initiative to take the Rubric from the classmates they knew had received it. One student, unaware of the Rubric or of the chaos the Rubric had caused, said she had started working on her response essay. I also observed that the 25 students expected to develop their essays using the Rubric. The rubric was to serve as a checklist for their task-specific response essay so as to maximize their score. None of the 25 students took the initiative to work without the rubric.

To further illustrate the influence of rubric dependency over holistic rubric use for learning growth, I collected data from my 23 Argumentative Rhetoric students in 2017. That semester, the Course Coordinator had asked the instructors to share rubrics with their students when assigning tasks. Sharing rubrics was not a common practice in that course. So, for the first assignment, I decided to give the students their task but not have a rubric accompany the assignment; I wanted to see how my students would behave, knowing that students in other sections were given a rubric with their homework. Seventeen students reacted and asked about the rubric the other students had received. They questioned the fairness of the assigned task, and they found it discriminatory that the other students were given guidance—the specific descriptors to pave "the way" to answer their homework and receive higher grades on the task. Again, what transpired was a focus on the specific guidance to complete the task and not on the holistic value of using the rubric to gain knowledge. The second illustration showed that the students' concern was on acquiring the rubric to complete the task.

b. Did students use rubrics to assist them in monitoring their learning development?

When 15 graduate students in my Advanced Educational Psychology course were asked if they used rubrics to measure learning growth, 10 answered they did not. They used the rubrics to do requested assignments. The rubrics were not considered as self-reflection tools. However, when the same 15 students were given a task as a Portfolio assignment to discuss their weekly learning growth, they, then, used the rubric to monitor progress.

The use of the rubric to monitor learning development was dependent on the task itself. Developing a task that required students to monitor their progress and forced them to self-reflect as part of the task, made the students partake in the exercise. Students were not able to self-evaluate their learning and monitor their progress.

In addition, over the past four years, 300 students' performances were assessed in simulated job interviews as part of my English in the workplace course. Using the Job Interview Scoring Rubrics, students were assessed on skills, content, verbal and behavior expectations, and dress code (see Fig. 27.1).

98% of the students received full scores. The students had acted with intent: they had researched companies, practiced interview questions, watched interviews on YouTube, and came prepared to the interview. During the interview, they did what the indicators (stimulus) prescribed and received maximum scores. The task was straightforward. I had assumed that they had performed effectively for real-life learning. This was not the case.

Since the simulated interviews, around 25% of those students have passed by my office during their senior year to help them prepare for an authentic job interview. Thus, this is an indication that the simulation interview—perfectly assessed semesters prior—was not assimilated for future real-life use.

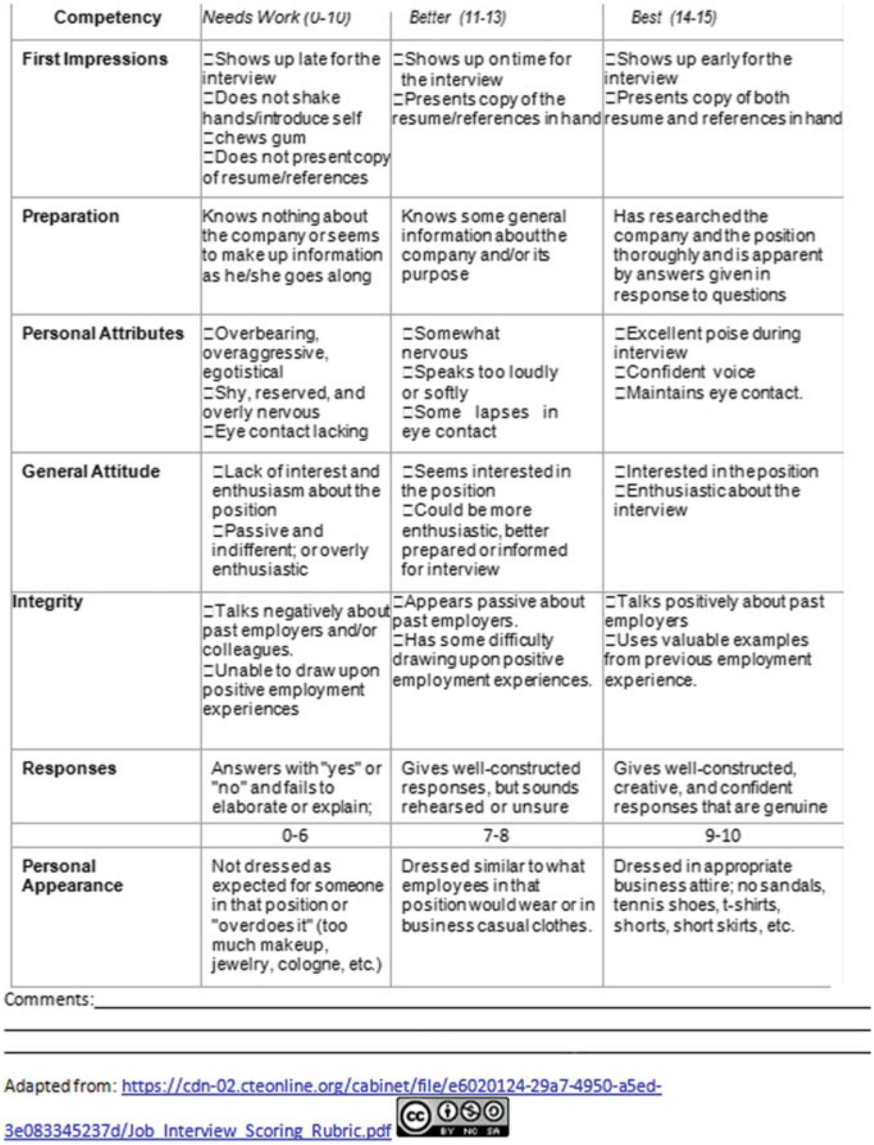
In short, the value of acquiring job-interview skills and preparing for the assignment were two segmented chunks needed for the short-term course task. Even though they had received full interview scores, those students had not stored the information for future more authentic use. They used the rubrics equivocally to construct their behavior during the interview.

c. Did students use rubrics with single-point descriptors to facilitate tangible holistic real-life learning?

Using simple rubrics in a business English course, I helped students construct knowledge for short- and long-term use. Students were asked to use PAIBOC—Purpose, Audience, Interest, Benefit, Objection, and Context—when writing (see Fig. 27.2) and assess using the simple checklist, scaled 1 or 2 for each present indicator (see Fig. 27.3).

I had my 23 students describe themselves as products/brands on a supermarket shelf waiting to be bought. (See examples of written assignments in Figs. 27.4, 27.5, and 27.6).

These three writing samples reflected use of PAIBOC, but each criterion (P, A, I, B, O, C) reflected a separate complete entity in itself. Fragmented, the communication points of P, I, B, O were there but lacked unity and coherence. Thus, students had



Competency	*Needs Work (0-10)*	*Better (11-13)*	*Best (14-15)*
First Impressions	□Shows up late for the interview □Does not shake hands/introduce self □chews gum □Does not present copy of resume/references	□Shows up on time for the interview □Presents copy of the resume/references in hand	□Shows up early for the interview □Presents copy of both resume and references in hand
Preparation	Knows nothing about the company or seems to make up information as he/she goes along	Knows some general information about the company and/or its purpose	Has researched the company and the position thoroughly and is apparent by answers given in response to questions
Personal Attributes	□Overbearing, overaggressive, egotistical □Shy, reserved, and overly nervous □Eye contact lacking	□Somewhat nervous □Speaks too loudly or softly □Some lapses in eye contact	□Excellent poise during interview □Confident voice □Maintains eye contact.
General Attitude	□Lack of interest and enthusiasm about the position □Passive and indifferent; or overly enthusiastic	□Seems interested in the position □Could be more enthusiastic, better prepared or informed for interview	□Interested in the position □Enthusiastic about the interview
Integrity	□Talks negatively about past employers and/or colleagues. □Unable to draw upon positive employment experiences	□Appears passive about past employers. □Has some difficulty drawing upon positive employment experiences.	□Talks positively about past employers □Uses valuable examples from previous employment experience.
Responses	Answers with "yes" or "no" and fails to elaborate or explain;	Gives well-constructed responses, but sounds rehearsed or unsure	Gives well-constructed, creative, and confident responses that are genuine
	0-6	7-8	9-10
Personal Appearance	Not dressed as expected for someone in that position or "overdoes it" (too much makeup, jewelry, cologne, etc.)	Dressed similar to what employees in that position would wear or in business casual clothes.	Dressed in appropriate business attire; no sandals, tennis shoes, t-shirts, shorts, short skirts, etc.

Comments:____

____

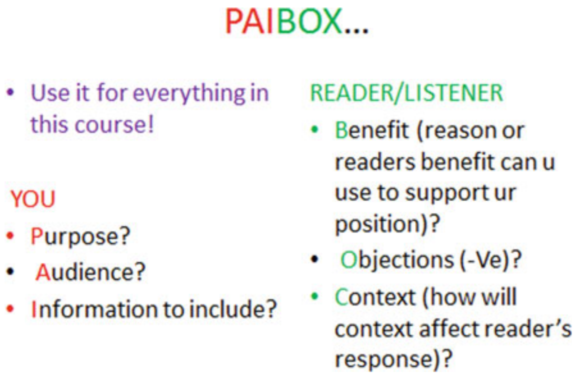
____

Adapted from: https://cdn-02.cteonline.org/cabinet/file/e6020124-29a7-4950-a5ed-3e083345237d/Job_Interview_Scoring_Rubric.pdf

Fig. 27.1 Job Interview Scoring Rubric (2014) used in English in the workplace course

retrieved the primitive rubric from their storage and used the checklist to prompt segregated production of criteria indicators. Students did not connect the criterion to assimilate message unity.

Fig. 27.2 PAIBOC indicators



PAIBOX…

- Use it for everything in this course!

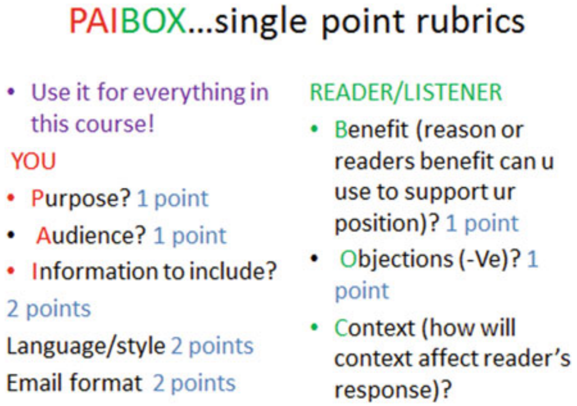
YOU

- Purpose?
- Audience?
- Information to include?

READER/LISTENER

- Benefit (reason or readers benefit can u use to support ur position)?
- Objections (-Ve)?
- Context (how will context affect reader's response)?

Fig. 27.3 Single-point rubric tool



PAIBOX…single point rubrics

- Use it for everything in this course!

YOU

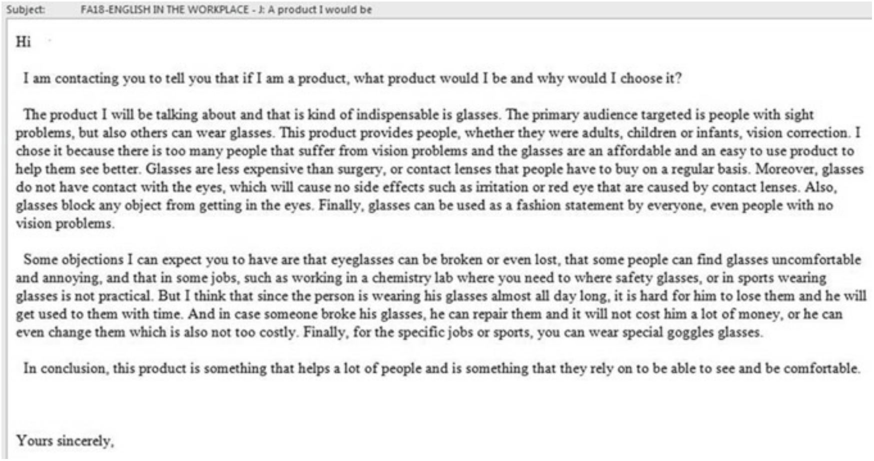
- Purpose? 1 point
- Audience? 1 point
- Information to include? 2 points

Language/style 2 points

Email format 2 points

READER/LISTENER

- Benefit (reason or readers benefit can u use to support ur position)? 1 point
- Objections (-Ve)? 1 point
- Context (how will context affect reader's response)?



Subject: FA18-ENGLISH IN THE WORKPLACE - J: A product I would be

Hi

I am contacting you to tell you that if I am a product, what product would I be and why would I choose it?

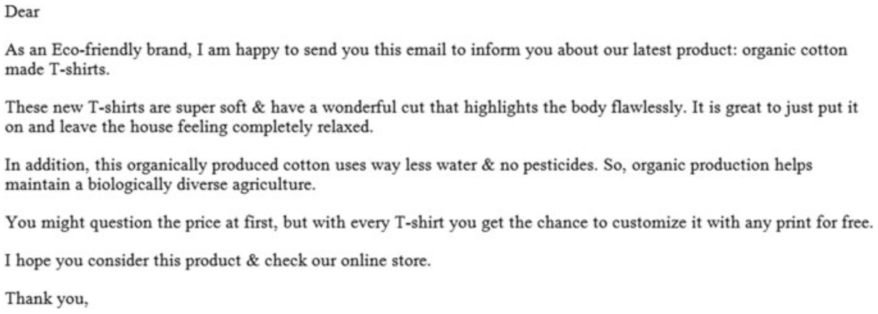
The product I will be talking about and that is kind of indispensable is glasses. The primary audience targeted is people with sight problems, but also others can wear glasses. This product provides people, whether they were adults, children or infants, vision correction. I chose it because there is too many people that suffer from vision problems and the glasses are an affordable and an easy to use product to help them see better. Glasses are less expensive than surgery, or contact lenses that people have to buy on a regular basis. Moreover, glasses do not have contact with the eyes, which will cause no side effects such as irritation or red eye that are caused by contact lenses. Also, glasses block any object from getting in the eyes. Finally, glasses can be used as a fashion statement by everyone, even people with no vision problems.

Some objections I can expect you to have are that eyeglasses can be broken or even lost, that some people can find glasses uncomfortable and annoying, and that in some jobs, such as working in a chemistry lab where you need to where safety glasses, or in sports wearing glasses is not practical. But I think that since the person is wearing his glasses almost all day long, it is hard for him to lose them and he will get used to them with time. And in case someone broke his glasses, he can repair them and it will not cost him a lot of money, or he can even change them which is also not too costly. Finally, for the specific jobs or sports, you can wear special goggles glasses.

In conclusion, this product is something that helps a lot of people and is something that they rely on to be able to see and be comfortable.

Yours sincerely,

Fig. 27.4 Student X's writing using PAIBOC



Dear

As an Eco-friendly brand, I am happy to send you this email to inform you about our latest product: organic cotton made T-shirts.

These new T-shirts are super soft & have a wonderful cut that highlights the body flawlessly. It is great to just put it on and leave the house feeling completely relaxed.

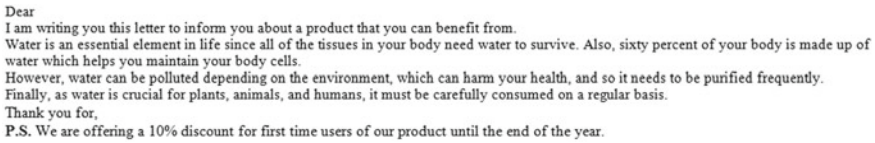
In addition, this organically produced cotton uses way less water & no pesticides. So, organic production helps maintain a biologically diverse agriculture.

You might question the price at first, but with every T-shirt you get the chance to customize it with any print for free.

I hope you consider this product & check our online store.

Thank you,

Fig. 27.5 Student Y's writing using PAIBOC



Dear
I am writing you this letter to inform you about a product that you can benefit from.
Water is an essential element in life since all of the tissues in your body need water to survive. Also, sixty percent of your body is made up of water which helps you maintain your body cells.
However, water can be polluted depending on the environment, which can harm your health, and so it needs to be purified frequently.
Finally, as water is crucial for plants, animals, and humans, it must be carefully consumed on a regular basis.
Thank you for,
P.S. We are offering a 10% discount for first time users of our product until the end of the year.

Fig. 27.6 Student Z's writing using PAIBOC

So to minimize rubric criterion use as S-R, promote message unity, and increase production monitoring, I mentored students in this way: consider content, comprehend, view rubric criteria/indicators, comprehend, write task with no rubric, assess task, provide constructive feedback, recall rubric to retrieve criteria, rewrite task, and assess writing using the self-generated rubric. The scaffold had students focus on content construction, meaning, and rubric criteria recall, creating unified and coherent messages. Figures 27.7, 27.8, 27.9, 27.10, 27.11, and 27.12 illustrate pre-post scaffold writing progression and self-created rubric prompts to show message construction.

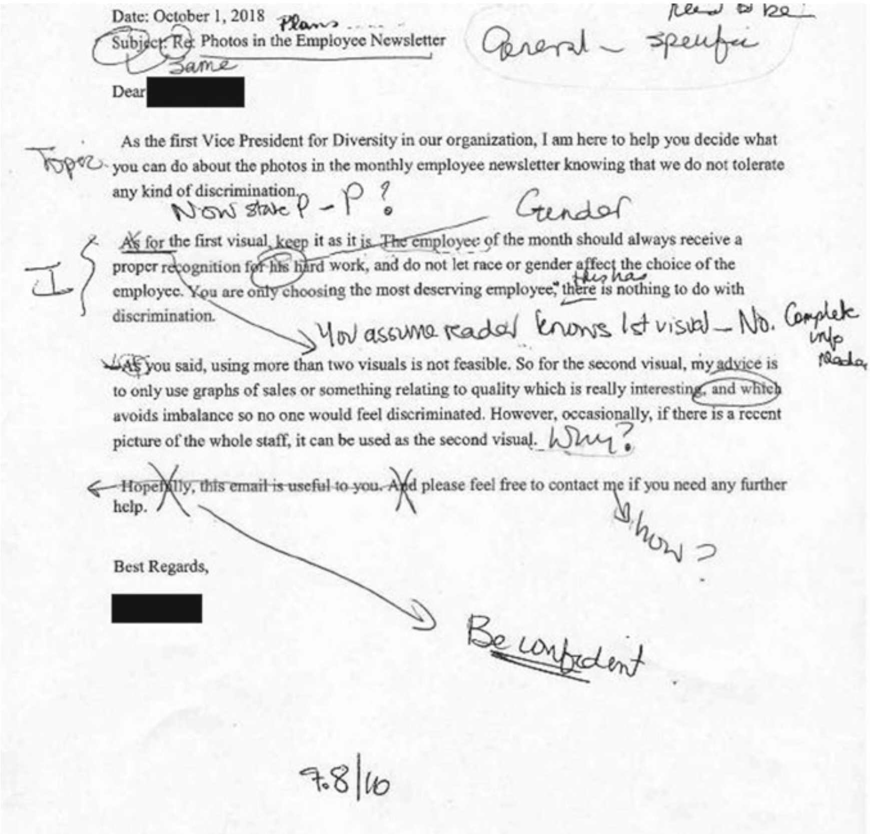
Students re-produced the single-point rubrics as checklists based on how they understood content and on what information they felt needed focus. Notice how student A's and student B's rubric checklists differed, reminding them that missing content still needed mastery and measurement.

As a group of learners, they cultivated their knowledge in a dynamic class environment with the educator as chaperone. In line with Social Constructivism, I mentored the students through imitation, collaborative, and discovery learning and had them use the rubrics to measure learning (Baker 2018; Bandura 1986; Vygotsky 1978).

I share two more student-created rubrics in Figs. 27.13 and 27.14.

Notice that all four rubrics had different pointers. This demonstrates that students learned differently, and the prompts recalled information based on its relevance to them and the context. The criteria and indicator details were absent.

d. Did students' use of self-generated rubric criteria or indicators facilitate identifying their missing knowledge and increase their learning efficacy?



Date: October 1, 2018 Plans ....

Subject: Re: Photos in the Employee Newsletter

Same

Need to be

General – specific

Dear [redacted]

Topic As the first Vice President for Diversity in our organization, I am here to help you decide what you can do about the photos in the monthly employee newsletter knowing that we do not tolerate any kind of discrimination.

Now state P – P ?

Gender

I As for the first visual, keep it as it is. ~~The employee~~ of the month should always receive a proper recognition for his hard work, and do not let race or gender affect the choice of the employee. You are only choosing the most deserving employee; this has there is nothing to do with discrimination.

You assume reader knows 1st visual – No. Complete info Reader

As you said, using more than two visuals is not feasible. So for the second visual, my advice is to only use graphs of sales or something relating to quality which is really interesting, and which avoids imbalance so no one would feel discriminated. However, occasionally, if there is a recent picture of the whole staff, it can be used as the second visual. Why?

~~Hopefully, this email is useful to you.~~ And please feel free to contact me if you need any further help.

how?

Best Regards,

[redacted]

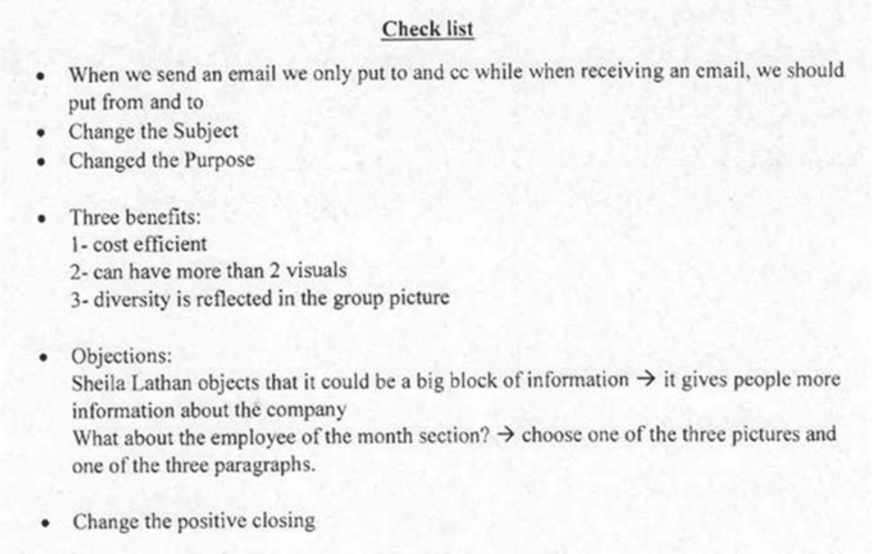
Be confident

9.8|10

Fig. 27.7 Student A's first draft with feedback

Students' use of self-generated rubric criteria or indicators facilitated identifying their missing knowledge and increased their learning efficacy. Advocating for this to happen in all teaching/learning settings, I have observed that learners are forced to take on active roles. Learners are committed to complete learning responsibilities. Acknowledging that these rubrics were self-generated from memory, I note that they did not enable learners to measure what they did not know. The list prompted learners to remember what needed to be included in the context.

Rubrics with pre-defined criteria, indicators and scales identified what learners did not know. Although the learning experience was not authentic, had limited content exploration and self-regulated student learning when the rubrics were provided to guide tasks, as post-task assessment tools, they may have been used to brainstorm and explore what learners did not know (e.g. Palacios et al. 2018). Looking at the self-generated rubrics in the figures, we see the different cues the four students identified to achieve the same task. Students did not provide details, just prompts. Rubric details were needed in the learning process but not in the recall process.



Check list

- When we send an email we only put to and cc while when receiving an email, we should put from and to
- Change the Subject
- Changed the Purpose

- Three benefits:
  1- cost efficient
  2- can have more than 2 visuals
  3- diversity is reflected in the group picture

- Objections:
  Sheila Lathan objects that it could be a big block of information → it gives people more information about the company
  What about the employee of the month section? → choose one of the three pictures and one of the three paragraphs.

- Change the positive closing

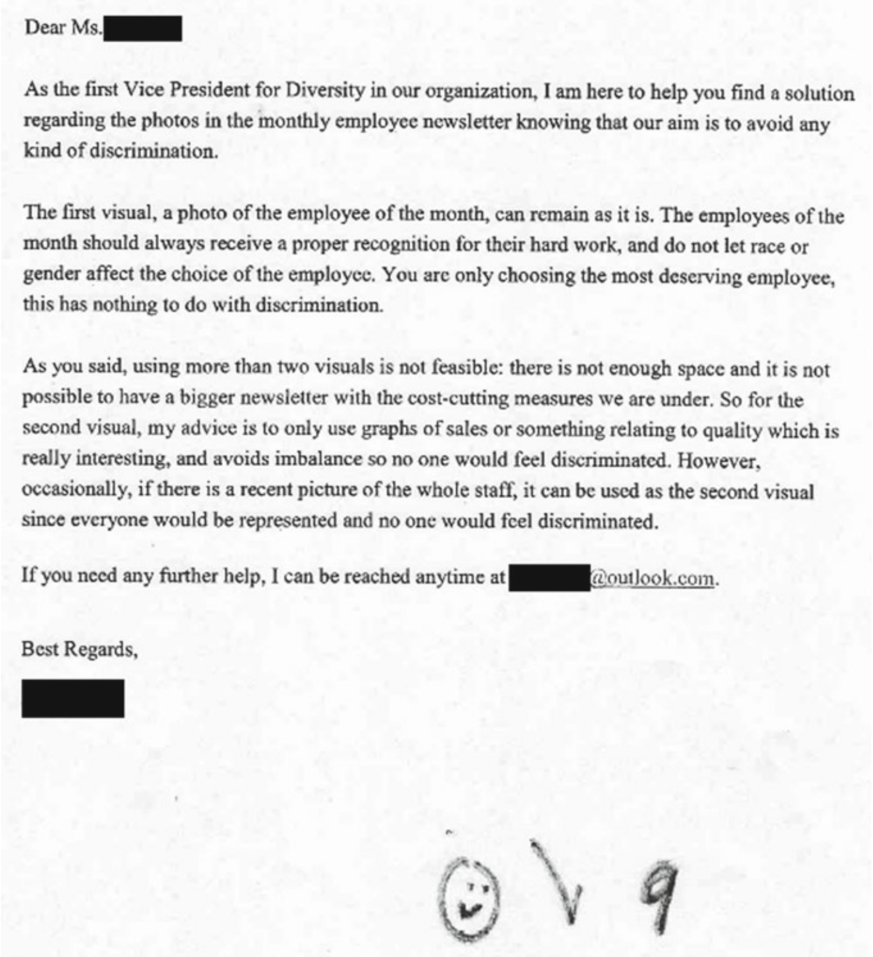
Fig. 27.8 Student A's self-created rubric tool prompt

e. Did students' use of rubrics that provide criteria but describe indicators posited at opposite ends of a scale assist in developing real-life learning?

Students had to brainstorm to determine the end product. Rubrics that provided criteria and described indicators posited at opposite ends of a scale assisted them in developing real-life learning.

I have illustrated that the detailed rubrics did not challenge short-term learning or task completion. However, did rubrics that did not include details provoke a different learning environment? With limited guidance, learners brainstormed to optimize the achievement scale. This initiative will increase meaning, solidify knowledge acquisition, and transfer learning into long-term memory. This semester, 25 students were provided with such rubrics to write critical reflections. The students were able to do the needed brainstorming based on the two indicators at opposite ends of the scale and generate their critical reflections successfully. The rubric demarked the criteria and included scales to describe two conditions—mastery and below-average learning (see Fig. 27.15). The students created their assignments keeping the indicators in mind.

More detailed than single-point rubrics, the rubric still guided the learners; however, the students were forced to demonstrate in-depth content synthesis to generate the outcome. The 25 students' work had mirrored interactivity with criteria and indicators to reflect their personal learning growth. The critical reflections were not developed to meet specific indicators but to meet the overall criteria. The outcomes were not shallow and fragmented to warrant inflated achievement scores.



Dear Ms.[redacted]

As the first Vice President for Diversity in our organization, I am here to help you find a solution regarding the photos in the monthly employee newsletter knowing that our aim is to avoid any kind of discrimination.

The first visual, a photo of the employee of the month, can remain as it is. The employees of the month should always receive a proper recognition for their hard work, and do not let race or gender affect the choice of the employee. You are only choosing the most deserving employee, this has nothing to do with discrimination.

As you said, using more than two visuals is not feasible: there is not enough space and it is not possible to have a bigger newsletter with the cost-cutting measures we are under. So for the second visual, my advice is to only use graphs of sales or something relating to quality which is really interesting, and avoids imbalance so no one would feel discriminated. However, occasionally, if there is a recent picture of the whole staff, it can be used as the second visual since everyone would be represented and no one would feel discriminated.

If you need any further help, I can be reached anytime at [redacted]@outlook.com.

Best Regards,

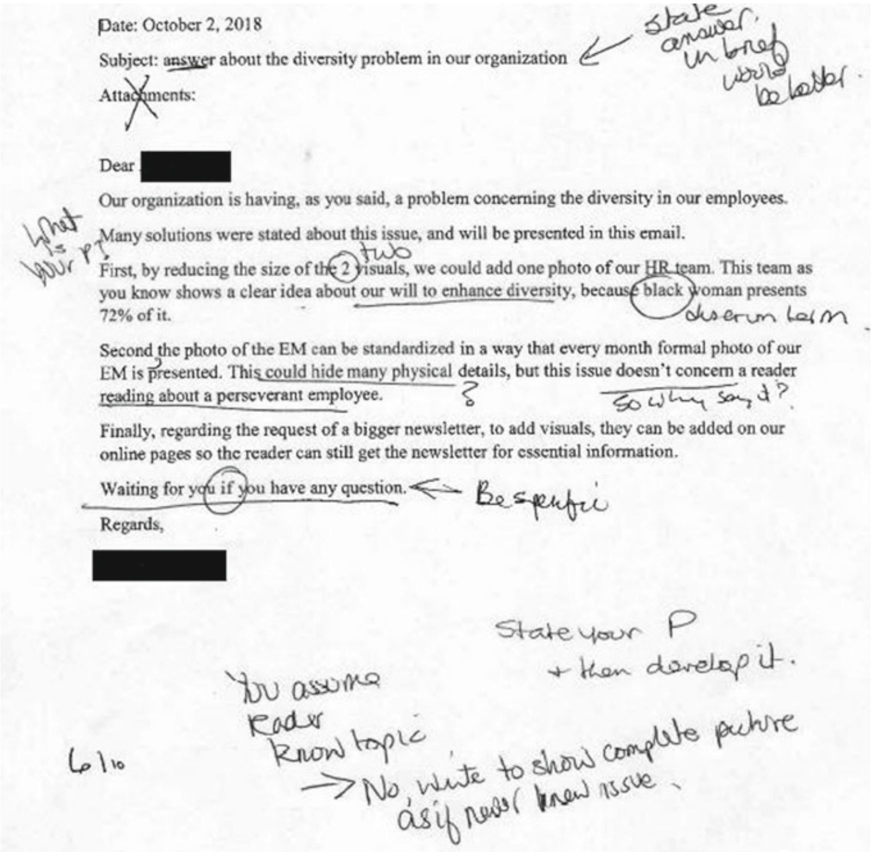
[redacted]

☺ ✓ 9

Fig. 27.9 Student A's second draft

Thus, rubrics needed to provide criteria, but they did not need to provide numerous indicators.

When rubrics do not provide content context cues that involve learning details, learners are forced to take on more responsible roles in aligning their learning growth to their needs. Rubrics facilitate learning growth when students and educators use them as a means to enhance the learning.



Date: October 2, 2018

Subject: answer about the diversity problem in our organization ← State answer, in brief would be better.

~~Attachments:~~

Dear [redacted]

Our organization is having, as you said, a problem concerning the diversity in our employees.

What is your pt? Many solutions were stated about this issue, and will be presented in this email.

First, by reducing the size of the 2 visuals, we could add one photo of our HR team. This team as you know shows a clear idea about our will to enhance diversity, because black woman presents 72% of it.

two

diverse team

Second the photo of the EM can be standardized in a way that every month formal photo of our EM is presented. This could hide many physical details, but this issue doesn't concern a reader reading about a perseverant employee. ? So why say it?

Finally, regarding the request of a bigger newsletter, to add visuals, they can be added on our online pages so the reader can still get the newsletter for essential information.

Waiting for you if you have any question. ← Be specific

Regards,

[redacted]

State your P + then develop it.

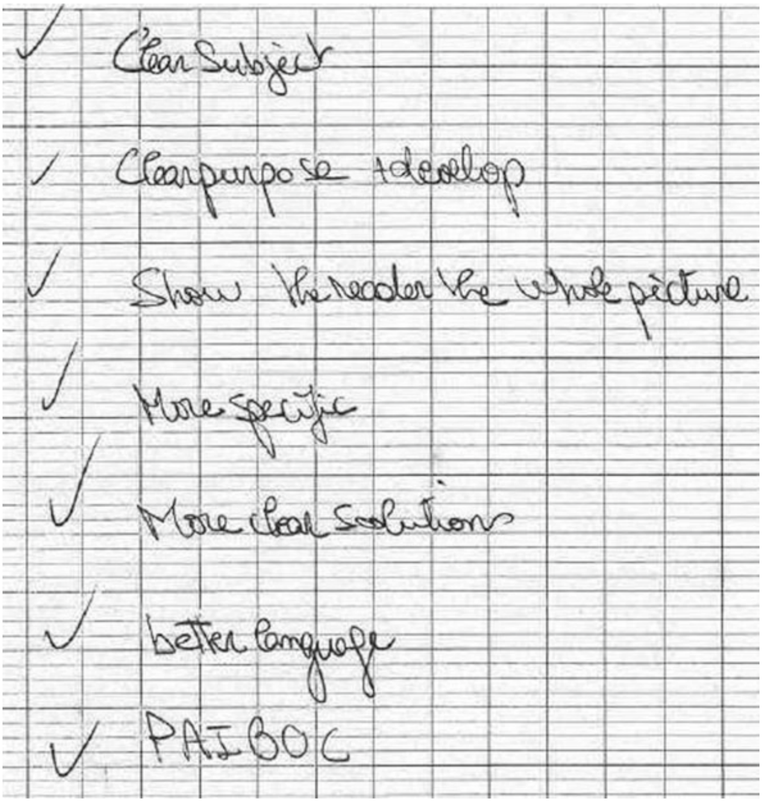
You assume reader know topic → No, write to show complete picture as if never knew issue.

6/10

Fig. 27.10 Student B's first draft with feedback

27.6 Insights Gained

Rubrics with detailed, multiple-level descriptors inflated scores and measured minimal critical thinking. The use of the rubric limited what Constructivists advocated to be free-willed connectivity and learning growth since students used the rubric descriptors as the guide to supply their answers. Based on the completed assignments, it is evident that students tended to create answers based on rubric criteria indicators: They responded to meet the desired scale and receive the achievement score as a reward. The focus became what Skinner (1989) explained to be a stimulus-response arrangement (Zhou and Brown 2015). Thus, the behavioral act had no need for meaningful free-willed assembly to respond to the rubric-based task. The students processed the requirements and provided me with the outcomes based on the expectancies defined by the rubric descriptors. The rubric stimulated the production of rubric-level-focused learning. The rubric I provided had not given



✓ Clear Subject

✓ Clear purpose to develop

✓ Show the reader the whole picture

✓ More specific

✓ More clear Solutions

✓ better language

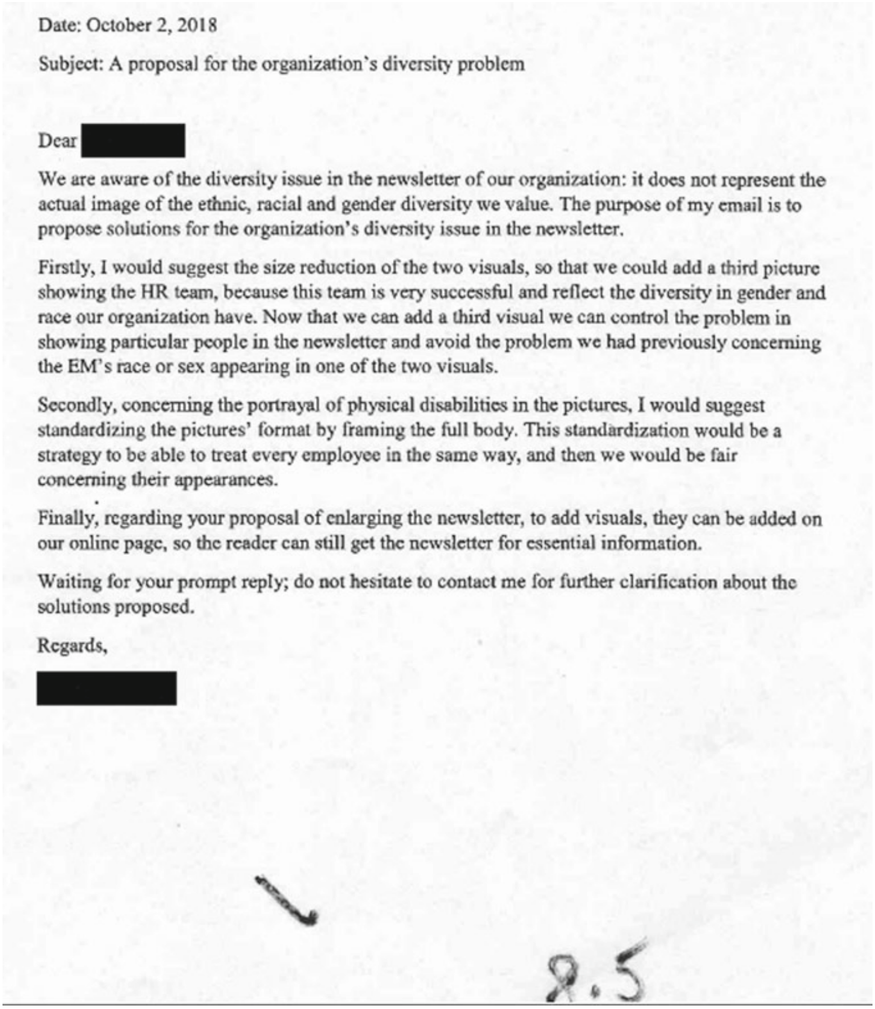
✓ PAIBOC

Fig. 27.11 Student B's self-created rubric tool prompt

room for creative diverse associations of varied answers that showed a transformed understanding of content knowledge. Their answers were not "freely" designed to show self-initiated construction.

Drawing from the literature on learning theories, I concluded that the students' content learning was for short-term use and decayed or was displaced due to limited storage capacity, since the detailed rubric guidance did not require them to think and create associations. The students focused and produced pathways for successful task completion; however, they were not able to retrieve content successfully to respond error-free in the test items. The learning was short-term since the grades on the test were low; content assimilation had not sufficiently established long-term connections for retrieval success. Thus, the rubric-focused guidance resulted in score inflation—a picture that did not prove to be a valid measure of real-life learning. The task outcomes were produced based on rubric indicator pathways.

In general, the impact of a rubric as a guide to promote learning and to more accurately measure what students actually knew as they engaged successfully for real-world language use was best accomplished through scaffolding. Students built



Date: October 2, 2018

Subject: A proposal for the organization's diversity problem

Dear [redacted]

We are aware of the diversity issue in the newsletter of our organization: it does not represent the actual image of the ethnic, racial and gender diversity we value. The purpose of my email is to propose solutions for the organization's diversity issue in the newsletter.

Firstly, I would suggest the size reduction of the two visuals, so that we could add a third picture showing the HR team, because this team is very successful and reflect the diversity in gender and race our organization have. Now that we can add a third visual we can control the problem in showing particular people in the newsletter and avoid the problem we had previously concerning the EM's race or sex appearing in one of the two visuals.

Secondly, concerning the portrayal of physical disabilities in the pictures, I would suggest standardizing the pictures' format by framing the full body. This standardization would be a strategy to be able to treat every employee in the same way, and then we would be fair concerning their appearances.

Finally, regarding your proposal of enlarging the newsletter, to add visuals, they can be added on our online page, so the reader can still get the newsletter for essential information.

Waiting for your prompt reply; do not hesitate to contact me for further clarification about the solutions proposed.

Regards,

[redacted]

✓

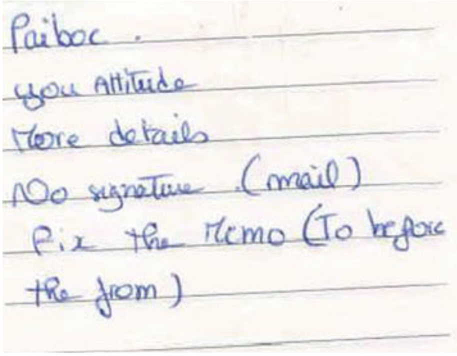
8.5

Fig. 27.12 Students B's second draft

self-reflective judgments of when to use the rubric as a means to accomplish the task to show holistic learning.

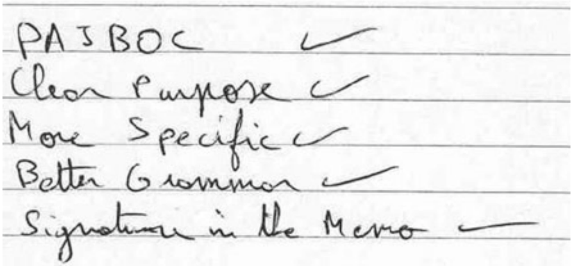
Students did not use rubrics holistically to measure learning development when they were focused on using the rubrics to complete tasks. The rubrics were used by the students to complete tasks based on adhering to the task-specific category pointers. Thus they could ensure completion of the task, resulting in assessment that they had accurately fulfilled assignment requirements. It is true that rubrics should be used because of their power to help develop learning and to facilitate specific guidance to make the specific learning accountable. However, in the observed assignments, the rubric descriptors provided the teacher or the learners with such detailed information

Fig. 27.13 Student C's self-created rubric checklist prompt

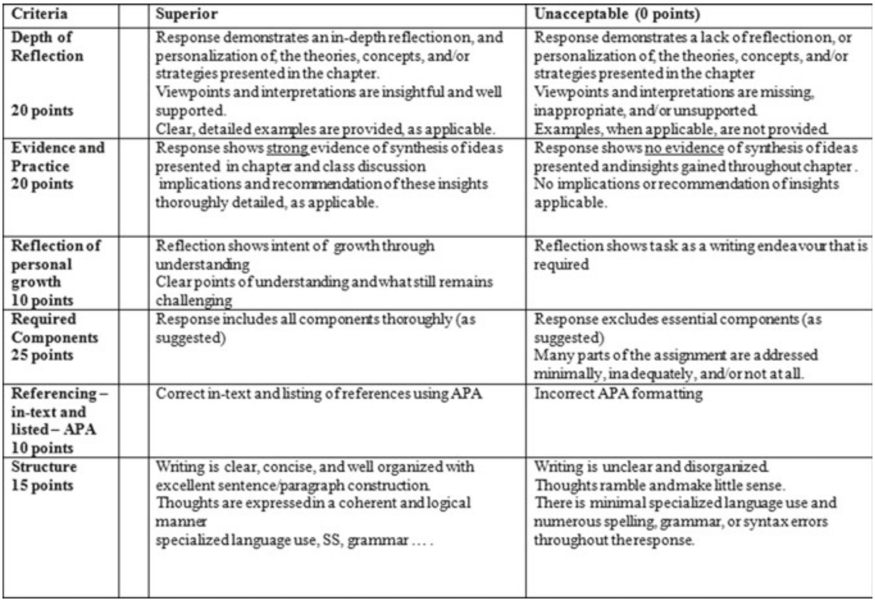


Paiboc .
you Attitude
More details
No signature (mail)
Fix the memo (To before
the from)

Fig. 27.14 Student D's self-created rubric checklist prompt



PAIBOC ✓
Clear Purpose ✓
More Specific ✓
Better Grammar ✓
Signature in the Memo ✓



Criteria		Superior	Unacceptable (0 points)
Depth of Reflection 20 points		Response demonstrates an in-depth reflection on, and personalization of, the theories, concepts, and/or strategies presented in the chapter. Viewpoints and interpretations are insightful and well supported. Clear, detailed examples are provided, as applicable.	Response demonstrates a lack of reflection on, or personalization of, the theories, concepts, and/or strategies presented in the chapter Viewpoints and interpretations are missing, inappropriate, and/or unsupported. Examples, when applicable, are not provided.
Evidence and Practice 20 points		Response shows strong evidence of synthesis of ideas presented in chapter and class discussion implications and recommendation of these insights thoroughly detailed, as applicable.	Response shows no evidence of synthesis of ideas presented and insights gained throughout chapter. No implications or recommendation of insights applicable.
Reflection of personal growth 10 points		Reflection shows intent of growth through understanding Clear points of understanding and what still remains challenging	Reflection shows task as a writing endeavour that is required
Required Components 25 points		Response includes all components thoroughly (as suggested)	Response excludes essential components (as suggested) Many parts of the assignment are addressed minimally, inadequately, and/or not at all.
Referencing – in-text and listed – APA 10 points		Correct in-text and listing of references using APA	Incorrect APA formatting
Structure 15 points		Writing is clear, concise, and well organized with excellent sentence/paragraph construction. Thoughts are expressedin a coherent and logical manner specialized language use, SS, grammar	Writing is unclear and disorganized. Thoughts ramble and make little sense. There is minimal specialized language use and numerous spelling, grammar, or syntax errors throughout the response.

Fig. 27.15 Rubric tool for critical reflection paper

so as to facilitate direct scoring by the teacher or guarantee successful meeting of task requirements. The focus on promoting student learning growth and using the rubric holistically to assist in students' learning of the content long-term became insignificant (Moskal 2000). As evidenced in the literature, both students and teachers want rubrics to align the assignment work with task completion (Popham 1997, 2005).

Students did not initially know how to use the rubrics to assist them in monitoring their learning. When the students were told how to use the rubric for that purpose with an assignment, they were able to use the rubric to monitor their learning. Students needed mentoring to practice monitoring, but they also needed less detailed rubrics to minimize dependency. When rubrics were detailed enough to facilitate content completion, students always opted for the use of the rubric as an end-in-itself facilitator.

Left alone, learners did not self-evaluate and monitor their learning because they were never taught the value of doing so (Popham 2006; Bandura 1977). The use of the rubrics remained popular for successful task completion.

Students used rubrics with single-point descriptors to facilitate tangible holistic real life learning because detailed descriptors were not present and the single-point descriptors did not make sense on their own. When rubrics were made simple, students had to become responsible learners in charge of their own learning challenges. Rubrics with single-point descriptors were ideal to set up proactive learning situations. Using single-point rubrics, learners weighed their actions based on reason and had to identify, add, and modify criteria to self-generate completion of the rubric (Fluckiger 2010).

The common use of the rubrics as a directed checklist is limiting. Instead, educators should have students create self-generated rubrics that assimilate the criteria needed to create the outcome, measure knowledge, and measure real-world learning. In my view, rubrics should do more than assess a task and show the learning outcome. Rubric criteria, indicators and scale details should measure acquired real-life knowledge.

I appeal to educators to consider the impact of rubric use in their learning spaces. I argue for caution when using rubrics with detailed, multiple-level descriptors that inflate scores and measure minimal critical thinking when used as a checklist.

This action research resulted in a successful advocacy on the impact of the rubric. I shared data and observations of authentic cases to explore the two main research questions. I clarified and shared six quandaries:

- Compared to traditional assessments, rubric guidance and post-task scores create grounds for discussing short-term and long-term information processing.
- Rubrics may reflect shallow, short-term learning growth.
- Planning rubric components may reflect constructivism but promote behavioral S-R learning practice.
- Rubric use with scaffold mentoring portrays real-life learning accountability. Mentoring students will enable them to use the rubrics to monitor their learning progress and provide more challenging opportunities to compose. Thus, knowing how to use the rubric effectively becomes key.

- Critical thinking, self-directed rubrics promote self-reflected learning and measure constructive meaning.
- Rubric criteria, minimal indicators and scales specification promote effective real-life learning.

27.7 Conclusion: Implications for Test Users

Self-generated rubric criteria or indicators, single-point descriptors, and rubrics that facilitate identifying missing knowledge are the three best types of rubrics to use to measure learning. I believe the common practices of rubrics in education are minimizing the measure of real-world language learning since the outcomes are guided by descriptors that provide stimulus–response learning. In conclusion, it is essential to define the functionality of rubrics in the teaching-learning-assessment paradigm and determine their impact in education. When students use rubrics to complete tasks, the resulting grade is often inflated and does not measure authentic learning; the actual learning is fragmented, short-term, and shallow. When teachers plan dynamic discovery learning spaces, students should use rubrics to guide them in critical thinking initiatives to meet content criteria and construct in-depth learning reflections that indicate real-life learning.

As part of the learning curve, teachers should plan detailed rubrics to reflect standardized learning outcomes that illustrate short-term learning purposes. Then, teachers should modify rubric use and create tasks to measure self-monitoring and promote critical reflection on stored information and diverse task-production.

Based on the results of this action research, it is recommended that teachers use rubrics differently during certain phases of the learning process to maximize student learning and facilitate assessment.

References

Andrade, H. G. (2000). Using rubrics to promote thinking and learning: What do we mean by results? *Educational Leadership, 57*(5), 13–18. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.452.5684&rep=rep1&type=pdf. Accessed 30 Dec 2019.

Baker, A. (2017). Informational processing theory for the classroom. In M. Zhou & D. Brown (Eds.), *Educational learning theories* (pp. 117–120). https://oer.galileo.usg.edu/cgi/viewcontent.cgi?article=1000&context=education-textbooks. Accessed 30 Dec 2019.

Baker, A. (2018). Discovery learning: Zombie, phoenix, or elephant? *Instructional Science, 46*(1), 169–183.

Bandura, A. (1977). *Social learning theory*. New York: General Learning Press.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Upper Saddle River, NJ: Prentice Hall.

Brooks, J. G., & Brooks, M. G. (1993). *In search of understanding: The case of constructivist classrooms*. Alexandria, VA: Association for Supervision and Curriculum Development.

Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. Ann Arbor, MI: University of Michigan Press.

Coombe, C., Purmensky, K., & Davidson, P. (2012). Alternative assessment in language education. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 147–155). New York: Cambridge University Press.

Ewing, J. C., Foster, D. D., & Whittington, M. S. (2011). Explaining student cognition during class sessions in the context: Piaget's theory of cognitive development. *NACTA Journal, 55*(1), 68–75.

Fluckiger, J. (2010). Single point rubric: A tool for responsible student self-assessment. *Education Faculty Publications, 5*. https://digitalcommons.unomaha.edu/tedfacpub/5. Accessed 30 Dec 2019.

Gay, L. R., Mills, G. E., & Airasian, P. (2006). *Educational research*. Upper Saddle River, NJ: Pearson.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Series editor: J. Richards. New York: Cambridge University Press.

Greenberg, I. (2012). ESL needs analysis and assessment in the workplace. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 178–186). New York: Cambridge University Press.

Hanfstingl, B., Benke, G., & Zhang, Y. (2019). Comparing variation theory with Piaget's theory of cognitive development: More similarities than differences? *Educational Action Research*. https://doi.org/10.1080/09650792.2018.1564687.

Job interview scoring rubric. (2014). https://cdn-02.cteonline.org/cabinet/file/e6020124-29a7-4950-a5ed-3e083345237d/Job_Interview_Scoring_Rubric.pdf. Accessed 30 Dec 2019.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 33*(1), 49–64.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation, 7*(3). https://pareonline.net/getvn.asp?v=7&n=3&sa=U&ei. Accessed 30 Dec 2019.

Palacios, M. G., Shabel, P., Horn, A., & Castorina, J. A. (2018). Uses and meanings of "context" in studies on children's knowledge: A viewpoint from anthropology and constructivist psychology. *Integrative Psychological and Behavioral Science, 52*(2), 191–208.

Popham, W. J. (1997, October). What's wrong—and what's right—with rubrics. *Educational Leadership*, 72–75. http://skidmore.edu/assessment/handbook/Popham_1997_Whats-Wrong_and-Whats-Right_With-Rubrics.pdf. Accessed 30 Dec 2019.

Popham, W. J. (2005). *Classroom assessment: What students need to know* (4th ed.). Boston: Pearson Education Inc.

Popham, W. J. (2006). *Assessment for educational leaders*. Boston: Pearson Education Inc.

Sabieh, C. (2017). Flip! Just make sure you assess learning effectively. In C. Coombe, P. Davidson, D. Boraie, S. Hidri, & A. Gebril (Eds.), *Language assessment in the Middle East and North Africa: Theory, practice and future trends* (pp. 260–274). Dubai, UAE: TESOL Arabia.

Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis*. New York: Appleton-Century Crofts.

Skinner, B. F. (1989). *Recent issues in the analysis of behavior*. Columbus, OH: Merrill.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychiatry and Psychology, 17*(2), 89–100.

Wright, V. (2018). Vygotsky and a global perspective on scaffolding in learning mathematics. In J. Zajda (Ed.), *Globalisation and education reforms* (pp. 123–135), *Globalisation, comparative education and policy research*, 19. Dordrecht: Springer.

Zhou, M., & Brown, D. (2015). *Educational learning theories* (2nd ed.), *Education open textbooks*, 1. https://oer.galileo.usg.edu/education-textbooks/1. Accessed 30 Dec 2019.

Chapter 28
A Mixed-Methods Approach to Study the Effects of Rater Training on the Scoring Validity of Local University High-Stakes Writing Tests in Spain

Julia Zabala-Delgado

Abstract Spanish universities within the European Higher Education Area have developed a language accreditation framework. The nature of these examinations and the context of higher education place rater training at center stage, but institutions must be aware of this need for scoring validity to re-allocate already stretched resources. This chapter uses a mixed-methods approach to ascertain the longitudinal effect of training of raters at a Spanish university on their recollection of their rating process, the reliability of their scores, and their use of the scale. The study was carried out in three stages over a period of six months with an experimental group and a control group, who took part in three scoring sessions, and two training modules. The data obtained consisted of 150 scripts rated and 30 recorded and transcribed semi-structured interviews of rating sessions. The data were analyzed quantitatively and qualitatively to discern changes of patterns in the rating processes. The results give some insights into the effects of training on scoring validity and reliability. Furthermore, they hint at the possibility of identifying "expert raters" for each context, which could allow small institutions to turn their drawback of limited resources into the advantage of a controlled testing environment.

28.1 Introduction: Purpose and Testing Context

The Bologna process and the creation of the European Higher Education Area (Allegre et al. 1998; Bologna Declaration 1999; European Commission 2018) have changed Higher Education in Europe and have given internationalization, transferability and multilingualism a prominent role in the academic and professional development of university graduates. European universities face the challenge of implementing systems to accredit language knowledge and of doing so in a way that is transferable not only across institutions but also across national borders. As of

J. Zabala-Delgado (✉)
Universitat Politècnica de València, Valencia, Spain
e-mail: juzadel@upv.es

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_28

2007, the Spanish government and its regional delegations required their university graduates to certify their knowledge of a European language. Universities faced the difficult task of stretching their resources to implement accreditation systems that were not only valid and reliable but also susceptible to be implemented attending to practicality concerns and accountable for their effects on the path of higher education in Spain.

28.2 Testing Problem Encountered

In Spain, the Spanish Association of Language Centers in Higher Education (ACLES), comprises 61 universities and belongs to the European Confederation of Language Centres in Higher Education (CERCLES) with 290 universities in Europe. With the goal of implementing an internationally transferable language accreditation system, ACLES developed the CertAcles exam model (ACLES 2011). One of the main problems encountered by the association was to ensure international transferability, since as Alderson and Banerjee (2001) pointed out, there is a great deal of research on large-scale international tests but less about localized tests in which high stakes might be involved. Some studies have been published, among which the report published by the European Commission (2015) on the comparability of language testing in Europe is an example, but there is still a need for detailed studies on the processes followed by examinations that, although local, are still high stakes and have a great impact on national educational policies.

This research focuses on the scoring validity of CertAcles writing tests, from the point of view of rating quality as mentioned by Harsch and Martin (2013), which includes not only the reliability of the raters but also the validity of their decisions. Rater training is an important factor in ensuring rating quality, but it can often be disregarded or viewed as excessively costly, time-consuming, and ineffective. What this chapter explores is how training influences the rating process for raters and how their judgements and use of the writing scale are affected, with the goal of offering a solution for the optimization of training processes to ensure rating quality. With this purpose in mind, our research questions are as follows:

Research question 1: How do CertAcles Rater Training Modules for Writing Tasks affect the raters' recollection of the rating process?

Research question 2: To what degree do rater training modules for the CertAcles writing paper affect inter-rater reliability, rater severity, and consistency?

Research question 3: To what degree do rater training modules for the CertAcles writing paper affect how raters interpret and apply the CertAcles rating scale?

28.3 Review of Literature

The reliability of rating written tests has been one of the main concerns in the field of writing assessment dating back to Edgeworth (1890) and his study of competitive examinations. The number of studies on scoring procedures, however, has risen predominantly in the last 20 years (Alderson and Banerjee 2002); ranging from the design and validation of scoring schemes, to the rating process, the characteristics of the raters or the training required to ensure that test scores provide a basis for their suitable interpretation and use. Since the validity of a test cannot be understood without reliability, a writing test not only needs to fulfill its purpose by measuring what it is supposed to be measuring but it must also be consistent in doing so (Alderson et al. 1995; Hughes 1989).

Marker reliability is an aspect of scoring validity that is particularly relevant in the case of writing assessment, and the use of rater training as a means to achieve rating quality has been widely studied in the literature. Though there is consensus that inter-rater reliability is ultimately important, there are cases in which differences of opinions between raters might be legitimate (Alderson et al. 1995), as rater reliability is a broad concept that includes not only the numbers in the scores but also the reason behind those numbers (Kroll 1998). Clearly, a satisfactory amount of agreement is a key factor for considering that scoring validity is achieved (Weir 2005) but it is not enough to ensure it. In fact, forcing raters to agree on a score to force inter-rater reliability has come under attack on the grounds that it reduces validity, and most studies on foreign language writing tests focus on balancing validity and reliability in scores (Barkaoui 2007). In order to achieve this balance, the focus shifts from eliminating rater variability to studying it and understanding how it affects the validity of the inferences made (Deville and Chalhoub-Deville 2006). The goal is to reduce only the variability produced by unforeseeable random error (Lumley and McNamara 1995). This change of focus has encouraged the emergence of research that is not quantitative-centered but focuses on understanding rater decisions by using a mixed qualitative approach (Elder et al. 2007; Knoch 2009; Weigle 1999; Wiseman 2012; Yan 2014), or even a qualitative-centered approach (Barkaoui 2010; DeRemer 1998; Ellis et al. 2002; Sydorenko et al. 2015).

There are many studies in the literature on marking processes and rating frameworks (Crisp 2010, 2012; Cumming et al. 2002; Lumley 2002; Milanovic et al. 1996). However, the main problem, as identified by Lumley (2002), is the need for raters to reconcile their impressions of the text with its features and the rating scale provided. The rating scale cannot account for all possible situations encountered when rating a script, and during this reconciliation process, raters need to resolve these situations using different strategies that will differ depending on their personal characteristics and bias. The rater constructs an image of the text in their mind that is slightly different to that of the other raters as it is influenced by the rater's working memory and experiences. There are many rater characteristics that have been studied as having an impact on rater behavior, such as 1st language (e.g., Elder and Davies

1998; Huang 2013; Kachchaf et al. 2012), experience with rating (e.g., Barkaoui 2010) and pre-existing training (e.g., Elder et al. 2007; Shohamy et al. 1992).

The reliability of writing performance assessment has improved as a result of a combination of training, better specifications of scoring criteria and better tasks (Lumley 2002), but training is still paramount when dealing with rating quality. Regardless of the criticisms, the advantages of rater training outweigh its disadvantages, as training helps raters understand the rating criteria while adjusting their expectations and making them aware of the need for agreement. However, for training to be effective, there is a need to understand the rater as the center of the process. In spite of the use of mixed-methods to understand the reasons behind the raters' decisions, many findings are limited, inferential, inconclusive or contradictory, which is still a general concern in the literature (Brown 2012). Furthermore, there are questions concerning the long-term effect of training, which have motivated longitudinal studies, such as Lumley and McNamara's (1995), Congdon and McQueen's (2000), Knoch's (2011), or Lim's (2011). The complexity of the process and the implications as regards the validity of rater decisions, imply that for institutions to ensure rating quality, they need to implement training and understand its effects on rater decisions, while taking into consideration the effect of time, which makes quality control more consuming as regards time and resources. Only by understanding rating processes and the effects of training on rating quality, can institutions reconcile their practicality requirements with the need for valid, reliable, and consistent scores in their writing tests.

28.4 Methodology

28.4.1 Design

This study uses a mixed-methods approach with two groups of raters (experimental and control) and a design involving three steps carried out over six months and with three scoring sessions divided into two training modules for the experimental group. A longitudinal design allowed for the observation of the effects of the training and also the duration and variation of such effects over time. A description of the steps follows:

Step 1 Both groups of raters were provided with the first set of five scripts to rate, a marking sheet for each script and a CertAcles B2 rating scale. Individual interviews took place immediately after rating, for reasons of practicality as well as to obtain a more recent recall of the process. An interview was conducted for each individual performance rated, that is, allowing the participants to express themselves and asking for clarification when needed. All materials used during the scoring session were used as stimuli for recall. All interviews were recorded for posterior transcription.

Step 2 Raters in the experimental group followed a training session for CertAcles raters. The week after the training, raters in both groups were again called individually and given a second set of scripts, repeating the marking process and interviews, which were again recorded for transcription.

Step 3 Four months after the second rating session, raters in the experimental group underwent a second training session and again in the following week, the process was repeated and the interviews recorded and transcribed.

28.4.2 Participants

Participants in both groups were university lecturers working for the Universitat Politècnica de València (UPV), a Spanish university administering and delivering CertAcles exams. They had more than 10 years' experience in teaching English as a second language, with the exception of one of the raters whose experience was of 2 years. All raters had MAs in language-related fields with the exception of one who had an MA in engineering and a Certificate of TEFL. Raters in the experimental group were 5 females and 1 male, 3 within the 36–45 age range and 3 within the 46–55 age range. None of them had experience in rating standardized written exams. Their nationalities were Spanish (2), German (1), Irish (1), Canadian (1), and Russian (1). Participants in the control group were within the 36–45 (2) and 46–55 (2) age range and all female. Their nationalities were Spanish (3) and American (1).

28.4.3 Instruments and Procedures

28.4.3.1 Questionnaire

The questionnaire included standard bio-data on participants (gender, age, languages taught) and characteristics, such as first language, years of experience and training received.

28.4.3.2 Semi-Structured Interview

A semi-structured interview schedule was chosen as a compromise between practicality and flexibility. The interview focused on the raters' reading behavior when marking and both the features that influenced their marks and their use of the scale (focus and understanding). It was piloted with an experienced CertAcles rater to avoid ambiguity, double-barreled questions and leading questions.

28.4.3.3 Sets of Candidate's Writing Scripts

Three sets of scripts to be marked by both groups at each of the steps were selected for the study. Each of the sets included five scripts belonging to a B2 CertAcles administration in July 2015 and had been previously marked by four experienced CertAcles markers, whose combined ratings rendered a benchmarked score for each script.

28.4.3.4 Rater Training Modules

The training followed by CertAcles raters was designed based on the guidelines given by Weigle (2002). Due to practicality constraints, the duration of the on-site training was limited to four hours and individual rating sessions were carried out in an asynchronous manner by the raters. The following steps were followed:

- Raters received a "gold standard" script of the level and five scripts exemplifying the different points of the scale, and they were asked to use the scale to order them;
- After a CEFR familiarization exercise, they came together to share their ordering and the motivations for their decisions by using the scale and providing examples;
- A set of borderline scripts were distributed and rated individually. A group discussion ensued to compare individual marks to the benchmark and comments provided by the leading team of raters.
- Questions and answers were encouraged to help raters understand the benchmark but agreement was not forced. Raters who deviated from the score by more than one point out of five were recalled individually to discuss their reasons.

28.4.3.5 CertAcles B2 Rating Scale

The UPV CertAcles B2 rating scale was developed by the UPV for the CertAcles writing paper by combining an empirical development procedure based on scripts representative of CertAcles candidature and an a priori theoretically driven development by using the Common European Framework of Reference. The scale has four main criteria: Task Achievement, Coherence and Cohesion, Grammatical Range and Accuracy, and Lexical Range and Accuracy. It also includes a section for an impressionistic score in an Overall Writing Performance criterion. The rating scale is used together with a marking sheet where raters introduce the scores. A brief summary of each criterion is included in the marking sheet for quick reference.

Table 28.1 Length of interviews and transcriptions

Experimental group						
	Step 1		Step 2		Step 3	
R1	13’56”	1271 words	16’22”	1366 words	12’20”	1217 words
R2	14’58”	1209 words	17’02”	1424 words	12’53”	1137 words
R3	18’55”	1646 words	16’57”	1377 words	14’06”	1107 words
R4	23’06”	2227 words	18’26”	1670 words	17’48”	1793 words
R5	13’10”	1456 words	11’38”	1154 words	9’42”	1109 words
R6	12’	1416 words	13’49”	1496 words	8’40”	969 words
Control group						
R1C	16’ 10”	1553 words	13’ 54”	1521 words	11’30”	1151 words
R2C	18’18”	1499 words	18’ 20	1908 words	13’43”	1542 words
R3C	13’39”	1979 words	12’25”	1536 words	12’50”	1883 words
R4C	17’	1239 words	13’33”	1559 words	9’56”	1107 words

28.4.4 Data Collection

After completion of the three steps, the data obtained consisted of individual markings of the 10 raters for one set of scripts at step 1, one set at step 2 and one set at step 3, and a total of 30 recorded interviews, one for each rating session of the 10 raters at each of the steps. The order of the scripts was altered for each rater to avoid order effect. Raters were identified as R1, R2, etc., for raters in the experimental group, and R1C, R2C, etc., for raters in the control group to provide anonymity. An orthographic transcription of the interviews was carried out. The length of the interviews and transcriptions can be seen in Table 28.1.

28.4.5 Data Analysis

Qualitative analysis was used to answer RQ1 and examine raters’ recollections of the process, as well as to answer RQ3 as regards changes in the use of the scale. Quantitative analysis was used to answer RQ2 in terms of changes in inter-rater reliability, rater severity, and consistency, as well as to answer RQ3 in terms of changes in the ranges of scores in relation to the benchmark.

28.4.5.1 Qualitative

Nvivo was used for qualitative analysis to transcribe semi-structured interviews and code them. The approach followed for coding was mixed, with a deductive, top-down approach based on the literature and on the initial goals of the research and an

Table 28.2 Coding scheme for qualitative analysis

Top down	Bottom up
Process to reach a mark Number of readings Steps followed Holistic marking Analytic marking	**Strategies to reach a mark** Self-monitoring Qualitative analysis of problem *(evaluating both strengths and weaknesses)* Comparison between candidates for grading Arbitrating between their impression and the scale
Use of marking criteria Post-judgment During judgment Pre-judgment	**Use of the scale** Focus on external factors Lack of use Use -Mention non-specific -Mention with internalized quotes Indication that the scale is misunderstood or misused Mention of scale shortcomings
Focus on criteria Task achievement Coherence and cohesion Grammar range and accuracy Lexical range and accuracy Overall written production	**Feeling toward their process** Certainty about their process Uncertainty about their process

inductive, bottom-up approach that allowed for new themes to appear. The coding scheme was intended to help answer the research questions by showing the sequence of rating as well as the raters' recollections of the scoring categories applied to each of the scripts, together with the challenges faced during the process. The scheme can be seen in Table 28.2.

28.4.5.2 Quantitative

Descriptive statistics were calculated for all the scores and Cronbach's alpha was obtained to examine inter-rater reliability. Furthermore, shared variance between the groups and the benchmark scores were calculated by running a Spearman correlation between their means. Since the study focused on a longitudinal analysis, a mixed between-within subjects analysis of variance (ANOVA) was carried out to analyze the statistical significance of the impact of the training.

28.5 Findings

28.5.1 How Do CertAcles Rater Training Modules for Writing Tasks Affect the Raters' Recollection of the Rating Process?

The interviews carried out with the raters in both groups included a question about the process they followed. An analysis of the answer to the question is represented in Tables 28.3, 28.4, 28.5, 28.6, 28.7, 28.8, 28.9, 28.10, 28.11, and 28.12 for both groups, organized per step and rater.

28.5.1.1 Rating Processes of Experimental Group

According to their own recollections of the process, raters in the experimental group did alter their rating processes after receiving training. This is particularly evident between steps 1 and 2 when they had completed the first training module, with changes in the process maintained and still occurring in step 3. The most relevant changes according to their own accounts of the process were:

Table 28.3 Interpretation of R1's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Scale familiarization • Read scripts • Re-read each script, give overall mark, check scale, and give marks per criterion • Compare candidates and adjust mental scheme to the scale	• Scale familiarization • Read scripts and give overall mark • Re-read the scale • Re-read each script, mark each criterion looking for examples of the features in the scale descriptors	• Read all scripts • Re-read each script, mark each criterion while checking descriptors in the scale • Compare scripts and give overall mark

Table 28.4 Interpretation of R2's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Read scripts, annotate negative and positive features • Re-read scripts iteratively, occasionally checking the scale when in doubt	• Read scripts for general impression and give overall mark • Re-read scripts, mark per criterion while checking descriptors in the scale • Adjust overall mark	• Read scripts for general impression and give overall mark • Re-read scripts, mark per criterion while checking descriptors in the scale and identifying examples of the descriptors on the scripts • Adjust overall mark by comparing all candidates

Table 28.5 Interpretation of R3's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Scale familiarization • Read scripts, identify mistakes, give overall mark • Check descriptors in case of doubt • Re-read scripts and mark per criterion, looking for examples of descriptors • Re-read, skimming to check marks	• Read scripts, mark mistakes, give general mark • Check descriptors one by one and give a mark per criterion, looking at the annotations on the script • Re-read (skimming), search for specific examples to justify marks	• Read scripts, mark mistakes (add annotations from the scale) • Give a mark per criterion, checking scale and looking at the annotations on the script • Re-read (skimming), search for specific examples to justify marks • Give overall mark

Table 28.6 Interpretation of R4's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Read script for overall idea • Re-read, annotate features and mark each criterion • Re-read weak scripts • Check the scale to confirm marks	• Read script for overall idea • Check descriptors • Re-read iteratively and mark per criterion • Check the scale to confirm marks and search for examples of descriptors on the script	• Read script for overall idea • Check descriptors • Re-read iteratively and mark per criterion • Check scale to confirm marks and search for examples of descriptors on the script

Table 28.7 Interpretation of R5's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Scale familiarization • Read to give overall mark • Re-read to mark per criterion • Re-read to check the marks given	• Read to get overall idea • Check the scale • Re-read to mark per criterion • Give overall mark • Re-read to check the marks given	• Read to get a general idea • Read to mark per criterion while checking the scale • Give overall mark

Table 28.8 Interpretation of R6's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Scale familiarization • Read to give general mark and make annotations • Re-read, mark per criterion and add annotations	• Scale familiarization • Read to get a general idea • Read, marking per criterion and making annotations while checking scale • Give overall mark	• Read to get a general idea • Read, marking per criterion and making annotations while checking scale • Give overall mark • Re-check descriptors

Table 28.9 Interpretation of R1C's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Read scripts, annotate positive and negative features • Re-read scripts for an overall mark, comparing candidates (on a scale of 1–10) • Mark per criterion, check descriptors, re-read specific sections to confirm marks	• Read all the scripts, compare scripts and give overall mark while underlining mistakes (on a scale of 1–10) • Mark per criterion, check descriptors, re-read specific sections to confirm marks	• Read scripts, annotate positive and negative features • Re-read scripts for a general impression, give overall mark by comparing performances (on a scale of 1–10) • Mark per criterion without checking descriptors, only scale of reference (1–5)

Table 28.10 Interpretation of R2C's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Scale familiarization • Read script iteratively • Mark per criterion and check the scale for reassurance • Give overall mark	• Scale familiarization • Read script iteratively • Mark per criterion and check the scale for reassurance • 2nd and 3rd readings for weak productions • Give overall mark	• Scale familiarization • Read script iteratively • Mark per criterion. Check scale with weak scripts • 2nd and 3rd readings for weak scripts • Give overall mark

Table 28.11 Interpretation of R3C's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Read scripts, annotate positive and negative features, decide if pass or fail • Read scripts, give marks per criterion and write feedback for candidate • Read scripts, compare candidates and adjust marks accordingly • Quick check of descriptors	• Read scripts, get a general idea • Read scripts, give marks per criterion and write feedback • Read scripts, compare candidates and adjust marks accordingly • Check scale and adjust marks accordingly	• Read scripts, get a general idea • Read scripts, give marks per criterion and write feedback • Read scripts, compare candidates and adjust marks accordingly • Scale only checked for weak productions

Table 28.12 Interpretation of R4C's account of the rating process followed at steps 1, 2, and 3

Step 1	Step 2	Step 3
• Scale familiarization • Read scripts, overall grade, annotations on the text and mark per criterion • Check the scale for weak productions in search of reassurance	• Read script, overall grade, annotations on the text and mark per criterion • Check scale of reference and description of criteria on marking sheet	• Read script, overall grade, annotations on the text and mark grammar and spelling • Read script, mark content and structure • Scale only checked for weak productions

The change in the order for giving the overall mark, which is altered for R1, R2, R3, R5, and R6, from giving the overall mark before the individual marks to giving it at the end of the process. The focus goes from a holistic overall mark to systematically giving marks per criterion.

The scale goes from being checked before the process to being checked during the marking process. There is a change from grading based on intuition and previous experiences and then a quick check of the scale, to grading criterion after criterion while checking the scale.

There is indication of a *heightened awareness of the process* since the training seems to improve the match between the process they describe as being followed and the process actually followed according to the analysis of the interviews.

There is *a greater uncertainty about their process and their decisions*, which takes raters to check the scale more often for reassurance.

28.5.1.2 Rating Processes of Control Group

Contrariwise, raters in the control group do not alter their process very much in the three steps. There are minor changes and mostly related to what seems heightened confidence in their process, with less checking of the scale, and fewer reservations in admitting to the researcher that they do not check the scale.

28.5.2 To What Degree Do Rater Training Modules for the CertAcles Writing Paper Affect Inter-Rater Reliability, Rater Severity, and Consistency?

Cronbach's alpha was calculated for the combined set of scores awarded by the raters in both groups at each of the steps, as well as for the scores achieved for the individual criteria. Results can be seen in Tables 28.13 and 28.14.

Raters in both groups had a high consistency level, between .76 and .96, the majority of them $\alpha > .8$, which are good reliability values for a high-stakes test. There seems to be a drop in the Cronbach's alpha of TA and CC in the experimental group that does not occur in the control group, but besides that, there seems to be no clear pattern differentiating the behaviors of the experimental and control groups across the three steps. Nevertheless, and to be sure no patterns could be found, a sub-analysis was carried out comparing the means of both groups at each step and for each of the criteria. A graphical representation of the evolution of the means per rater per criterion and in both groups showed that the behavior and means of both groups are similar at the three steps, with a negligible difference in the case of task achievement, since both groups move in the range of 3.

Spearman correlations were carried out between the means of the two groups and the benchmark scores at the three steps. The results can be consulted in Table 28.15.

Table 28.13 Cronbach's alpha for experimental group

Experimental group ($N = 6$)		
Step 1	Step 2	Step 3
Combined scores		
$\alpha = .92$	$\alpha = .89$	$\alpha = .93$
Task achievement (TA)		
$\alpha = .89$	$\alpha = .86$	$\alpha = .76$
Coherence and cohesion (CC)		
$\alpha = 90$	$\alpha = .86$	$\alpha = .85$
Grammatical range and accuracy (GRA)		
$\alpha = .94$	$\alpha = .77$	$\alpha = .93$
Lexical range and accuracy (LRA)		
$\alpha = .89$	$\alpha = .84$	$\alpha = .94$
Overall written production (OWP)		
$\alpha = .92$	$\alpha = .90$	$\alpha = .94$

Table 28.14 Cronbach's alpha for control group

Control group ($N = 4$)		
Step 1	Step 2	Step 3
Combined scores		
$\alpha = .94$	$\alpha = .90$	$\alpha = .94$
Task achievement (TA)		
$\alpha = .80$	$\alpha = .84$	$\alpha = .95$
Coherence and cohesion (CC)		
$\alpha = .93$	$\alpha = .92$	$\alpha = .94$
Grammatical range and accuracy (GRA)		
$\alpha = .94$	$\alpha = .78$	$\alpha = .93$
Lexical range and accuracy (LRA)		
$\alpha = .96$	$\alpha = .93$	$\alpha = .95$
Overall written production (OWP)		
$\alpha = .89$	$\alpha = .92$	$\alpha = .94$

Table 28.15 Spearman correlations between the means of the two groups and the benchmark scores at the three steps

	Step 1	Step 2	Step 3
Experimental group and the benchmark	$r_s = .70$, $n = 5$, $p < .001$	$r_s = .90$, $n = 5$, $p < .001$	$r_s = .90$, $n = 5$, $p < .001$
Control group and the benchmark	$r_s = .71$, $n = 5$, $p < .001$	$r_s = .70$, $n = 5$, $p < .001$	$r_s = .80$, $n = 5$, $p < .001$

The Spearman correlations carried out at each of the steps between the benchmark and the means of the control and experimental groups show strong correlations in all of the cases, although the strength of the correlation increases in the experimental group in step 2 and is maintained in step 3. In the case of the control group, the strength remains the same in step 2 and increases, although to a lesser extent, in step 3. Figure 28.1 represents the means of the three groups at steps 1, 2, and 3.

As for the severity of the raters, the evolution of their means across the steps was represented graphically and a visual inspection of the graphs agreed with the previous analysis in that the performance of the raters does not seem to be altered by the training modules.

To further analyze the statistical significance of the impact of the training on the experimental group, a mixed between-within subjects analysis of variance was conducted to look into the total scores of the control and experimental groups across the three time periods, steps 1, 2, and 3. No significant interaction was found between group and time, Wilks' Lambda = .79, F (2.7) = .913, p = .44, partial eta squared = .20. There was no substantial main effect for time, Wilks' Lambda = .52, F (2.7) = 3.25, p = .10, partial eta squared = .48. The main effect comparing the two

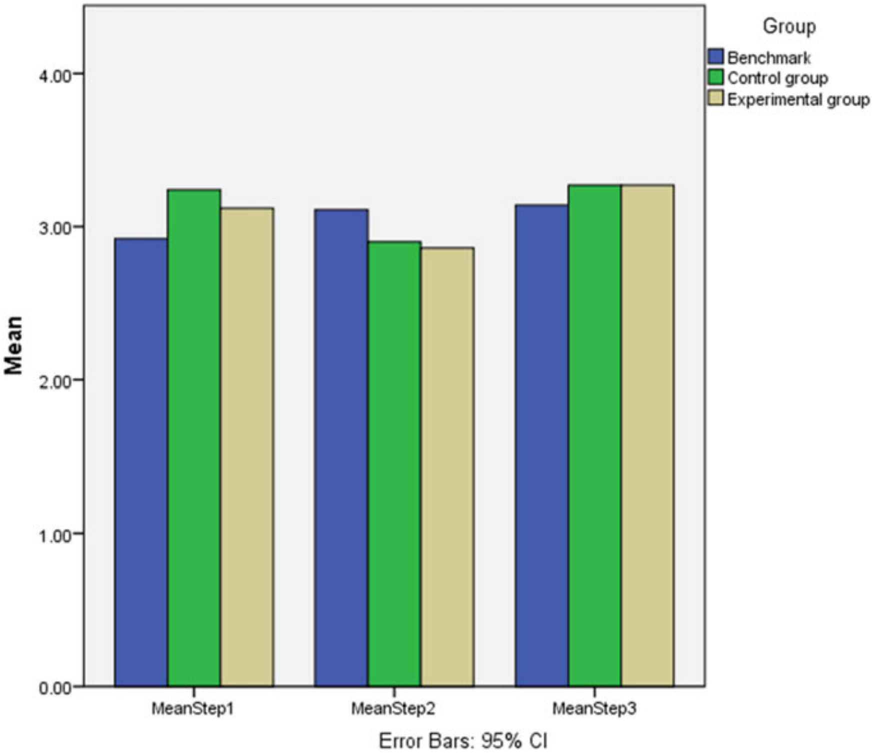


Fig. 28.1 Comparison of means of experimental and control groups with benchmark scores at each of the steps

interventions was not significant, F (1.8) = .084, p = .78, partial eta squared = .01. These calculations thus confirm the above findings in indicating that the training did not have an effect on the scores given by the raters in the experimental group.

28.5.3 Research Question 3. To What Degree Do Rater Training Modules for the CertAcles Writing Paper Affect How Raters Apply the CertAcles Rating Scale?

28.5.3.1 Range of Scores

To analyze the way the scale was applied at the three steps, an analysis of the ranges of scores used by the raters in both groups was carried out and the results compared to the range of the benchmarked scores. The result of the analysis is illustrated in Table 28.16. R5 is the rater with most limited range at the three steps and R4 has the widest range. The most relevant result when looking at the performance of the raters as a group is a restriction of range in step 2 for both the experimental (Mean range = 2.7) and the control (Mean range = 1.8) groups that are not present in the benchmark score (Mean range = 3.4), but the restriction is less pronounced in the experimental group where R3 and R6 actually increase their ranges and get close to that of the benchmark score. As a result, the combined scores of the experimental group are closer to the benchmark range with each step.

28.5.3.2 Focus on Criteria

As for use of the scale in rating, it was already mentioned in the results to RQ1 that the scale was checked more often. However, there is also a change in the attention paid to the different criteria. In the experimental group, previously underrepresented criteria are mentioned more often after the training. This increased balance in the consideration of criteria can be observed by looking at Figs. 28.2, 28.3, and 28.4 illustrating the times each of the criterion is mentioned by the raters when asked about the ratings at each step (Figs. 28.5, 28.6, and 28.7).

Upon visual inspection, raters in the control group report similar behavior as regards their focus on the criteria in the scale for the three steps. However, raters in the experimental group report variations and describe a more balanced approach.

28.5.3.3 Quoting the Scale

Another example of the variation in the use of the scale is represented by the raters quoting the scale when marking in steps 2 and 3. R4 is a particularly good example, since she goes from using the wrong scale range (1–10) at step 1, to quoting the scale

Table 28.16 Range of scores of experimental group (EG) and control group (CG) at steps 1, 2, and 3

	R1	R2	R3	R4	R5	R6	R1C	R2C	R3C	R4C	Mean range EG	Mean range CG	Benchmark
Range of scores at step 1	3	3.5	2.4	4.2	1.7	1.7	2.4	2.5	2.2	1.7	2.4	2.2	3.6
Range of scores at step 2	3	2.8	3.2	3.5	.8	3	1.5	1.1	2.2	2.4	2.7	1.8	3.4
Range of scores at step 3	2.9	3.4	2.4	4.2	1.2	2.6	3.3	3.3	3.2	3.4	2.7	3.3	2.3

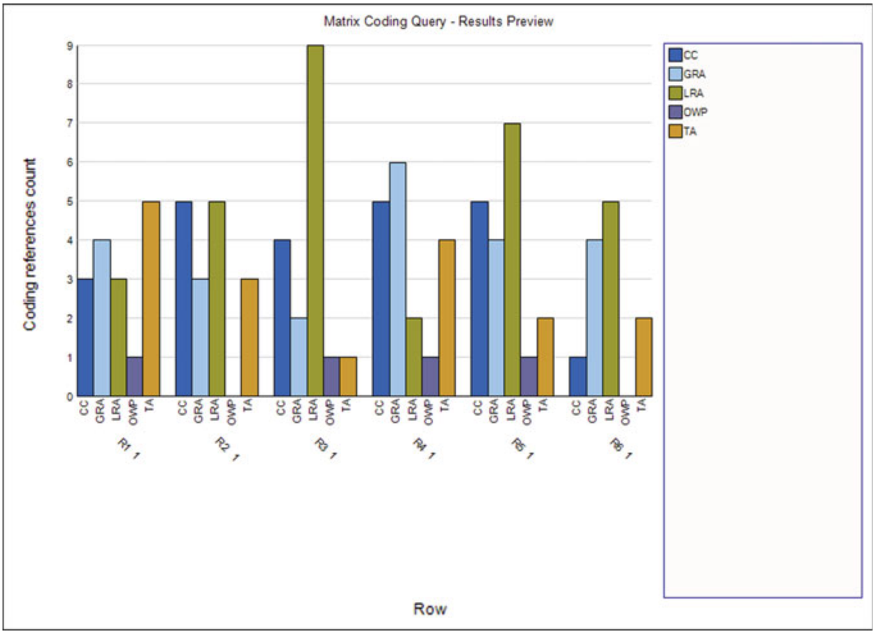


Fig. 28.2 Use of criteria for experimental group at step 1

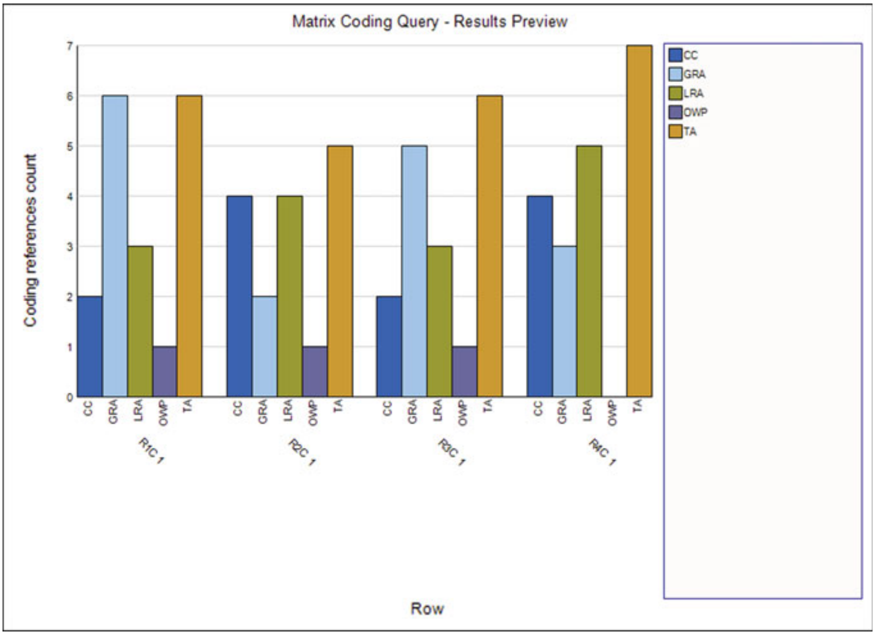


Fig. 28.3 Use of criteria for control group at step 1

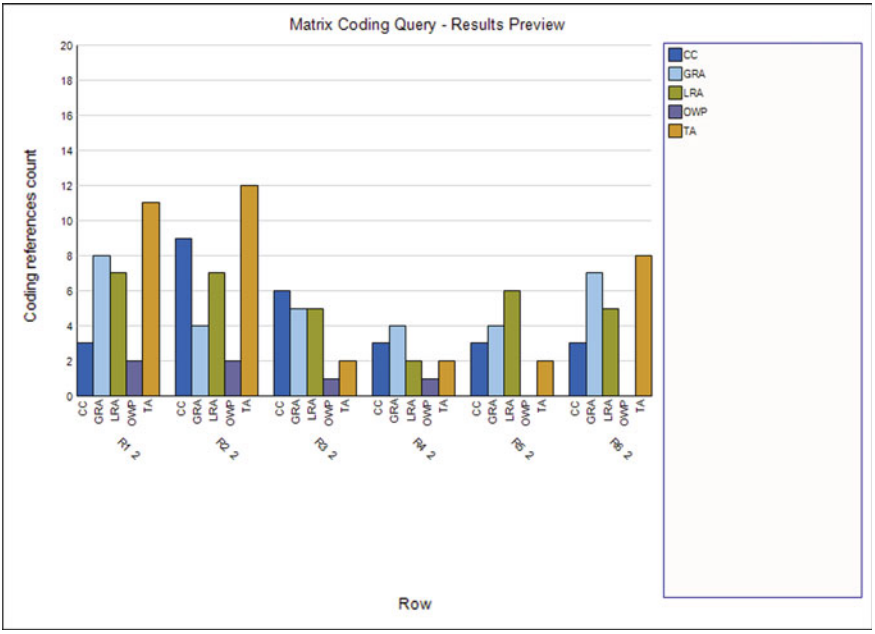


Fig. 28.4 Use of criteria for experimental group at step 2

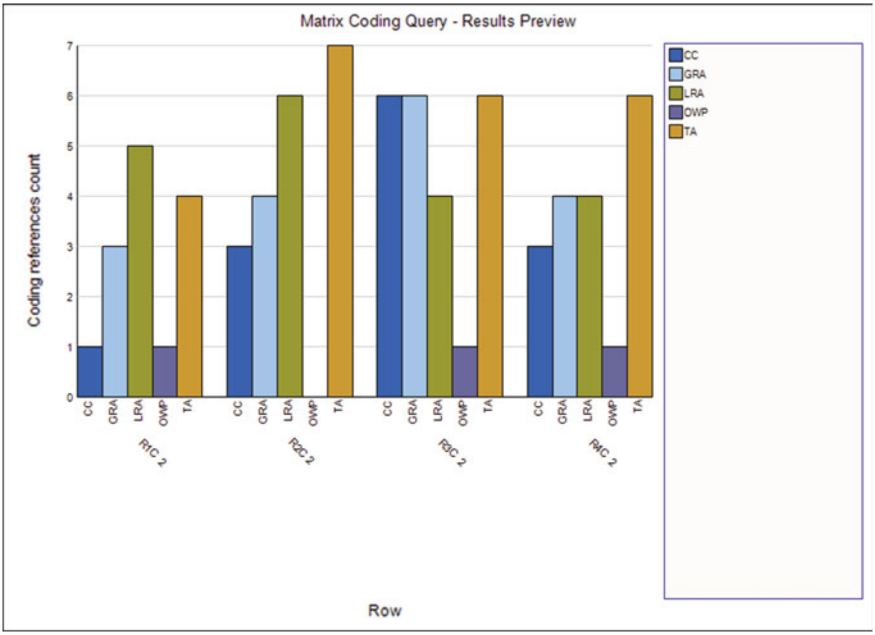


Fig. 28.5 Use of criteria for control group at step 2

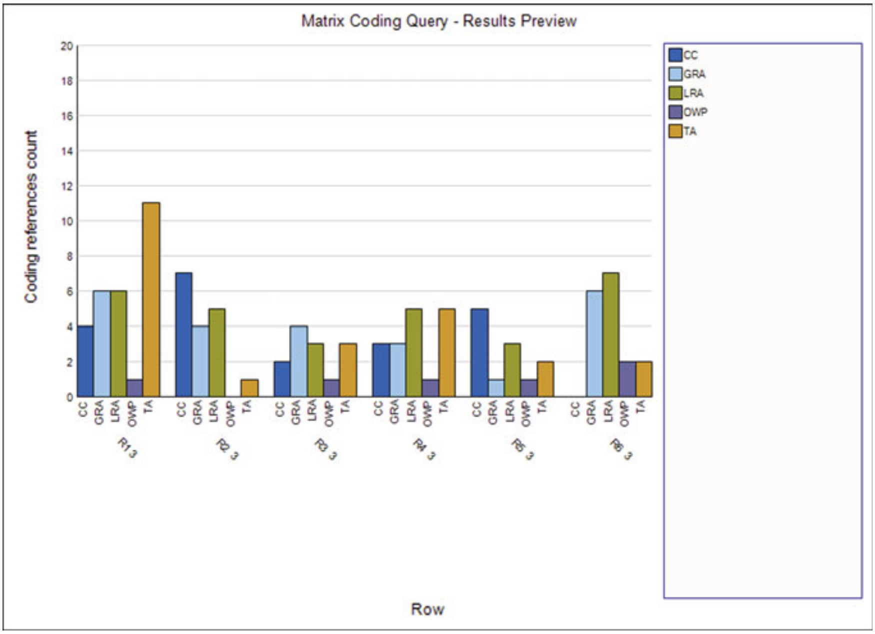


Fig. 28.6 Use of criteria for experimental group at step 3

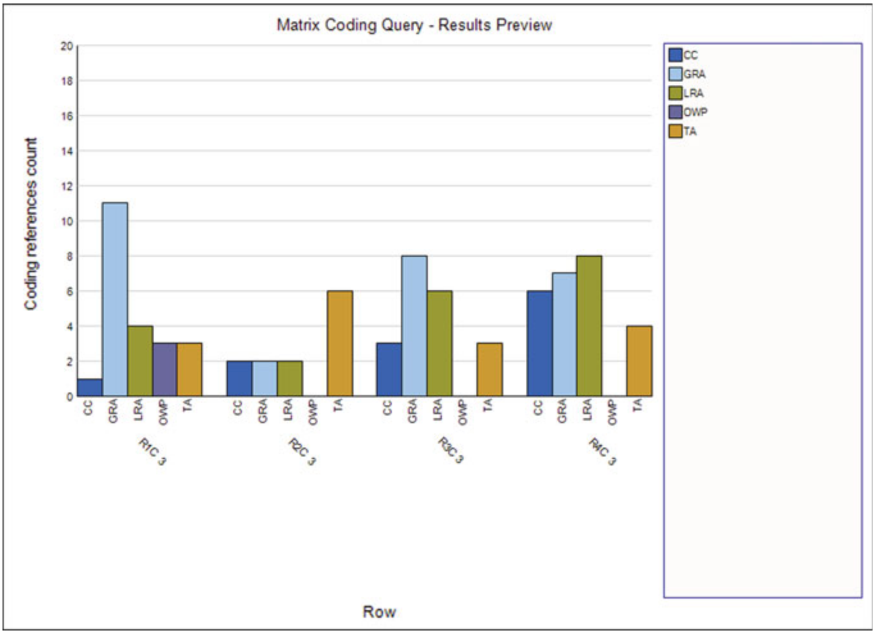


Fig. 28.7 Use of criteria for control group at step 3

and giving examples exemplifying the quote. Conversely, in the control group, the scale is checked less often, which could be understood as a sign of confidence in their marking process or in the interview process as well, with less need to justify their decisions. In fact, although all raters mention checking the descriptors in the scale during the rating session at step 1 and step 2, in step 3, R1C talks about not checking the scale at all, while raters R2C, R3C, and R4C mention checking the scale only in the case of weak productions.

28.6 Insights Gained

The insights gained will be discussed by analyzing the overlapping implications of the results obtained from the reliability standpoint and the validity standpoint. This is in line with the study focusing on rater quality, understood as a combination of the reliability of the ratings and the validity of the decisions.

28.6.1 Insights Gained from a Reliability Standpoint

From the reliability standpoint, these results, although inferential, seem to agree with Lumley and McNamara (1995) and Weigle (1994) in that training does not seem to eliminate variance in raters, but does seem to have an effect on raters' consistency, which will be analyzed from the validity point of view further on. As mentioned before, raters that deviated from the scores by more than one point received individual feedback, which, according to the results of this study, were not effective in eliminating such deviations. This corroborates what was reported by Knoch (2011) in a study of the effectiveness of individualized feedback to raters, who seemed to be less prone to incorporating suggestions when highly experienced. It also agrees with research conducted by Cumming et al. (2002) on the difficulty found by raters to unlearn set rating practices and on their rating being affected by their previous experiences.

The fact that the reliability of all raters seemed to be very good before any intervention was carried out further corroborates the above and agrees with Weigle (2002) and Winke and Gass (2013) on the effect of raters' background. Raters from similar backgrounds and with a similar knowledge of the L1 are affected by similar factors and thus perform similarly. The use of a detailed rating scale can also increase the reliability of ratings, which together with the experience of the raters in the study and their level of studies, would confirm Barrett's (2001) consideration that raters that are highly knowledgeable in the domain they are rating can be expected to achieve a high degree of consistency. This is further corroborated by the fact that R5, one of the raters who presented more issues from the reliability point of view, was the rater with least training in the field of linguistics and least experience. Likewise, the results of this study also corroborate Lim (2011), who found similar results in observing that raters maintained their quality of rating over time, which advocated for the existence of a category of expert raters.

28.6.2 Insights Gained from a Validity Standpoint

Rater training from a validity standpoint needs to help raters understand and apply the scoring criteria in a way that reflects what the test is intended to measure. From the qualitative analysis of the interviews, there is an indication that the training does affect the recollection of the rating process of raters and the steps show they applied less intuition and more structure: formulating an idea, confirming it with the descriptors and revising their decisions. This can be seen in the raters' checking of the descriptors and their entering an adjustment stage in an organized manner after the training. There is evidence that the findings corroborate that raters need to reconcile their intuitive impression with the rating scale and the specific features of the task. Nevertheless, and in spite of training producing more uncertainty, this is not necessarily counterproductive but helps them in elaborating the process and seems to lead them to consult the scale more often and more systematically. Obtaining an insight into these conciliation processes can help guide future training as well as report the results to raters and improve the feedback they receive as indicated by Cumming et al. (2002). The attitudes of the raters to the feedback on their processes seems to be positive, in agreement with the results obtained by Elder et al. (2007), and seems to generate a better awareness of the whole process.

As mentioned above, although their similar backgrounds and experiences, together with their knowledge of the context, create a higher reliability than expected between raters who underwent training and raters who did not, when problematic situations occur, raters in the control group are more influenced by personal intuition, creating inconsistencies, and potential sources of construct irrelevance. This could ultimately result in unfair results and advocates for the use of training even if expert raters could potentially be pre-selected.

28.7 Conclusion: Implications for Test Users

Before commenting on the conclusions of the study, some limitations should be noted. The main limitation was the quantity of data, which has already been mentioned in the literature review—as it is a recurrent limitation in this type of study. Different results can be obtained when using FACETS analysis, as the results can give a deeper insight into the behavior of raters. Furthermore, semi-structured interviews were used to collect the data on the marking processes, and they were carried out after the rating process, bringing memory into play, albeit the use of stimulated recall. Notwithstanding this, the use of a mixed-methods approach allowed us to gain insights not only from the reliability standpoint, but also from the validity standpoint as shown above. The quantitative approach showed the variation of the ratings along the three steps for both group of raters, but it was the qualitative approach that provided insight into their process of reconciliation between their experience, the use of the scale and the written performances. It was the use of both approaches that allowed us to understand the process of rating beyond the scores and thus to place the rater at the center of the study. Quantitative data indicated that training had little effect on the ratings obtained, and would have led us to believe that raters in both groups behaved as experienced raters and performed similarly at each step, independently from the training received. However, qualitative data showed that the processes were in fact affected to a certain degree and allowed us to understand that a detailed scale together with training generated a more structured process and fewer inconsistencies in interpretations, ensuring a more valid interpretation of the performances.

The main implications of the study concern small-scale programs, such as the one analyzed. The study points toward placing the rater and not the scale at the center of the process. The raters in the study decided what to pay attention to, how to solve conflicts and how to bring together their impression of the text with the institutional requirements, the scale and the training. Therefore, by carefully selecting raters for training programs according to a hypothetical category of expert raters with certain characteristics, training efforts could be optimized and resources used more wisely, with considerable effects for small-scale high-stakes examinations. In fact, the analysis of their profiles and of their recollections of the process, together with a longitudinal analysis of their ratings, led us to infer that in both groups, most raters represent and behave as expert raters. In fact, they showed many of the characteristics defined in Chi et al. (2014): (1) knowledge of their domain, (2) ability to perceive meaningful patterns, (3) ability to perform rapidly in their domain, (4) capacity to perceive a problem at a deeper level, (5) willingness to spend time analyzing a problem qualitatively, (6) strong self-monitoring skills, (7) accuracy at judging problem difficulty, and (8) more reliance on semantic memory than on general reasoning.

Although Lim (2011) stated that few large-scale language tests could be rated by a small number of raters working in a single location with daily interaction, this is exactly the case in examinations, such as CertAcles exams, run by individual universities within a national framework. In this case, the downfall of small-scale high-stakes examination programs could also be their greatest fortune. The limitations inherent to financial and time constraints in small programs also entail the possibility of creating a group of like-minded individuals with a profound knowledge of the context. These individuals could focus on similar features when marking and potentially become expert raters for a particular context and exam based on their characteristics and even before being subjected to rater training. Without diminishing the crucial importance of the training of raters, their similar background, combined with double rating to reduce rater differences, could render acceptable reliability with fewer training resources, as long as this reliability was constantly monitored and the calibration of raters was performed in every administration.

Further research should, however, be carried out to see if the inferences obtained from this study could indeed be confirmed with a larger database and with raters with a different background and no knowledge of the context at hand. Furthermore, the effects of variations in the training modules, such as the use of more golden standards to reduce rater uncertainty, or the increase in the duration of the training as well as the delivery method (attendance-based, online or a combination or both), should also be explored. Nevertheless, it is still encouraging for smaller institutions to observe that the validity and reliability of scores are within reach despite their limitations, and that by shifting perspectives, weaknesses can be used as tools to solve testing problems by taking into consideration the particularities of the contexts.

References

ACLES. (2011). *Model of language accreditation.* http://www.acles.es/multimedia/enlaces/9/files/fichero_29.pdf. Accessed 9 July 2018.

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part I). *Language Teaching, 34*(4), 213–236.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*(2), 79–113.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Allegre, C., Berlinguer, L., Blackstone, T., & Rüttgers, J. (1998). Sorbonne joint declaration: Joint declaration on harmonisation of the architecture of the European higher education system. Accessed 18 July 2019.

Barkaoui, K. (2007). Participants, texts, and processes in second language writing assessment: A narrative review of the literature. *The Canadian Modern Language Review, 64,* 97–132.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31–57.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49–58.

Bologna Declaration. (1999). Retrieved July 18, 2019. https://www.eurashe.eu/library/bologna_1999_bologna-declaration-pdf/.

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 413–425). London: Routledge.

Chi, M. T., Glaser, R., & Farr, M. J. (2014). *The nature of expertise.* London: Psychology Press.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163–178.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36,* 1–21.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice, 31,* 10–20.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7–29.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 9–25). Amsterdam: John Benjamins Publishing.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53,* 644–663.

Elder, C., & Davies, A. (1998). Performance on ESL examinations: Is there a language distance effect? *Language and Education, 12,* 1–17.

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37–64.

Ellis, R., Johnson, K. E., & Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219–233.

European Commission Directorate General for Education and Culture. (2015). *Study on comparability of language testing in Europe.* http://ec.europa.eu/languages/library/documents/edl-report_en.pdf. Accessed 2 Nov 2017.

European Commission/EACEA/Eurydice. (2018). *The European higher education area in 2018: Bologna process implementation report*. Luxembourg: Publications Office of the European Union.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281–307.

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System, 41*(3), 770–785.

Hughes, A. (1989). *Testing for language teachers.* Cambridge, New York: Cambridge University Press.

Kachchaf, R., & Solano-Flores, G. (2012). Rater language background as a source of measurement error in the testing of English language learners. *Applied Measurement in Education, 25*(2), 162–177.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—A longitudinal study. *Language Testing, 28*(2), 179–200.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18,* 219–240.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543–559.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 92–114). Cambridge: Cambridge University Press.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76,* 27–33.

Sydorenko, T., Maynard, C., & Guntly, E. (2015). Rater behaviour when judging language learners' pragmatic appropriateness in extended discourse. *TESL Canada Journal, 32*(1), 19.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197–223.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145–178.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation.* Oxford: Palgrave.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762–789.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150–173.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501–527.

Part V
Learning from Tests, Teachers, and Language Assessment Literacy

The theme of this part relates to learning from tests, teachers, and language assessment literacy, with chapters 29–30 as experience-based papers and chapters 31–36 as data-based.

- Chapter 29, entitled "A Critical Evaluation of the Language Assessment Literacy of Turkish EFL Teachers: Suggestions for Policy Directions," by Ölmezer-Öztürk, Öztürk, and Aydın, explores problems in the language assessment practices in Turkish EFL teaching especially regarding teachers' language assessment literacy. The authors suggest possible solutions including the need for involving teachers, trainers, and policy makers in addressing the assessment literacy challenge and supporting those teachers who already have assessment background knowledge.
- Chapter 30, "Some Practical Consequences of Quality Issues in CEFR Translations: The Case of Arabic," by Norrbom and Zuboy, addresses problems with the quality (in terms of terminology, level descriptors, and style) of the CEFR Arabic translation used by the Council of Europe. The authors discuss the implications and suggest that, in the short term, users exercise care in utilizing any teaching or assessment items based on that Arabic translation, and in the long term an official Council of Europe Arabic CEFR translation be produced along with supporting documentation including a multilingual glossary.
- "Assessment Literacy and Assessment Practices of Teachers of English in a South-Asian Context: Issues and Possible Washback," by De Silva (Chapter 31), is a mixed-methods study which examined issues in assessment literacy among English teachers in Sri Lanka. The author found that some teachers were fairly knowledgeable about the basic principles of assessment, but most had problems applying such principles. The author suggests the need for training and support for setting, administering, and scoring tests.
- In Chapter 32, entitled "English Language Testing Practices at the Secondary Level: A Case Study from Bangladesh," by Rahman and Khan, the authors describe a study which explores issues raised by testing practices in Bangladeshi English language secondary schools and their washback on examinees and the educational system as a whole. They suggest a need for developing assessment literacy for teachers,

test developers, and scorers so that these groups as well as test users can identify and correct harmful practices with consequences in teaching and learning.

- The context of high-stakes testing in the UAE is examined in Chapter 33 "A New Model for Assessing Classroom-Based English Language Proficiency in the UAE." The author, Khemakhem, describes a study which examines problems raised by an IELTS band 6 graduation requirement at the end of a B.Ed. program in the UAE designed to ensure that students have the minimum English language proficiency needed to teach English in schools. The researcher suggests bridging the gap between what is and what should be by proposing a new assessment tool (i.e., the Classroom-Based English Language Proficiency Rubric) that merges IELTS descriptors with the principal features of classroom interaction.
- A team of authors moves us to Malta in Chapter 34 "Assessing Teacher Discourse in a Pre-Service Spoken English Proficiency Test in Malta." Vassallo, Xerri and Jonk address problems in the spoken English proficiency of pre-service teachers and describe the design and use of a spoken proficiency test that included teacher discourse as the first of five criteria and explain how assessing teacher discourse is a suitable way to address the needs of pre-service teachers of English and the effects of doing so on English teaching in Malta.
- Chapter 35 "High-Stakes Test Preparation in Iran: The Interplay of Pedagogy, Test Content, and Context," by Saif, is a study which explores issues related to high-stakes test preparation in Iran. The author finds that the culture of the test center shapes the test preparation courses in terms of test demands, and that instruction goes well beyond test-related activities to include contextual factors like student goals/needs, teacher experience, and second language learning beliefs.
- Authors Yan, Kim, and Kotnarowski (Chapter 36) examine the "Development of a Profile-Based Writing Scale: How Collaboration with Teachers Enhanced Assessment Practice in a Post-Admission ESL Writing Program at a USA University." Their study investigates issues that arose because raters of academic English writing samples in the USA approached the rating of argument development differently, which resulted in conflicting ratings on certain essays. The author showed how several rounds of discussion led to resolving these differences and creating separate criteria for argument development and lexico-grammar, which in turn led to a scale that more accurately reflected the examinees' range of writing performances.

The eight chapters in this part are connected to the theme of learning from tests, teachers and language assessment literacy and with each other in that they deal with challenges that arise in language assessment literacy levels in different contexts around the world, namely Turkey, Sri Lanka, Bangladesh, Malta, Iran, the United Arab Emirates and the USA (Chapters 29, 31–36). Two of the chapters in Part V deal with how aspects within tests have helped stakeholders learn from the tests themselves. These challenges stem from the quality of the CEFR translation to Arabic on the Council of Europe website and the varied and sometimes conflicting interpretations of raters on the argument writing scale in the USA (Chapters 30, 36).

The eight chapters in this part are related to each other in that they are experience-based (chapters 29–30) in the form of narratives and position papers and the remaining six of them are data-based (Chapters 31–36). In addition. all chapters are similar to each other in that they are about high-stakes English language tests, or aspects relating to high-stakes assessment like marking, language assessment literacy policies and practices.

Chapter 29
A Critical Evaluation of the Language Assessment Literacy of Turkish EFL Teachers: Suggestions for Policy Directions

Elçin Ölmezer-Öztürk, Gökhan Öztürk, and Belgin Aydın

Abstract English as a Foreign Language (EFL) teaching policies have long been a primary concern for Turkey with its constantly changing educational systems and policies. Although there have been several attempts in terms of program renewal in primary and secondary levels, there has been no reformation in terms of language testing and assessment policies, which have already been the most problematic aspect of language teaching. In Turkey, language teachers are highly responsible for assessing learners through formal and informal assessment practices. However, though the expectation of them is considerable, their exposure to training in language assessment is quite limited in both pre-service and in-service levels. Considering all these realities, this chapter scrutinizes the language testing and assessment practices in the Turkish EFL context with a special emphasis on English teachers' language assessment literacy. First, a short description of the context in terms of language testing and assessment practices is presented. Next, what language assessment literacy is and how literate language teachers in the world are examined. Language teachers' assessment literacy, how they are trained, and what kind of problems are experienced in Turkey are discussed in detail referring to the elements in teacher education and training processes both in the pre-service and in-service levels. New insights parallel with the current literature and potential solutions are presented upon identifying the weaknesses of the ongoing system for each stakeholder to develop EFL teachers' language assessment literacy in contexts having similar characteristics to those of Turkey.

E. Ölmezer-Öztürk · G. Öztürk (✉)
Anadolu University, Eskişehir, Turkey
e-mail: gokhanoztrk@gmail.com

B. Aydın
TED University, Ankara, Turkey

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_29

29.1 Introduction: Purpose and Testing Context

Teaching English is an important political and social phenomenon in Turkey, just like in many other parts of the world. Turkish students study English for 11 years, beginning with the second grade, until they start their education in the fields they choose at university. Yet, most of the students in the university try hard to learn English during their intensive language education in the preparatory schools. When we look at the big picture, we see that the ultimate result of these 11 years is not at the desired level.

One of the main reasons for this failure has always been pointed to as the inconsistencies in the assessment and evaluation system in the country, inappropriate decisions taken by policymakers, and low literacy levels of the teachers in evaluating their students' learning performance. There is major competition, especially in the high stakes national exams, for passing from one stage to another, from primary to secondary, then to high school and university levels. Turkish students and teachers are under great pressure to be successful in these exams, and although they try hard to survive in such an exam-oriented educational life, the results generally do not satisfy the majority of the society.

This overall picture represents English as a foreign language learning (EFL) process as well. The scores Turkish learners get from tests measuring their English language proficiency are of the lowest among Organization for Economic Cooperation and Development (OECD) countries (Savaşkan 2016). There is no doubt that constantly changing foreign language teaching policies have had a negative impact on this issue, but it is also undeniable that one of the major reasons underlying this failure is the lack of assessment knowledge among Turkish EFL teachers (Ölmezer-Öztürk and Aydın 2018; Hatipoğlu 2015) engaged in this process. Scrutinizing the Turkish EFL context in terms of English teachers' assessment literacy, this chapter, after defining what assessment literacy is, why it is important, and how literate the language teachers in the world are, mainly focuses on Turkish teachers of English, how they are trained, and the problems experienced. It also focuses on possible solutions to the problems in language assessment for different stakeholders, mainly including teachers, trainers, and policymakers, with the hope that the picture drawn of the Turkish context can present a model to other countries experiencing similar problems.

There has been an increasing interest in the assessment literacy of language teachers in recent years (Popham 2009), and "assessment literacy is seen as a sine qua non for today's competent educator" (p. 4). There exist many definitions of assessment literacy in the literature. To start with, Stiggins (1995), who coined the term, defined it as "knowing the difference between sound and unsound assessment" (p. 240). Another definition is that of Falsgraf (2005, p. 6), for whom assessment literacy is "the ability to understand, analyze and apply information on student performance to improve instruction." Rooted in assessment literacy, a new term "language assessment literacy" (LAL) has flourished, and this new area includes language teachers' educational assessment knowledge and practices specifically in language teaching and learning. Despite having certain overlapping features with assessment literacy (Taylor 2013, p. 405), language assessment literacy is regarded as a distinct area (Stiggins 1991; Inbar-Lourie 2017). LAL includes some unique features that only exist in language assessment, and it both includes assessment skills and language-related skills of teachers (Inbar-Lourie 2017). For Pill and Harding (2013, p. 382), many competencies that are necessary for "understanding, evaluating and creating language tests and analyzing test data" make up LAL. Another definition of LAL by Tsagari and Vogt (2017, p. 377) is "the ability to design, develop and critically evaluate tests and other assessment procedures, as well as the ability to monitor, evaluate, grade and score assessment on the basis of theoretical knowledge."

Since assessment literacy is regarded as a bridge connecting learner achievement and the quality of assessment (Mertler 2009), each and every language teacher has to have a solid background in language assessment. Only with this background, it is possible to talk about assessment literate teachers and their more informed decisions leading to better assessment practices. However, though it is argued that the exposure of teachers to educational assessment should not be restricted to one course on testing in teacher education programs (Popham 2009), a number of studies show that language teachers' education in terms of LAL is still problematic. For example, Mertler and Campbell (2005) found that teachers participating in the study were assessment illiterate, and they did not start their professions with enough knowledge in language assessment. Similarly, for Tsagari and Vogt (2017) and Xu and Brown (2017), the participants were not assessment literate enough to carry out their assessment-related practices effectively. In addition, some other studies in the field focused on revealing the needs of pre-service and in-service teachers in terms of language assessment literacy (Inbar-Lourie 2008; Malone 2013; Scarino 2013), and the possible effects of training, mostly workshops specifically designed for increasing the LAL level of the participants (Lam 2015; Baker and Riches 2017). These studies indicate that many teachers did not feel themselves ready for their professions as assessors. In other words, even though these language teachers had key roles in assessment practices, they were not well equipped with necessary assessment literacy skills. Thus, understanding LAL levels of teachers is necessary, because this understanding "is a good departure point for promoting both assessment literacy research and teacher development in education" (Xu and Brown 2017, p. 134).

29.2 Testing Problem Encountered

In Turkey, within the unified model of higher education executed by the Higher Education Council (HEC), EFL teachers are trained at university level during their pre-service period. This initial four-year program is highly standardized and run by HEC in a top-down manner which leaves no space for institutional flexibilities at universities. After graduation, EFL teachers in Turkey are appointed by the Ministry of National Education (MoNE) to work at state schools, are recruited by universities as instructors to work in schools of foreign languages, or are hired by private universities or schools to teach English. Since all the processes may significantly differ regarding the training of teachers in terms of their LAL, and the amount of work they engage in about testing and assessment practices in those institutions, focusing on each of them separately and in detail will present an overall picture of the Turkish context.

Initial English language teacher education programs consist of a four-year university education which aims to develop the professional knowledge of pre-service teachers and includes courses on pedagogical sciences, subject matter knowledge and a practicum process in the final year. In the last two decades, HEC, as the only decision-maker in the unified model of higher education in Turkey, has made several reforms in the content of the program, such as decreasing the number of literature and linguistics courses and increasing the number of pedagogical courses in the program with the intention to make the programs more compatible with international TESOL standards (Mahalingappa and Polat 2013). However, the only component of the program which is not influenced by these reform attempts is related to testing and assessment of the foreign language. Within the scope of the current program, which is quite similar to previous ones, pre-service EFL teachers take two courses on testing and assessment; one is testing and evaluation in education, and the other is assessment in foreign language teaching. The former focuses on general terms and concepts in educational sciences and aims to help teacher candidates gain an overall perspective in testing and evaluation. The latter, while mostly covering the principles on how to prepare tests and exams to assess language skills, is limited to a weekly three-hour course in one academic semester only. Although there might be slight differences among universities in terms of the content of this course and some practical elements of language testing and assessment are included in these three hours, they are very limited.

After graduation, the majority of EFL teachers are appointed at state schools as English teachers by Turkish MoNE or recruited as instructors by universities to work at schools of foreign languages and English preparatory programs. For the teachers working at schools, all training activities are organized by MoNE two times in an academic year on a broad range of topics, from classroom management to traffic rules. However, although Turkey's education system is mostly exam-oriented (OECD 2013), there have been few training opportunities on testing and assessment in general, and almost no specific training on language testing and assessment has been organized for English teachers. On the other hand, the situation at universities is slightly different, and more opportunities are provided to the teachers. One-year intensive English language teaching programs for large numbers of freshman students include a relatively systematic assessment organization. Testing and assessment offices are responsible for the preparation and administration of the assessment throughout the year. Teachers working in these testing offices are mostly graduates of ELT programs, or sometimes of other related programs such as translation, literature, or linguistics, and may receive training if provided by their administration. These potential trainings which teachers attend and the practices they engage in when they are assigned in testing offices, are the only opportunities for teachers working at university contexts to improve their LAL.

As explained previously, the main component that contributes to the LAL of Turkish EFL teachers is their pre-service education with two theory-oriented courses. When they start their professions, they receive limited in-service training opportunities organized by MoNE, or by the university administrations, depending on their working context. Although the case of the teachers at universities seems slightly better than the ones working for MoNe, recent studies (Öz 2014; Ölmezer-Öztürk and Aydın 2018) reveal they are not much different from each other in terms of how literate they are. A few studies have been conducted in the field, and they support this argument, indicating that Turkish EFL teachers' language assessment knowledge is quite low (Ölmezer-Öztürk and Aydın 2018), pre-service education is quite limited in terms of developing prospective teachers' LAL (Hatipoğlu 2015), their methods to test students' performance are highly traditional (Öz 2014), teachers have difficulties in transferring their knowledge into testing and assessment practices (Öz and Atay 2017), and they highly rely upon their own assessment preferences, resisting the suggestions they receive in training (Han and Kaya 2014).

As revealed in the studies mentioned above, the main source of the problem behind the illiteracy in teachers' assessment knowledge is the insufficiency of the input they receive. When this insufficient input is given mostly in a theory-oriented way, it is not surprising that novice teachers' experiences do not go beyond the exams they prepare mostly using their intuition. These exams are generally prepared in a multiple-choice format with the pressure of preparing their students for the national exams. Thus, it is possible to conclude that the largest problem of the country in terms of language testing and assessment, and accordingly for LAL of teachers, is the negative washback effect of tests on the language classes, students and teachers. In addition, the in-service training, which is quite insufficient and does not serve teachers' immediate needs, does not help with increasing practicing teachers' assessment literacy. One-shot trainings conducted by trainers who are not familiar with the context or the needs of the teachers do not help to solve the problem either. To illustrate, Ölmezer-Öztürk and Aydın (2019, p. 384) found that these one-shot trainings were unsuccessful in the eyes of the participating Turkish teachers, and the following comments by the participants of this study reveal how teachers feel about these one-shot trainings:

> Not all the information in the trainings is applicable. Thus, the trainings should be context-specific, and train us by taking our institutional factors into consideration. Thanks to this, we could convert all this theory into practice.
>
> The trainings are more beneficial when they are long-lasting and sustainable because it is not very easy to learn new things or to adapt to new information. So, with the help of the recurrent trainings, teachers firstly become more aware of their practices, and start to apply what they have learned in those trainings.

Finally, it can be stated that finding trainers who are experts in both language teaching and language assessment is very problematic. We have academicians who are trained in teaching the language, training teachers on how to teach the language, but the majority of these university-level teacher trainers are not very familiar with the assessment component of language teaching. Thus, the lack of competent trainers in language testing and assessment appears as another problematic point in improving the LAL of Turkish EFL teachers.

29.3 Solution of the Problem

As recently stated by Yastıbaş and Takkaç (2018), understanding the process of how language teachers develop language assessment appropriate for their teaching purposes is necessary in order to interpret the big picture of teachers' assessment literacy. Considering the aforementioned realities and problems in educating assessment literate teachers, it can be clearly seen that English language teachers in the Turkish context have low LAL levels and they need help and support in order to increase their assessment literacy.

In this regard, two points are important to discuss. The first one is related to teachers' resistance to new ways of assessing language. As demonstrated by two recent studies, Turkish EFL teachers have extensively used traditional testing and assessment methods (Öz 2014), rely upon their own preferences, and resist outsider suggestions (Han and Kaya 2014). Considering that resistance to change might be a rooted sociological reality in Middle-Eastern contexts like Turkey and this resistance might be a great obstacle in introducing and implementing innovative and contemporary instructional decisions, it would be a better idea to start with preparing teachers for a change before directly imposing on them trainings in which they do not find something practical for their own specific contexts. In other words, teachers should firstly be given the idea that they can do better in language testing and assessment, and their resistant attitudes should be addressed through pre-trainings so that they have the maximum readiness level for LAL trainings.

The second point that emerges in light of these studies is the ineffectiveness of one-shot training sessions given by outsiders. It is quite evident that language testing and publishing companies and private teacher training institutes give these one-shot sessions throughout the world to increase the LAL levels of teachers. However, these sessions were perceived as ineffective and not long-lasting by the participant teachers in a very recent study (Ölmezer-Öztürk and Aydın 2019). The major problems observed in these sessions are that they are primarily theory-oriented, they lack practical and contextual elements, and they are mostly trainer-centered, hindering collaboration and sharing among the trainees. For this reason, it is believed that instead of one-shot sessions, there should be long-lasting training sessions which focus on more practical elements rather than theory, integrate contextual elements into content, and provide collaborative and sharing opportunities for teachers. Doing so would lead to a greater impact among Turkish EFL teachers in terms of improving their LAL levels.

In addition to these to-the-point suggestions made for the problems identified by the recent studies, a few more general steps should also be taken in terms of instructional decisions related to the LAL of teachers in Turkey. First of all, a large-scale needs analysis study focusing on the EFL teachers in all educational contexts should be conducted to plan a systematic road map toward the improvement of their LAL. Parallel with the findings of this needs analysis, designing a lifelong training program which is in line with the national educational policy and language teaching aims is imperative. The trainings that would be included in the program should help teachers to gain assessment literacy, including all the necessary skills, ranging from choosing and developing appropriate assessment methods to administering, then scoring, and interpreting results for decision-making, and using these results for positive washback for their learners. These literacies can only be gained by sharing good examples using various assessment techniques appropriate to the teachers' contexts. In addition, as stated by Ölmezer-Öztürk and Aydın (2019), all the local needs and institutional factors should also be taken into consideration while designing these trainings, rather than one-shot trainings conducted by trainers who are unfamiliar with the teachers' contexts. Online platforms providing continuous interactive assessment trainings and sharing good practices of teachers can also be used for in-service training purposes.

Finally, during their four-year pre-service teacher training programs, teacher candidates should be equipped with the necessary teaching skills and strategies. Accordingly, they should be trained on how to assess language skills, not with a single three-hour course in an academic semester but with a carefully designed program balancing both theory and practice. Practicing opportunities including not only teaching but also assessment, should be a component of the practicum process in real school contexts, which is also a part of the training program. With the support of their mentor teachers and supervisors, teacher candidates can be encouraged to apply assessment practices appropriate to their teaching aims in addition to the lesson plans and materials they prepare. This will help future teachers to be better decision-makers in their own teaching and assessment practices when they start the profession.

29.4 Insights Gained

In Turkey both pre-service and in-service teacher education programs have problems in preparing language teachers with sufficient language assessment literacy. It seems inevitable that it would be necessary to reconsider the content of both the pre- and in-service programs and to include the courses necessary for providing professional development for the pre- and in-service teachers to help them become literate in language assessment. Considering the low levels of language assessment knowledge of EFL teachers identified in Ölmezer-Öztürk and Aydın's (2018) study, various efforts seem essential for the policymakers to consider. Professional development programs including hands-on experience for teachers, as well as theory addressing teachers' specific contexts and needs, would serve the purpose. These training programs should definitely include how teachers can assess their students' language proficiency in the four skills, not only in the grammar and vocabulary of the foreign language. In the center of these programs must be raising the consciousness of the teachers regarding how they can use various types of assessment tools to give feedback to their learners and how they can benefit from technology in doing so.

29.5 Conclusion: Implications for Test Users

This chapter presented an overall picture of the language assessment literacy of Turkish EFL teachers with primary emphasis on how they are educated in teacher education programs and their in-service training opportunities. Considering the problematic aspects of EFL teaching in the Turkish context, it is clear that one of the major parts is the testing and assessment field, and it is of the utmost importance that teachers who are responsible for teaching the foreign language should also be competent and knowledgeable in conducting the assessment practices in this process. Supporting teachers with sound background knowledge and practice opportunities seems essential to encourage them to build the indispensable bridge between teaching and assessment. Since professional development is a lifelong process, this support should be provided for in-service teachers as well. Thus, it is believed that the suggestions provided by this chapter will form a guideline in educating more assessment literate teachers in Turkey and other contexts having similar social and educational characteristics. It is then the policymakers' responsibility to establish trustable, transparent, and objective assessment policies and applications.

Further research is still necessary to determine the competency of Turkish EFL teachers in using effective assessment tools based on sound language assessment principles and whether they are supporting their own learners with the necessary feedback to help them become effective individuals in the twenty-first century. On the other hand, policymakers should also act in line with the research findings to make more effective and sustainable decisions to improve Turkish EFL teachers' LAL levels.

References

Baker, B. A., & Riches, C. (2017). The development of EFL examinations in Haiti: Collaboration and language assessment literacy development. *Language Testing, 35*(4), 557–581. https://doi.org/10.1177/0265532217716732.

Falsgraf, C. (2005). *Why a national assessment summit? New visions in action.* National Assessment Summit, Meeting conducted in Alexandria, VA. http://www.nflrc.iastate.edu/nva/worddocuments/assessment_2005/pdf/nsap_introduction.pdf. Retrieved April 14, 2018.

Han, K., & Kaya, Hİ. (2014). Turkish EFL teachers' assessment preferences and practices in the context of constructivist instruction. *Journal of Studies in Education, 4*(1), 77–93.

Hatipoğlu, Ç. (2015). English language testing and evaluation (ELTE) training in Turkey: Expectations and needs of pre-service English language teachers. *ELT Research Journal, 4*(2), 111–128.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402.

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment* (pp. 1–14). Cham: Springer International Publishing.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32*(2), 169–197.

Mahalingappa, L. J., & Polat, N. (2013). English language teacher education in Turkey: Policy vs academic standards. *European Journal of Higher Education, 3*(4), 371–383.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344.

Mertler, C. A. (2009). Teachers' assessment knowledge and the perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(2), 101–113.

Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the assessment literacy inventory*. Paper presented at the annual meeting of the American Research Association, Montreal, QC, Canada. https://eric.ed.gov/?id=ED490355. Retrieved May 23, 2018.

OECD. (2013). *Education policy outlook: Turkey*. Paris: OECD Publishing. http://www.oecd.org/education/EDUCATION%20POLICY%20OUTLOOK%20TURKEY_EN.pdf. Retrieved May 22, 2018.

Ölmezer-Öztürk, E., & Aydın, B. (2018). Investigating language assessment knowledge of EFL teachers. *Hacettepe University Journal of Education*. https://doi.org/10.16986/HUJE.2018043465.

Ölmezer-Öztürk, E., & Aydın, B. (2019). Voices of EFL teachers as assessors: Their opinions and needs regarding language assessment. *Journal of Qualitative Research in Education, 7*(1), 373–390.

Öz, H. (2014). Turkish teachers' practices of assessment for learning in the English as a foreign language classroom. *Journal of Language Teaching and Research, 5*(4), 775–785.

Öz, S., & Atay, D. (2017). Turkish EFL teachers' in-class language assessment literacy: Perceptions and practices. *ELT Research Journal, 6*(1), 25–44.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381–402.

Popham, J. W. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*, 4–11.

Savaşkan, İ. (2016). Turkey's place in the rankings of the English proficiency index. *Journal of Teacher Education and Educators, 5*(2), 192–208.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309–327.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72,* 534–539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77,* 238–245.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412.

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment, 6*(1), 41–64.

Xu, Y., & Brown, G. T. L. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment, 6*(1), 133–158.

Yastıbaş, A. E., & Takkaç, M. (2018). Understanding language assessment literacy: Developing language assessment. *Journal of Language and Linguistic Studies, 14*(1), 178–193.

Chapter 30
Some Practical Consequences of Quality Issues in CEFR Translations: The Case of Arabic

Björn Norrbom and Jacob Zuboy

Abstract Test users require interpretable test scores for Arabic L2 learners' proficiency. Relating scores to the Common European Framework of Reference for Languages (CEFR) is an effective way to meet this need and is a potentially powerful tool to enhance assessment literacy, facilitate communication among stakeholders in education, and drive educational reform efforts. Low quality and conflicting translations impede stakeholders' ability to do this. This chapter investigates the quality and possible implications of the Arabic CEFR translation referred to by the Council of Europe (COE). The analysis shows that the translation suffers from serious quality problems in terms of central terminology, level designations, and style. Implications for Arabic-speaking users of the Framework are that communication and dissemination of educational policies, learning goals, and assessment requirements are made more difficult, which may further impede development in a region broadly considered in need of educational reform. In the short term, users should exercise care in interpreting learning, teaching, and assessment products based on current Arabic translations of the CEFR. In the longer term, an official COE Arabic CEFR translation is needed along with other supporting materials such as a complete multilingual CEFR glossary, for the Framework to achieve its intended impact.

30.1 Introduction: Purpose and Testing Context

Since its 2001 publication, the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001a) has enjoyed considerable influence on language learning, teaching, and assessment across Europe and beyond. For language

B. Norrbom (✉) · J. Zuboy
National Center for Assessment, Riyadh, Saudi Arabia
e-mail: b.norrbom@etec.gov.sa

J. Zuboy
e-mail: j.zuboy@etec.gov.sa

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_30

testers, using the CEFR has several advantages such as comparing tests across countries and contexts and overall test improvement (Kantarcioğlu 2012). Linking examinations to the CEFR is increasingly becoming an expectation (Khalifa and Weir 2009).

The Council of Europe (COE) frequently cites the large number of CEFR translations into European and non-European languages, now over 40, (COE n.d.a) as evidence of its inherent merit (COE 2014; COE n.d.b) and, by extension, its validity. Two Arabic translations have been published (Dar Elias 2008; Umm al-Qura University 2016). Both of these were previously referred to on the COE webpage, but the 2008 version has at the time of writing been removed. The 2016 version—our focus here—remains available as a free download from the COE webpage.

The challenge of effectively disseminating a complex framework such as the CEFR in many languages has been documented (Little 2013), but remains underreported and under-researched. For guiding and regulatory documents impacting large numbers of people and requiring publication in multiple languages, detailed and methodical translation procedures are necessary to ensure quality. Such procedures typically include multiple drafts and reviews undertaken both individually and collectively by language and content professionals. Examples where such procedures have been successfully applied include the DIALANG project (A. Huhta, personal communication November 7, 2018) and the Spanish translation of the California Common Core State Standards (San Diego County Office of Education 2012). The PISA (2018) translation guidelines provide an excellent example in guidance and oversight of the translation process.

The key to CEFR linkage is flexibility and adaptability to context (COE 2009), and in order for institutions to meet these criteria, and for their qualifications to be comparable, they must be able to fully access the Framework, regardless of which language version they use. Obviously, low quality and conflicting CEFR translations impede their ability to do so.

Thus it is surprising that the COE does not provide any information regarding how translations were produced or what criteria apply for a translation to be referred to on its webpage. The only disclaimer is copyright-page boilerplate on the documents themselves, stating that the translation has been produced by arrangement with the COE but that it is the sole responsibility of the translator.

The permissiveness[1] of the COE's approach is difficult to explain given that it has officially validated a large number of European Language Portfolios (ELP) that are based on the CEFR (Alderson 2007). It also provides very strong suggestions as to how the standard-setting process should be carried out when linking examinations to the CEFR (COE 2009), but has not even acknowledged the question of whether the translations of the source texts from which the standards are drawn should be officially vetted, let alone who should be responsible for the vetting.

Turning now to the analysis of the 2016 Arabic translation, we argue that many or all of the problems we discuss would have been largely avoided had the COE

[1] See Trim (2012, p. 30) for discussion of "the permissive nature" of the COE's approach.

defined a set of quality criteria for the CEFR translations produced by arrangement with translators and promoted through its website.

30.2 Testing Problem Encountered

For the sake of brevity, we focus our analysis on Chapter 3, “The Common Reference Levels.” Common Reference Levels A1-C2 are the first noticed (North 2014) and most recognized (Figueras 2012) aspect of the Framework. As Coste (2007) so artfully illustrates, they are the synecdochic sail that has come to supplant the CEFR ship. They are now “common currency” in Europe and beyond (Figueras 2012, p. 479).

This analysis focuses on three main areas in the translation: central terminology, level designations, and style which encompasses punctuation and paratextual features. The data collected from the translation for the purpose of this analysis is very rich and can only be partially represented here.

30.2.1 Central Terminology

In order for frameworks and other guiding and regulatory documents to work effectively in learning, teaching, and assessment, a clear, common understanding of core terminology and concepts is required within and across languages (Milanovic 1998).

Among the most central terminology in Chapter 3 is the title of the chapter itself, “The Common Reference Levels.” Table 30.1 shows how this phrase is rendered in the 2016 Arabic translation.

As seen from the table, five different phrases are used. The potentially most confusing one is “The Different Reference Levels” which appears juxtaposed to the original.

Table 30.1 “The Common Reference Levels”: translations: Occurrences and back translations

Arabic	Back translation	No. of occurrences
al-mustawayāt al-marji‘īyyah al-mushtarakah	*“The Common Reference Levels”*	12
al-mustawayāt al-marji‘īyyah	*“The Reference Levels”*	3
mustawayāt al-iṭār al-marji‘ī	*“The Framework Levels”*	5
al-mustawayāt al-marji‘īyyah al-mukhtalifah	*“The Different Reference Levels”*	1
mustawayāt al-marji‘īyyāt al-mushtarakah	*“The Levels of Common References”*	1

Table 30.2 Translations of the term "Waystage" in Chapter 3

Arabic	No. of occurrences
mustawá al-asās	8
mustawá al-asās aw mā qabla al-bidāyah	2
mustawá al-asās (Waystage)	1
mustawá al-asās aw al-ikhtirāq (Waystage)	1
Waystage (English original, not transliterated)	1
al-ikhtirāq	1
[left out]	1

Likely consequences are self-evident: If such central terminology is not presented consistently, the Framework's user-friendliness, and thus its value as a reference, is significantly diminished. Good translation practice dictates the consistent use of the closest Arabic equivalent translation of the phrase, which we feel is the most commonly occurring one in the translation.

30.2.2 Level Designations: Waystage

"Waystage" is a neologism coined by van Ek and Trim (1998) to name the COE specification roughly corresponding to CEFR level A2. As such, it has proven resistant to translation (COE 2001a). The French source text (COE 2001b) includes the English terminology in juxtaposition with the French translation. And several translations (e.g., Goethe-Institut 2003; Nederlandse Taalunie 2008; Skolverket 2007) retain the English original throughout. Borrowing is common practice for such technical terms (Heim and Tymowski 2006). In languages with no graphological equivalence, transliteration is required.

Table 30.2 shows the Arabic translation of "Waystage" in Chapter 3:

As seen from the table, the term is rendered seven ways, the most common being *al-asās*.[2] Two different terms are used together in some of the entries, and the English word is sometimes retained either by itself or together with one or more Arabic terms. Also, the term is left out once. Finally, the translator does not once choose to transliterate the word in Arabic.

Remarkably, as shown in Table 30.3, the same word *al-asās* is actually used to denote three different CEFR levels ("Breakthrough"/A1, "Waystage"/A2, and "Threshold"/B1).

[2] Similar inconsistencies were also observed for Breakthrough (A1) and Threshold (B1). However, space does not permit us to report these in detail here.

Table 30.3 Use of the Arabic term “al-asās” in Chapter 3

Arabic	English	No. of occurrences
“al-asās”	“Waystage”	8
	“Threshold”	1
	“Breakthrough”	1
	“Basic”	1

This is perhaps most pronounced in the following example where both “Waystage” and “Threshold” are translated the same way (bold emphasis added in both English and Arabic): “[T]he set of Common Reference Levels: A1 (Breakthrough), A2 (**Waystage**), B1 (**Threshold**),[…]” (COE 2001a, p. 30) becomes “*Majmū‘at al-mustawayāt al-marji‘īyyah al-mushtarakah, wa hiya mustawá al-inṭilāqah, wa mustawá* ***al-asās*** *aw mā qabla al-bidāyah, wa mustawá al-‘utbah aw* ***al-asās***” (Umm al-Qura University 2016, p. 47).

In addition, *al-asās* is used to represent the term “Basic,” as in “Basic User,” which in the English original is an overarching term for levels A1 and A2.

It is very difficult to imagine a scenario where the inconsistencies and overlap in the translation would not cause confusion for readers with regards to CEFR level designations. Instead of championing the CEFR ideals of transparency and coherency (COE 2001a), the Arabic translation is opaque and incoherent. The instance where the same Arabic word is used for two consecutive CEFR levels may even cause a “communication breakdown” (COE 2001a, p. 93).

A more effective solution would be to maintain the original English term “Waystage” as a loanword, simply transliterating it. The word “*al-asās”* would serve well as a translation for “Basic” if used consistently and for this purpose only.

30.2.3 Style

A text’s style bears meaning that cannot be distilled from its message. Absent one, neither is complete. It is critical that style be carefully preserved in translation (Ghazala 1995).

The CEFR adheres to no manual of style. It includes neither a key to typographical features nor a glossary of technical terms. All definitions and reading instructions are embedded (e.g., Sections 2.1 and 3.7, respectively). The Framework relies heavily on stylistic devices such as punctuation and paratextual features to facilitate comprehension for readers and guide them in use. If these features are not rendered equivalently in translation, significant aspects of the CEFR’s meaning remain inaccessible, effectively untranslated.

30.2.3.1 Punctuation

Punctuation lends clarity, coherence, and cohesion to writing. It is an indispensable element of both English and Arabic texts (Ghazala 1995). Still, whereas punctuation is applied rigorously and methodically in English language prose (Ghazala 1995), its application in Arabic is haphazard, subject to authors' "taste and discretion" (Elewa 2015, p. 118). Scholars of translation have described the use of punctuation in Arabic as "erratic" (Williams 1989, pp. 89–90), "inconsistent and arbitrary" (Elewa 2015, p. 118), and "[at times bordering] on the chaotic" (Husni and Newman 2013, p. 235). It is "considered as an ornamentation […] and is, therefore, disregarded, sometimes completely" (Ghazala 1995, p. 272). This has resulted in a "quite unfortunate situation in Arabic writing" (p. 272).

Practical consequences of this unfortunate situation are evident throughout the 2016 Arabic translation of the CEFR. A point among the most salient, recurring, and broadly referenced throughout the Framework is Common Reference Level A2. Table 30.4 shows the different representations of the 19 occurrences of "A2" in Chapter 3 of the Arabic translation.

As can be seen from the table, whereas the A2 label appears 19 times in the same format (e.g., A2) in the English version, it appears only 18 times in the Arabic translation, and is rendered six different ways. The various representations of this singular reference point make it much more difficult to identify. Considering that the primary audience of the Arabic translation consists of people with no knowledge of English and no previous knowledge of or experience with the CEFR, it is clear that the Common Reference Levels as presented here are not transparent—not readily understandable and usable. This undermines the criterion of user-friendliness, which is critical to the CEFR's aim of helping "partners to describe levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems or qualifications" (COE 2001a, p. 21). There can be no accurate description, no comparison, without a consistent basis of reference. In this instance, the translator need only consistently use the lexically equivalent Arabic designation.

Table 30.4 Common reference level A2: Variations in Arabic translation

English	Arabic	No. of occurrences
A2	(A) 2	7
	A2	3
	[English original: A2]	3
	A (2)	2
	A/2	2
	2A	1
	[left out]	1

Table 30.5 Italicized words in CEFR section 3.6, English and Arabic versions

	Total words in 3.6	Words italicized	% of text italicized
English	1832	1121	61.2%
Arabic	1777	0	0%

30.2.3.2 Paratextual Features

Leeuwen (2006, p. 139) submits that “[m]uch of the cohesive work that used to be done by language is now realised, not through linguistic resources, but through layout, colour and typography.”

The resources Leeuwen refers to are termed “paratextual features” in CEFR terminology (COE 2001a, p. 90). They function in writing but are “tied to the spatial medium and not available to speech” (p. 94). The CEFR depends heavily upon paratextual features to mediate complex linguistic information for users.

Section 3.6 “[d]escribes the *salient* features of the levels, as made up by the illustrative descriptors”[3] (COE 2009, p. 18, *emphasis added*). Trim (2001, p. 17) states that 3.6 not only describes the Common Reference Levels but “justifies” them. Second to the Global Scale, it is likely the most known and referenced portion of the Framework. Understanding 3.6 is critical to the use and understanding of the CEFR (COE 2009).

In the English source text, the salience of the descriptors is reflected typographically through italicization. In the Arabic translation, the descriptors are typographically indistinguishable from the surrounding text.

As Table 30.5 illustrates, the contrast is profound:

This, of course, lead us to ask why. Suffice it to say that we have speculated at great length and depth in separate discourse. What is clear is that italics are a common feature of modern Arabic typography (Rjeily 2011) that is used (Adobe Systems Incorporated 2012; Microsoft 2017; World Bank 2004) or at least understood and accounted for (Alqinai 2010) in English-Arabic translation. Moreover, it is clear that the translators themselves were at least aware of the term, as they accurately translated both of its two appearances in the Framework with “*al-khaṭṭ al-māil”* (Umm al-Qura University 2016, pp. 364, 366). It is mystifying that they never once applied the technique.

Authors and editors often use paratextual features to plan and present texts in accordance with the manner in which they intend them to be read by their target audience[4] (COE 2001a). The use of italics in 3.6 is not a typographical adornment to be discarded in translation without consequence. It is a meaning-bearing authorial choice to be as accurately as possible preserved.

[3] The illustrative descriptors “focus on **aspects** [of competence] **that are new and salient**” at the different levels (COE 2018, p. 41, *emphasis in original*).

[4] See Jones (2012) for a critical discussion of the paratextual features of 3.6.

30.3 Resolution of the Problem

While the analyses presented here merely touch upon the issues brought to light, they indicate that the existing Arabic translation suffers from quality and content problems that compromise its utility as an external standard or reference.

Resulting implications are that communication and dissemination of educational policies, learning goals, and assessment requirements are made more difficult, which may further impede development in a region broadly characterized as in need of educational reform (Galal et al. 2008).

Soliman (2018) calls for a clear framework that outlines the language competences required in Arabic L2 teaching and learning, specifically referring to the CEFR.

Pursuant to that, test developers and users require interpretable test scores for Arabic L2 learners' Arabic proficiency, and relating scores to the CEFR is an effective way to meet this need. This, however, assumes that professionals have complete access to the CEFR's contents.

One very tangible effect experienced by the authors first hand is that we have been unable to use the Arabic CEFR descriptors to develop writing assessment rubrics. This has led to a situation where raters recruited need to be fully proficient in both English (to be able to use the rubrics) and Arabic (to rate the performances), forcing us to exclude expert users of Arabic to whom the Framework is not accessible in English.

Given the constantly touted high-number and diverse linguistic range of the current CEFR translations, and the difficulty users in numerous countries and contexts have reported in understanding and incorporating the Framework into their practice (Martyniuk and Noijons 2007), it appears indefensible that the COE has taken the translation process for granted.

30.4 Insights Gained

We agree with the COE when it says that, "Member states are responsible for guaranteeing the quality and fairness of testing and assessment on the basis of the *existing guidelines*" (COE n.d.b *emphasis added*). And for the translations, "Relevant national authorities and/or publishers are responsible for dissemination" (COE n.d.b). However, for languages other than English and French, the CEFR translations form a central part of the existing guidelines. While the COE can rightfully disclaim responsibility for the link between the existing guidelines and actual national or other assessment, it cannot, in our estimation, disclaim responsibility for the quality of CEFR translations, as these are the very basis for guaranteeing quality and fairness.

It seems the COE wants it both ways: It celebrates the notoriety and universal relevance in the field the CEFR has garnered, encouraging its adaptation and use beyond Europe, but declines to govern the power and influence this global reach entails.

30.5 Conclusion: Implications for Test Users

The COE's permissiveness (Trim 2012) in this regard risks initiating a self-reinforcing contravalidity cycle: the lack of governance increases the CEFR's fame, influence, and distribution while simultaneously compromising its transparency and coherency and hence its validity. Each successive iteration of the cycle furthers its reach and influence and further diminishes its validity. If validity is compromised in furtherance of reach and influence, the dissemination of the Framework will ultimately defeat its own purpose.

A minimum requirement would be that the COE vet and approve a translation before referring to it on its website or at least clearly disclaim association. Its current practice is quite the opposite as it simply states on its website that "The CEFR is available in 40 languages" (COE n.d.a) and provides links to these translations. This is hard to interpret as anything other than an endorsement of these translations. All COE publications, and publications endorsed by the COE, directly or indirectly, would generally be viewed as legitimate parts of the "existing guidelines."

In the balancing act of standardization versus local contextualized dissemination, CEFR translations appear to have been placed in the incorrect category as they essentially constitute local contextualization and dissemination while being endorsed by the COE as part of its "existing guidelines."

The Council of Europe has, among others, issued an official guide for linking examinations to the CEFR (COE 2009), a guide for the production of Reference Level Descriptions (COE 2005), and a guide for users (Trim 2001). We suggest the COE issue CEFR translation guidelines and strongly encourage their use, preferably in an official recommendation from the Committee of Ministers. Guidelines should include a key to typographical conventions.

Moreover, a CEFR bibliography for translators listing resources to develop a deeper understanding of the Framework should be published.

Based on the results presented above, we recommend that the COE remove the 2016 Arabic translation from its website, adding an explanatory note on its decision to do so.

The production of a CEFR glossary is long overdue. While many technical terms are defined and explained within the CEFR text or accompanying documents, they are not easy to extract or (re)contextualize, often requiring mediation by expert users. A comprehensive glossary of terms used in the CEFR approved by the COE and accompanying all materials would benefit translators and other users greatly (Martyniuk and Noijons 2007), and do much to promote assessment literacy (Taylor 2009).

The development of a provisional glossary, including full Arabic translations, is developed parallel to the current chapter (National Center for Assessment (NCA), forthcoming). After COE vetting, it will be published online and accessible for free. In light of the serious quality issues demonstrated for the Arabic translation, it might be time well spent for the COE and other key stakeholders to review and analyze the quality of other CEFR translations; the Chinese version, for example, where cumulative anecdotal evidence suggests serious quality concerns.

Ultimately, a professionally developed and official (or vetted) COE Arabic CEFR translation is needed for the Framework to achieve its intended impact in the Arabic-speaking world, where there is limited evidence of its reach (Phipps 2012). A pilot translation of CEFR Chapter 3, applying documented professional procedures, is currently in progress (NCA, forthcoming) and will be submitted to the COE and published for free access in due course. With its publication a report describing the process will also be made available, provisioning valuable insights to users.

References

Adobe Systems Incorporated. (2012). *Myriad Arabic*. https://www.adobe.com/content/dam/acom/en/products/type/pdfs/Myriad-Arabic-Online-Specimen.pdf. Retrieved Feburary 26, 2019.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659–663.

Alqinai, J. B. (2010). Mediating punctuation in English Arabic translation. *Journal of Applied Linguistics, 5*(1), 5–29.

Coste, D. (2007, February). *Contextualizing uses of the Common European Framework of Reference for Languages.* Strasbourg: Language Policy Division.

Council of Europe. (2001a). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2001b). *Cadre Europeen Commun De Reference Pour Le Langues: Apprendre, Enseigner, Evaluer*. Paris: Les Éditions Didier.

Council of Europe. (2005). *Reference level descriptors for national and regional languages (RLD): Guide for the production of RLD version 2 November 2005.* Strasbourg: Language Policy Division.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching and assessment (CEFR): A manual.* Strasbourg: Language Policy Division.

Council of Europe. (2014). *Languages for democracy and social cohesion: Diversity, equity and quality.* Sixty years of European co-operation. Strasbourg: Language Policy Unit.

Council of Europe (2018). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* Companion volume with new descriptors. Strasbourg: Council of Europe.

Council of Europe. (n.d.a). https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806d8893. Retrieved Feburary 26, 2019.

Council of Europe. (n.d.b). https://www.coe.int/en/web/common-european-framework-reference-languages/. Retrieved Feburary 26, 2019.

Dar Elias. (2008). *al-Iṭār al-marjiʿī al-ūrūbbī al-mushtarak lil-lughāt dirāsah tadrīs taqyīm* (The Common European Framework of Reference for Languages: Learning, teaching, assessment). Cairo: Dar Elias.

Elewa, A. (2015). *Levels of translation.* Cairo: Qalam for Translation and Publication.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477–485.

Galal, A., Welmond, M., Carnoy, M., Nellemann, S., Keller, J., Wahba, J., et al. (2008). *The road not traveled: Education reform in the Middle East and North Africa (English).* Washington, DC: World Bank.

Ghazala, H. (1995). *Translation as problems and solutions: A coursebook for university students and trainee translators.* Beirut: Dar wa Maktabat Al-Hilal.

Goethe-Institut. (2003). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. München: Goethe-Institut.

Heim, H. M., & Tymowski, W. A. (2006). *Guidelines for the translation of social science texts*. New York, NY: American Council for Learned Societies.

Husni, R., & Newman, D. L. (2013). *A-Z of Arabic-English-Arabic translation*. London: Saqi Books.

Jones, N. (2012). Defining an inclusive framework for languages. In E. D. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks*. Proceedings of the ALTE Kraków conference, July 2011 (pp. 105–117). Cambridge, UK: Cambridge University Press.

Kantarcioğlu, E. (2012). *Relating an institutional proficiency exam to the CEFR: A case study*. London: University of Roehampton.

Khalifa, H., & Weir, C. J. (2009). *Examining reading*. Cambridge: Cambridge University Press.

Leeuwen, T. (2006). Towards a semiotics of typography. *Information Design Journal, 14*(2), 139–155.

Little, D. (2013). Review of the book The Common European Framework of Reference: The Globalisation of language education policy, ed. by M. Byram and L. Parmenter. *The Canadian Modern Language Review/La Revue Canadienne des langues vivantes*, *69*(4), 514–522.

Martyniuk, W., & Noijons, J. (2007, February). Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe Member States. In *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities*. Council of Europe. Strasbourg: Language Policy Division.

Microsoft (2017). *Arabic Style Guide*. Microsoft.

Milanovic, M. (Ed.). (1998). *Multilingual glossary of language testing terms*. New York, NY: Cambridge University Press.

Taalunie, Nederlandse. (2008). *Gemeenschappelijk Europees Referentiekader voor Moderne Vreemde Talen: Leren, Onderwijzen, Beoordelen*. Den Haag: Nederlandse Taalunie.

North, B. (2014). Putting the Common European framework of reference to good use. *Language Teaching, 47*(2), 228–249.

Phipps, A. (2012). Series editor's preface. In M. Byram & L. Parmenter (Eds.), *The Common European framework of reference, The globalisation of language education policy* (pp. ix–x). Bristol: Multilingual Matters.

PISA. (2018, March). *PISA 2018 translation and adaptation guidelines*. https://www.oecd.org/pisa/pisaproducts/PISA-2018-TRANSLATION-AND-ADAPTATION-GUIDELINES.pdf. Retrieved Feburary 26, 2019.

Rjeily, R. A. (2011). *Cultural connectives*. New York, NY: Mark Beaty Publishers.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29,* 21–36.

San Diego County Office of Education. (2012). *California Spanish language development standards kindergarten through grade 12 EN ESPAÑOL*. https://commoncore-espanol.sdcoe.net/CaCCSS-en-Espanol/SLA-Literacy. Retrieved Feburary 26, 2019.

Skolverket, S. (2007). *Gemensam europeisk refernsram för språk: Lärande, undervisning och bedömning*. Stockholm: Skolverket.

Soliman, R. (2018). The implementation of the Common European Framework of Reference for the teaching and learning of Arabic as a second language in higher education. In K. M. Wahba, L. England, & Z. A. Taha (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (Vol. II, pp. 118–137). New York, NY: Routledge.

Trim, J. (2001). Guidance to all users. In J. Trim (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment: A guide for users*. Language Policy Division: Strasbourg.

Trim, J. L. (2012). The Common European Reference for Languages and its background: A case study of cultural politics and educational influences. In M. Byram & L. Parmenter (Eds.), *The Common European Framework of Reference, The globalisation of language education policy* (pp. 13–34). Bristol: Multilingual Matters.

Umm al-Qura University. (2016). *al-Iṭār al-marji'ī al-ūrūbbī al-mushtarak lita'allum al-lughāt wa ta'līmuhā wa taqyīmuhā* (The Common European Framework of Reference for Languages: Learning, teaching, assessment). Mecca: Umm al-Qura University.

van Ek, J., & Trim, J. (1998). *Waystage 1990*. Cambridge: Cambridge University Press.

Williams, M. P. (1989). *A comparison of the textual structures of Arabic and English written texts. A study in the comparative orality of Arabic*. Doctoral dissertation. https://etheses.whiterose.ac.uk/469/. Retrieved Feburary 26, 2019.

World Bank. (2004). *World Bank translation style guide: Arabic edition.* https://siteresources.worldbank.org/TRANSLATIONSERVICESEXT/Resources/Translation_Style_Guide_Arabic.pdf. Retrieved Feburary 26, 2019.

Chapter 31
Assessment Literacy and Assessment Practices of Teachers of English in a South-Asian Context: Issues and Possible Washback

Radhika De Silva

Abstract Teachers of English are usually involved in setting and scoring assessments but their assessment literacy and assessment practices are rarely studied. The purpose of this chapter is to discuss the findings of a study which was conducted with secondary level school teachers of English in Sri Lanka on their assessment literacy and assessment practices. The present study adopted a mixed-methods research design. The sample consisted of 150 teachers from different types of state and private schools in Sri Lanka. A questionnaire which consisted of both closed- and open-ended questions was administered to these teachers, and interviews were conducted with a sub-sample of teachers. Assessments set by these teachers were also studied. The findings revealed that teachers in the sample were literate about assessment principles to a fair extent. However, their application of that knowledge when designing and administering assessments was not so satisfactory. There were some teachers who had no training in assessment principles but were expected to design tests. The practices those untrained teachers adopt when setting assessments were also discovered through this study. The study revealed many factors that affect teachers' choice of content, number of questions, testing techniques, and scoring methods.

31.1 Introduction: Purpose and Testing Context

Teachers of English need to be aware of the continuous developments in the field of language teaching and assessment. Many teachers usually get updated about new approaches to teaching through in-service teacher-training programs and professional development courses. In addition to teaching, almost all language teachers in schools and universities engage in formal and informal assessment of their students. While informal assessments are embedded in other classroom activities, formal assessments are usually designed to elicit students' attainment of the objectives of a course. All formal assessments are not considered as tests and, according to

R. De Silva (✉)
The Open University of Sri Lanka, Nugegoda, Sri Lanka
e-mail: krsil@ou.ac.lk

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_31

Brown and Lee (2015), “tests are usually relatively time-constrained …and draw on a limited sample of behavior” (p. 490). Like language teachers all over the world, teachers of English in Sri Lanka are involved in setting and scoring both informal and formal language assessments, and these formal assessments include teacher-made tests as well. However, there is no formal mechanism in place to monitor the assessment literacy and assessment practices of the teachers who are engaged in testing frequently. The present study attempted to identify the assessment literacy and assessment practices of a sample of secondary level teachers of English in Sri Lanka, and the purpose of this chapter is to discuss the findings of that study which revealed their assessment literacy and assessment practices.

31.2 Testing Problem Encountered

Sri Lanka is a country which has different types of schools: non-fee-levying government schools, fee-levying semi-government and private schools and international schools. While international schools provide education in the medium of English, all the other schools use either the mother tongue or bilingual education as the medium of instruction. In these schools, the English language is taught by teachers with different educational and professional qualifications, and their experience in teaching at different levels also vary considerably. Irrespective of their qualifications and experience, almost all English language teachers in Sri Lankan schools are expected to set English language assessments mostly in the form of test papers. A common issue that can be observed in most of these assessments is that they do not conform fully to standard practices. When samples of test papers were analyzed, it was observed that the objectives behind different assessments were unclear and the scoring procedures adopted by some teachers were not reliable. Hence, it is important to discover the practices adopted by teachers and the reasons for those practices and to identify possible constraints faced by these teachers when constructing, administering, and scoring assessments.

Research question 1: What is the assessment literacy of secondary school teachers of English in Sri Lanka?

Research question 2: What are the assessment practices of secondary school teachers of English in Sri Lanka?

31.3 Review of the Literature

According to Coombe et al. (2009), assessment literacy is a vital component of a language teacher’s professional knowledge. Assessment literacy is defined as the knowledge about assessment principles and the ability to develop and score tests and interpret test results (Green 2014). Webb (2002) defines assessment literacy as the knowledge a teacher possesses about how to assess what students know and what they can do, the ability to interpret the results of assessments, and to use those results in improving the effectiveness of the program and students' learning. The importance of understanding the role of language assessment literacy in the language teaching profession is stressed by Fulcher (2012), who emphasizes the need "to articulate its role in the creation of new pedagogic materials and programs in language testing and assessment to meet the changing needs of teachers and other stakeholders for a new age" (p. 113). The importance of collaboration between teachers and testing specialists in developing and administering tests keeping test takers' needs in mind has been stressed by researchers (Hamp-Lyons 2003, p. 182).

Studies on assessment literacy and assessment practices in other contexts reveal that teachers lack knowledge about effective assessment procedures and they do not receive adequate training on good assessment practices (Djoub 2017). Assessment literacy is also defined as the ability to use assessment procedures effectively in a given educational context, and Taylor (2009) stresses the need for equipping teachers with the knowledge of effective assessment procedures.

As Elsshawa et al. (2016) point out, teachers usually postpone addressing assessment issues until the final stage of an academic year. This affects the main purpose of assessment, which is to help students learn (assessment for learning) and results in measuring students' achievements at the end of their study period (assessment of learning). Djoub (2017) attempted to gauge teachers' assessment literacy using questionnaire data, and the researcher found that the assessment literacy of the teachers was low and they had not received any training in testing and assessment. However, generalizing of the findings is not possible due to small sample size and lack of triangulation of data. A large-scale study by Vogt and Tsagari (2014) which investigated the assessment literacy of foreign language teachers revealed that the majority of the teachers did not possess adequate training in testing and assessment.

An empirical tool to study language teachers' assessment literacy needs was developed by Fulcher (2012) and was administered via the internet. Fulcher mentions several limitations of the study which were self-selecting participants, the inability to identify the needs of different subgroups, and the coding of qualitative data. Despite these limitations, Fulcher (2012) states that the results of his study could be used to show "how a research base can be constructed and used to support pedagogic decisions in the structuring and delivery of materials for teaching language testing and improving assessment literacy" (p. 15).

31.4 Methodology

The present study used a mixed-methods research design. The quantitative data were obtained through a questionnaire (see Appendix 31.1), and qualitative data were obtained through open-ended questions in the questionnaire and in-depth interviews of a subsample of teachers. The interview guide can be found in Appendix 31.2. A content analysis of English Language test papers set by these teachers was also conducted to check the application of their knowledge in assessment principles when setting test papers. The questionnaire was administered to 150 teachers representing four provinces in the country, namely Western, Northern, North-Western, and Central provinces, and the response rate was 70%. Sixty English language test papers set by the teachers of English in the sample for Grades 6, 7, and 8 for the second term test in the year 2016 were collected. A content analysis of the papers was conducted for weight allocation for each language skill, type of activities, allocation of marks for individual activities, and to study the extent to which the test papers have met good principles of assessment.

Twenty teachers who were involved in setting the selected test papers were interviewed. The aim of the semi-structured interviews was to identify the problems teachers face when setting, administering, and marking English language tests, the practices they use when designing English language tests and to gauge their knowledge about assessment principles. The interviews were transcribed and coded for problems they face when setting, administering, and scoring assessments, for practices they adopt when setting, administering, and scoring and for evidence of their knowledge of principles of assessment.

31.5 Findings

The teachers in the sample varied in their qualifications, experience, and prior training on testing and assessment. Nearly 49% of the teachers had received some training in testing, and they displayed substantial knowledge of the principles of assessment, while others showed a lack of basic knowledge of testing and assessment (see Table 31.1). Appendix 31.3 shows the analysis of responses to the closed-ended items in the questionnaire. The open-ended questions in the questionnaire and the interview data revealed the problems faced by teachers when setting, administering,

Table 31.1 Profile of the teachers who participated in the study

Teachers	Gender	Qualifications	Years of experience	Type of school	Training in testing and assessment
105	Female-74 Male-31	Masters-11 Graduates-35 National Diploma in Teaching English-41 Other Diplomas-3 English trained-14 A/L qualified-1	1–4 years-36 5–9 years-41 Over 10 years-28	Government-55 Private-23 International-37	No training received-33 Some training received-51 Adequate Training received-21

and marking tests, and these data also showed the practices adopted by teachers when doing the above activities.

31.5.1 Problems Faced by Teachers

31.5.1.1 Problems Faced When Setting Assessments/Tests

Almost all the teachers in the sample, irrespective of their qualifications, training, and experience, reported that setting test papers is a difficult task and marking them and releasing marks within a limited time frame is extremely challenging. Some teachers claimed that setting a test paper to parallel grades where the students are taught by several teachers is a problem.

> Some teachers in parallel grades have not taught all the lessons… We find it hard to set a common paper to measure the intended competencies. (T1)

Another problem they mentioned was the difficulty in setting papers for classes with mixed-ability students. Balancing the test items to challenge good students while not discouraging the weaker ones is a challenge for them. Some teachers confessed that they have no knowledge in testing techniques and they just recycle the questions that appeared in the past papers.

Experienced teachers elaborated on the problems which were mainly related to setting and marking. These teachers displayed their knowledge in terminology in testing and assessment when discussing their problems.

> Content validity of the test is less if it does not contain a representative sample of the syllabus. As the time allocated to conduct classroom tests is limited, we cannot include enough samples to test students' language skills. (T3)

The interviews also revealed that the majority of the teachers had not included test items to test speaking and listening skills when they set language tests.

Teachers who teach in international schools reported difficulty in setting papers to meet the needs of local students. Their students struggle with unfamiliar vocabulary and socio-cultural settings that appear in reading passages, and they attempt to address this issue by using texts with familiar content when setting papers locally. However, they are worried about the negative consequences of not preparing students for foreign examinations.

The monopoly in setting tests in some schools was also revealed. One experienced teacher (T15) from a government school stated that a single teacher in her school sets all the papers and she never gets an opportunity to contribute to setting tests.

31.5.1.2 Problems Faced When Administering Tests

All the teachers who were interviewed, mentioned the problem of non-availability of spacious classrooms for administering tests. "The number of students in a class is above 45 and arranging classrooms for tests is a daunting task," one teacher said. Cheating is found to be a problem for many of the teachers in the study. Many were of the view that Multiple Choice, True or False, and Gap-filling type questions are the most copied test items. Certain measures adopted by them to avoid cheating were also presented by the teachers in this study. One strategy adopted was to put two grades in one hall and allow students from the same grade to sit in every other row, but conducting listening tests in that setting was found to be problematic.

Lack of physical resources, i.e., bad lighting, noise in the surrounding classrooms, were also mentioned as factors that cause problems when administering English language tests.

31.5.1.3 Problems Faced When Scoring Tests

The scoring of essays seemed a challenge to many teachers. The teachers who had training in testing mentioned scorer subjectivity and inconsistency as problems when marking essays. They spoke about how the mistakes in test papers become problematic when allocating marks. Their struggle when marking keys are not provided by the setter was also a problem stated by the teachers.

Scoring of picture descriptions is a challenge for them since some students write simple and correct sentences while others try to be more creative and make mistakes attempting to do so.

Another teacher reported that she faces problems when awarding marks for productive skills as she was uncertain whether to give more marks for content or for language.

Many teachers complained that no marking could be done during school hours and that has badly affected their personal lives.

31.5.2 Practices in Setting and Scoring Tests

Most of the teachers confessed that they limit the number of writing tasks to one, as scoring is time consuming; some stated that they avoid essay writing if possible. In schools where teachers are not highly qualified, teachers work as a team when designing tests. While this collaborative setting is done in some schools, a single setter takes the responsibility in other schools. The sources of texts for reading comprehension were mainly the internet and past papers. An international school teacher revealed that their promotions are based on students' performance and as a result, the marking becomes subjective. According to her, some teachers give pass marks to students even if their performance is poor, to avoid criticism from the board of management.

31.5.3 Assessment Literacy of Teachers as Revealed by the Tests Constructed by the Selected Sample

The content analysis done on sixty English language test papers set by the teachers in the sample showed wide variations in their assessment literacy. The assessments conducted by most of the teachers did not conform to basic assessment principles. The papers were analyzed for the skills tested, weighting given to each skill, how far the test items matched with the objectives of the syllabus, and to see the extent to which the test had met the principles of assessment.

31.5.3.1 Weighting Given for Each Skill

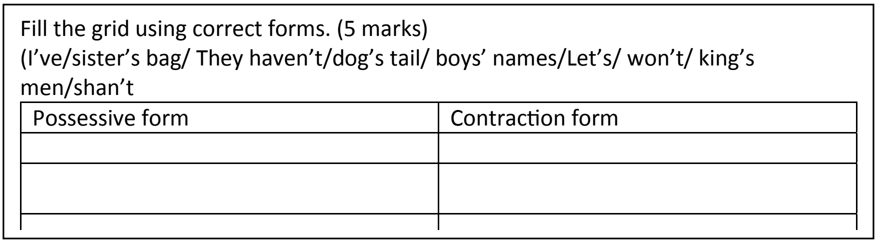
It was observed that 80% of the papers had been designed to assess Reading and Writing skills only, and Speaking and Listening skills were neglected to a great extent. Some test papers tested the speaking ability indirectly via dialogue completion tasks. Language components of grammar and vocabulary were assessed in all the test papers either as discrete-point test items or as integrative test items. Hence, there was a mismatch between what is in the syllabus and what is assessed in teacher-made tests. Another finding of the content analysis was that even though writing is tested in some form in these tests, the weighting given to writing when compared to that of Reading, Grammar, and Vocabulary is very low.

31.5.3.2 How Far the Test Items Matched with the Objectives of the Syllabus

The objectives stated in the Grade 6, 7, and 8 syllabi are to develop four language skills and other language competencies necessary for effective communication. However, the tests designed by many teachers did not meet the objectives of the syllabus. The papers consist of many questions on grammar, and those items test grammar in isolation without testing their use in communicative situations. This issue was observed in many of the test papers designed by the teachers in the sample (see Examples 1 and 2) (Figs. 31.1 and 31.2).

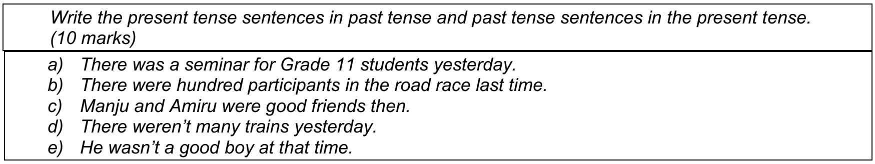
31.5.3.3 How Far the Tests Have Met the Principles of Assessment

Many test papers set by teachers showed that the principles of assessment, mainly validity and reliability had not been considered when setting papers. Many tests set by the teachers in the sample had many items to assess grammar, and less priority had been given to other language skills, and the principle of content validity had been violated. Some teachers had not included essay writing activities, as they thought that scoring essays is time consuming.


Fill the grid using correct forms. (5 marks)
(I've/sister's bag/ They haven't/dog's tail/ boys' names/Let's/ won't/ king's men/shan't

Possessive form	Contraction form

Fig. 31.1 Example 1—Grade 8, private school

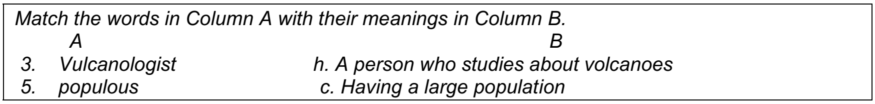

Write the present tense sentences in past tense and past tense sentences in the present tense. (10 marks)

a) There was a seminar for Grade 11 students yesterday.
b) There were hundred participants in the road race last time.
c) Manju and Amiru were good friends then.
d) There weren't many trains yesterday.
e) He wasn't a good boy at that time.

Fig. 31.2 Example 2—Grade 7

In Example 2, the instructions are inappropriate, and the test takers may get confused as all the sentences given are in the past tense. This affects the reliability of the test. Another test item showed ambiguity as many answers were possible for some items which may affect the reliability of the test items. Some items set by teachers gave clues to the correct answer (see Examples 3 and 4). Since the answers could be guessed easily, the validity of the above test items would be affected (Figs. 31.3 and 31.4).

It was observed that some setters had paid attention to competencies stated in the curriculum but had failed to meet the assessment principle of construct validity. The main teaching approach used in the second language classroom is the Communicative Approach. However, the setters have used discrete-point testing when testing the knowledge of vocabulary, as shown in Example 5 (Fig. 31.5).


Match the words in Column A with their meanings in Column B.

	A	*B*
3.	*Vulcanologist*	*h. A person who studies about volcanoes*
5.	*populous*	*c. Having a large population*

Fig. 31.3 Example 3

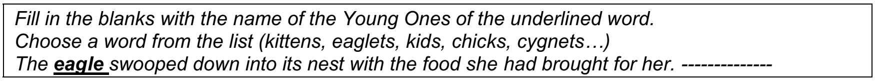
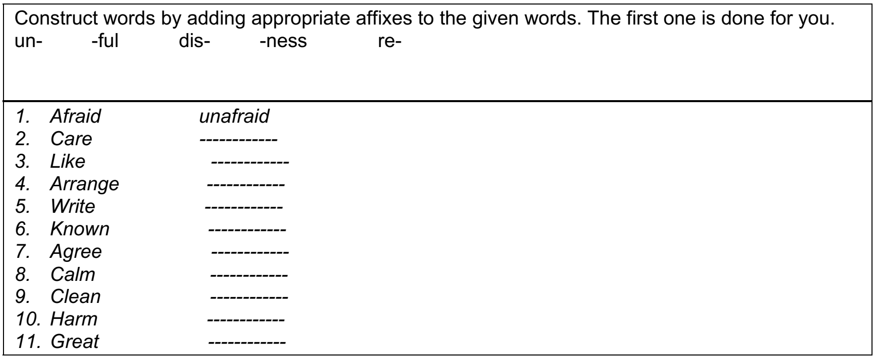

Fill in the blanks with the name of the Young Ones of the underlined word.
Choose a word from the list (kittens, eaglets, kids, chicks, cygnets…)
*The **eagle** swooped down into its nest with the food she had brought for her. --------------*

Fig. 31.4 Example 4



Construct words by adding appropriate affixes to the given words. The first one is done for you.
un- -ful dis- -ness re-

1. *Afraid* *unafraid*
2. *Care* ------------
3. *Like* ------------
4. *Arrange* ------------
5. *Write* ------------
6. *Known* ------------
7. *Agree* ------------
8. *Calm* ------------
9. *Clean* ------------
10. *Harm* ------------
11. *Great* ------------

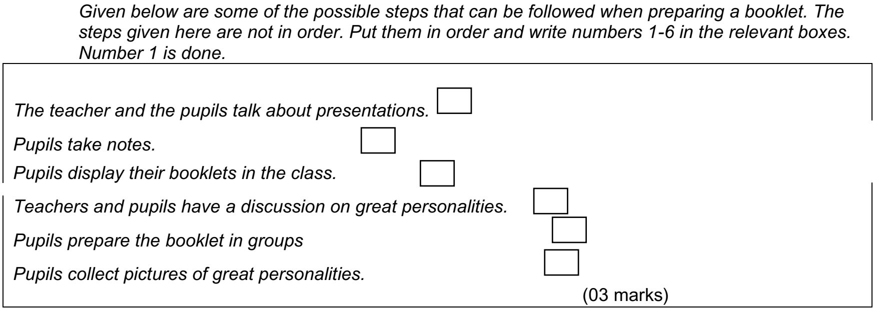
Fig. 31.5 Example 5

In order to increase the construct validity of this item and to make it more communicative, it is necessary to improve this question by including these words in sentences or in a paragraph and allow students to guess the meaning from context and add affixes accordingly. This will also have a beneficial backwash on students as they will not just memorize the words but learn to use them meaningfully.

Another task type which needs attention is the re-ordering task activity. The following re-ordering task appeared in a test paper set by a teacher (see Example 6) (Fig. 31.6).

Given below are some of the possible steps that can be followed when preparing a booklet. The steps given here are not in order. Put them in order and write numbers 1–6 in the relevant boxes. Number 1 is done.

Even though the instructions given state “Number 1 is done,” no box is marked. The test takers may have found it difficult to select the first sentence as more than one answer is possible. The instructions are inappropriate, and the marks allocated are inadequate. Reliability, validity, and other principles are violated in the test.



Given below are some of the possible steps that can be followed when preparing a booklet. The steps given here are not in order. Put them in order and write numbers 1-6 in the relevant boxes. Number 1 is done.

The teacher and the pupils talk about presentations. □

Pupils take notes. □

Pupils display their booklets in the class. □

Teachers and pupils have a discussion on great personalities. □

Pupils prepare the booklet in groups □

Pupils collect pictures of great personalities. □

(03 marks)

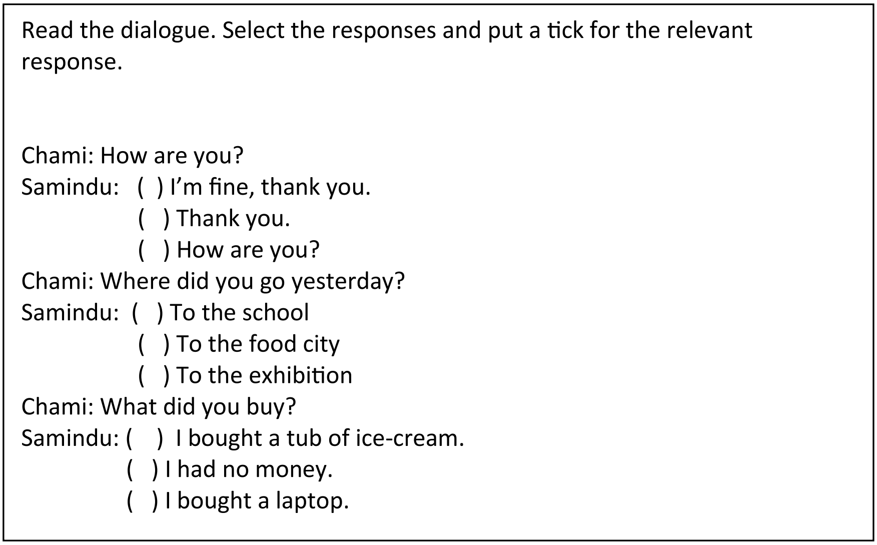
Fig. 31.6 Example 6—Grade 8, government school

Teachers showed poor knowledge in designing Multiple Choice Questions as well. The following item set by a teacher shows that there is ambiguity in most of the test items affecting the reliability of the test (see Example 7) (Fig. 31.7).

Some teachers displayed their lack of knowledge in assessment principles as some essay topics were found to be inappropriate for Sri Lankan students. This may have affected the face validity of the test. The topics given for selection assess different skills as some topics ask for the writer's point of view and their argumentative essay writing skills while another topic given assesses their creative writing skills, and this can affect the consistency of scoring (see Examples 8 and 9) (Figs. 31.8 and 31.9).

The above test items show that some setters lack basic literacy in assessing writing. The assessments designed by teachers varied in the number of tasks given, their difficulty level, mark allocation, and the consistency in marking. Grade level competencies have not been considered by some.

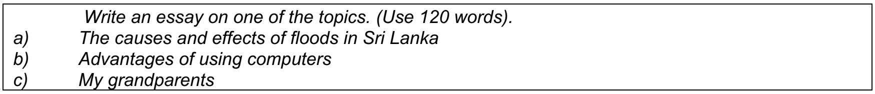
However, some teachers who have had training in testing and assessment had been careful not to violate the assessment principles as far as possible. They displayed their assessment literacy in the test papers designed by them.



Read the dialogue. Select the responses and put a tick for the relevant response.

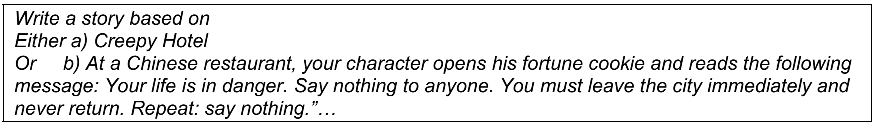
Chami: How are you?
Samindu: ( ) I'm fine, thank you.
( ) Thank you.
( ) How are you?
Chami: Where did you go yesterday?
Samindu: ( ) To the school
( ) To the food city
( ) To the exhibition
Chami: What did you buy?
Samindu: ( ) I bought a tub of ice-cream.
( ) I had no money.
( ) I bought a laptop.

Fig. 31.7 Example 7—Grade 8



Write an essay on one of the topics. (Use 120 words).
a) The causes and effects of floods in Sri Lanka
b) Advantages of using computers
c) My grandparents

Fig. 31.8 Example 8—Grade 7—Mid-term test



Write a story based on
Either a) Creepy Hotel
Or b) At a Chinese restaurant, your character opens his fortune cookie and reads the following message: Your life is in danger. Say nothing to anyone. You must leave the city immediately and never return. Repeat: say nothing."…

Fig. 31.9 Example 9—Grade 8—International school

31.6 Insights Gained

The research showed that the problems faced by secondary school teachers in the sample when setting, administering, and scoring tests varied considerably. The practices teachers adopted when engaging in the above tasks also varied according to their level of experience and training received. Some were satisfied with the on-the-job training they had received while many teachers were interested in developing their assessment literacy.

Hence, it is important to plan and conduct comprehensive teacher-training programs which would provide adequate knowledge in the principles of assessment plus hands-on experience in designing and scoring assessments.

Prior to any training program, it is necessary to study teachers' test design practices, problems faced by them, and their assessment literacy. It is also important that education authorities identify competent setters and moderators in each education zone and strengthen testing mechanisms in the school system by providing necessary facilities for administering tests. Since teachers have difficulties in attending face-to-face training sessions, it would be beneficial if online or distance education courses in testing and assessment were offered by experts in the field.

31.7 Conclusion: Implications for Test Users

The study collected data on the problems faced by teachers when designing, administering, and scoring tests in secondary schools in Sri Lanka. It also attempted to collect data on teachers' assessment practices and about their assessment literacy. Assessment literacy has been investigated using questionnaires with closed-response items and constructed responses in previous studies (Fulcher 2012). The present study, in addition to a questionnaire which collected demographics of the sample and their background in assessment, used semi-structured interviews and teacher-made tests to determine the teachers' assessment literacy. Even though some teachers claimed that they possessed adequate knowledge and experience in assessment when responding to the questionnaire, this knowledge was not evident in the test papers designed by them. Hence, the present research reveals the importance of using other instruments like content analysis of teacher-made assessments as measures of teachers' assessment literacy, which would display their true assessment literacy and its application in real situations.

Appendix 31.1

Questionnaire for Teachers
Part A

Name (Optional):
Age:
Gender:
Current school:
Type of school:
Years of experience as a teacher:
Grades taught:
Highest Qualifications:
Training received in test construction: Yes/No
If Yes, give details:

Part B

i. Please read each item carefully and select the response you think is the best one by putting a tick (√) in the appropriate box.

	Strongly Agree	Agree	Disagree	Strongly Disagree
1. I have a good knowledge of principles of assessment				
2. I have received adequate training in language testing and assessment				
3. I use a variety of testing techniques in my assessments				
4. Setting tests is a very difficult task for me				
5. I get my tests moderated by a senior teacher				
6. I use a scoring rubric when assessing essays				
7. My tests assess all four language skills, grammar, and vocabulary				
8. I include items which help me to discriminate students well				
9. I prepare test specifications before setting a test				
10. I edit and proof-read my test papers until they are error-free				

ii. What are the challenges you face when setting, administering, and scoring assessments?
iii. What practices do you adopt when setting, administering, and scoring assessments?

Appendix 31.2

Interview Schedule

1. Can you tell me about your experience in testing and assessment?
2. What are the challenges you face when setting tests?
3. What practices do you adopt when setting tests?
4. What challenges do you face in administering and scoring tests?
5. How do you overcome these challenges?
6. Would you like to have further training in testing and assessment?
7. What are the areas you need training in?

Appendix 31.3

Responses to the Closed-ended Items in the Questionnaire

	SA%	A%	D%	SD%
1. I have a good knowledge of principles of assessment	9.52	52.38	27.62	10.47
2. I have received adequate training in language testing and assessment	16.19	48.57	23.81	17.14
3. I use a variety of testing techniques in my assessments	57.14	42.85	00.00	00.00
4. Setting tests is a very difficult task for me	19.05	60.95	20.00	00.00
5. I get my tests moderated by a senior teacher	5.71	36.19	58.09	00.00
6. I use a scoring rubric when assessing essays	24.76	21.90	38.09	15.24
7. My tests assess all four language skills, grammar, and vocabulary	46.66	36.19	17.14	00.00
8. I include items which help me to discriminate students well	23.81	57.14	19.05	00.00
9. I prepare test specifications before setting a test	4.76	11.43	83.81	00.00
10. I edit and proof-read my test papers until they are error-free	72.38	27.62	00.00	00.00

SA = Strongly Agree, A = Agree, D = Disagree, SD = Strongly Disagree

References

Brown, H. D., & Lee, H. (2015). *Teaching by principles: An interactive approach.* New York: Pearson Education.

Coombe, C., Al-Hamly, M., & Troudi, S. (2009). Foreign and second language teacher assessment literacy: Issues, challenges and recommendations. *Research Notes, 38,* 14–18.

Djoub, Z. (2017). Assessment literacy: Beyond teacher practice. In R. Al-Mahrooqi, C. Coombe, F. Al-Maamari, & V. Thakur (Eds.), *Revisiting EFL assessment, second language learning and teaching* (pp. 9–27). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-32601-6_2.

Elsshawa, N., Heng, C. S., Abdullah, A. N., & Rashid, S. (2016). Teachers' assessment literacy and washback effect of assessment. *International Journal of Applied Linguistics and English Literature, 5*(4). http://hdl.handle.net/20.500.12358/26637.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041.

Green, A. (2014). *Exploring language assessment and testing: Language in action.* New York: Routledge.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162–189). Cambridge: Cambridge University Press.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29,* 21–36.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402.

Webb, V. (2002). *Language in South Africa: The role of language in national transformation, reconstruction, and development*. Amsterdam: John Benjamins.

Chapter 32
English Language Testing Practices at the Secondary Level: A Case Study from Bangladesh

Arifa Rahman and Rubina Khan

Abstract This chapter presents a research study that investigates the testing practices in English language secondary schools in Bangladesh and their impact on test takers and on education in general. It focuses on the test specifications and test items currently in use and the extent to which the testing practices are aligned to the curriculum objectives. The findings indicate that syllabus content is confused with the curriculum objectives and validity and reliability requirements of test items are flouted. As tests are not aligned to the objectives of the curriculum and valid testing principles are not applied to test design, teachers and learners find alternative ways of coping with the tests, thus spawning practices of teaching to the test, often relying on private tutoring and rote learning. The insights drawn from this study discuss the significance of the washback effect and the impact of testing and the crucial need to develop *assessment literacy* among teachers, test-setters, and scorers. Assessment literacy would facilitate test designers/users to recognize the consequences of flawed practices that are detrimental to teaching and learning.

32.1 Introduction: Purpose and Testing Context

Bangladesh, situated in South Asia, is the world's eighth most densely populated nation occupying an area smaller than Great Britain. With the fall of the British Empire in India in 1947, the region became part of the state of Pakistan from which it eventually seceded in 1971, with language rights playing a major role in its struggle for an autonomous state.

With its history of British colonialism, English has been part of the secondary education system of this region, alongside the vernacular language, since the early twentieth century. After independence, however, driven by a heightened sense of

A. Rahman (✉) · R. Khan
University of Dhaka, Dhaka, Bangladesh
e-mail: arifa73@yahoo.com

R. Khan
e-mail: rkhan@agni.com

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_32

identity and nationalism, *Bangla*, the first language of 98% of the population, became the official language and took precedence in education, bureaucracy, the media, and in a variety of domains. Schools and tertiary institutions were ordered to use *Bangla* as the medium of instruction but interestingly enough, although English was marginalized, it remained as a mandatory subject in the school curriculum.

From the early 90s, however, and within a matter of two decades, this initial displacement of English gradually gave way to its expanding role in the education sector (Rahman 2007). This is a reflection of the trend in several Asian nations of updating their language and language-in-education policies (Tsui and Tollefson 2007). Bangladesh brought about a series of language-in-education planning directives that focused primarily on giving more space to English in the curriculum. The changes in language policy in general and English language education policy in particular during the last two decades may be perceived as a neoliberal narrative in a globalized world (Hamid and Rahman 2019).

The country follows the traditional three-tier education system (primary, secondary, tertiary). With English as a mandatory subject, learners get 12 years of schooling with roughly 1800 contact hours of class time. There are two mandatory public examinations at the secondary level, at the end of years 10 and 12, respectively. Administered through eight examination boards nation-wide, both examinations are considered high-stakes as they perform a gate-keeping function for university entrance and career paths (Khan 2010). Thus, these two public examinations wield tremendous power over the education system and affect the teaching and learning that takes place in school. It affects not only individuals, institutions, and systems but also society in general.

This chapter investigates the testing practices in English language learning at the secondary level in Bangladesh. The main purpose is to see whether school and institutional testing practices reflect the objectives of the curriculum. Studies have shown that there is usually a "mismatch" between the intended English language learning objectives and current assessment practices (English in Action Research Report 2009; Das et al. 2014). Our study will investigate this further in the Bangladesh secondary school context.

The study thus attempts to address the following research questions:

1. What are the English language testing practices that exist at the secondary level in Bangladesh?
2. To what extent is there an alignment between curriculum objectives and testing practices?

32.2 Testing Problem Encountered

The testing problems encountered may be categorized under three issues relevant to our study. The first is the professed curriculum objectives at the secondary level and the extent to which they are met through current test practices in secondary education. The second deals with the existing Secondary School Certificate (SSC) test specifications, question setting and scoring while the third, which is over-arching, is the failure to consolidate the various English Language Education (ELE) reforms into teaching and assessment practices.

With regard to the first issue, according to the 2012 National Curriculum, the declared objectives for English Classes 9–10 are to:

- acquire competence in all four language skills;
- use the competence for effective communication in real-life situations;
- acquire necessary grammar competence in English;
- develop creativity and critical thinking through English;
- become independent learners of English;
- use language skills for utilizing information technology;
- be skilled human resources by using English language skills.

In spite of these admirable objectives on paper, the ground reality is unfortunately quite different. The Examination Board exam results show that failure rates in English are higher while the pass scores are generally lower than in other subjects. 2017 and 2018 SSC examination results showed a fall in English grades (and in Math) in all eight Board examinations (Islam 2018), thus bringing the pass rate to drop to a minimum in nine years. The percentage of failure in English at the Higher Secondary Certificate (HSC) Examination of the Dhaka Board from 2013 to 2017 revealed a similar trend.

Additionally, studies have shown that learning outcomes are rather dismal and learner competence in general remains poor, as mentioned earlier, and disconnected to real-life needs (Ahmed et al. 2006; Rahman 2007; Hamid and Baldauf 2008; English in Action Report 2009; Hamid and Erling 2016). Poor language competence is not limited to students alone; in fact, secondary school teachers of English appear to have the same limitations. The Baseline Survey of Secondary School English Teaching and Learning (1990) had found English teachers' language proficiency far below the required level. Nearly two decades later, the English in Action Baseline Study (2009) yielded similar findings. So when teaching is undertaken in such conditions, learning outcomes naturally tend to be inadequate.

In terms of the second issue related to SSC test specifications, question setting and scoring, we will cover this in detail in Sect. 32.5, but it may be stated here that there are problems with the nature and validity of the questions, the marking/scoring system and their impact on the results.

A third of the testing problems encountered at the secondary level is the failure to consolidate various ELE reforms into teaching and assessment practices (Imam 2005; Hamid and Baldauf 2008; Rahman 2015). For example, the English Language Teaching Improvement Project (ELTIP) was initiated in 2000 to develop communicative language teaching through CLT-based textbooks for classes 9–12, to train teachers in CLT, and to devise an appropriate assessment system that would test real-life language skills. The last objective, however, was rendered futile as the secondary examination boards proved to be citadels of resistance (Quader 2001). As a result, the public tests maintained their traditional mode with their focus on grammar-based questions until the National Curriculum 2012 brought about a number of changes. Even this has not had a desirable impact, as our current study indicates.

Since 2001 to 2017, successions of donor-aided ELT projects were implemented. The largest to date was *English in Action* (EIA), a nine-year program with primary, secondary, and adult language support initiatives which ended in 2017. In addition, two significant policy documents were introduced. First, the Bangladesh National Education Policy (2010) advocated modern, quality, suitable education for all. Delivery strategies were proposed through the development of a learning environment, methodology, and materials that would be attractive and enjoyable. Second, the Report of the Ministry of Education Advisory Committee for the Development of English (2010) clearly recognized that the weakness of the current tests lay in the question setting in the public examination as it was not always in accordance with the curriculum. The Report recommended reforming the examination system but without actually laying down a framework for action. Teachers at primary and secondary levels needed professional training for which proposals were made to revise current training programs to be more purposeful and effective.

However, it has been repeatedly seen that proposals made at the macro-level of educational policy depend for their effectiveness on the interpretation by teachers at the micro-level of pedagogic practice and on their abilities and commitment to understand and carry out these proposals (Hargreaves and Fullan 1992). Not surprisingly, matters did not move in the right direction. More recent studies have shown that the learning and teaching of English in Bangladesh is still quite inadequate (Akteruzzaman and Islam 2017).

The language assessment landscape of Bangladesh is not much different from other countries where English language teaching/learning reforms have been proposed. Studies have highlighted the difficulties of introducing a new assessment approach, which includes resistance from different stakeholders (Hasan 2004; Khan 2010). Hasan (2004), through an extensive document analysis of test formats and question papers of all secondary education boards, states that despite efforts to revise the question papers, there is a significant gap between what is intended to be taught and what is measured. Although the teaching and testing of speaking and listening skills are clearly stated in the curriculum, in reality, there is far less balanced emphasis on the development of the four language skills than required.

To sum up, the studies reviewed above suggest that ELE reforms have met with challenges despite their worthwhile intentions. Most of these challenges appear to result from policies being implemented by following a somewhat top-down approach where some key stakeholders, namely teachers, at least at the initial stages, were not involved. English language assessment reforms seem to have encountered more problems than the ELE reform itself due to the difficulty in carrying out communicative language assessment and the lack of planned effort to prepare teachers for the reform.

32.3 Review of the Literature

In general, “assessment” refers to the broad area of monitoring or taking stock of the performance of students or programs while “testing” refers to a set of specified uniform tasks to be performed by students, these tasks being an appropriate sample from the knowledge or skills in a broader field of content (Cumming 2009). In this study, we use the terms *testing* and *assessment* interchangeably to refer to testing. Tests are thus used as a yardstick and an indicator of a person’s ability to perform in a particular area.

Although a variety of issues are related to testing and assessment including testing practice, theory, ethics, and philosophy (Fulcher and Davidson 2007), we will be looking at the literature related to three issues as relevant to this chapter. They are:

- The guiding principles of good tests.
- Summative and formative assessment.
- Washback and test impact.

32.3.1 The Guiding Principles of Good Tests

The guiding principles that govern good test design are *validity, reliability,* and *practicality* (Hughes 2013). *Validity* refers to the extent to which a test measures what it says it measures. In early definitions, validity related to content, criterion, and construct but since the mid-nineties, the issue of construct validity has been critically debated (Messick 1996) that led to a more complex interpretation of constructs. *Reliability* refers to the consistency of test scores which means the test would give the same results if it were administered at another time or in another setting. Factors that promote test reliability are consistency in test format, content of questions, and the length of the examination. *Practicality* includes cost, time, and the resources needed, the ease of marking and the availability of trained markers. Practicality also points to teachers’ ability to develop, administer, and mark a test within the available time and having resources that are needed. It further emphasizes the need for students to get comprehensible feedback in the quickest possible time.

In addition, Bachman and Palmer (1996), with the advent of communicative language teaching, advocate *authenticity* and *usefulness* while Coombe and Hubley (2009) include *transparency* in the list of key assessment principles. *Authenticity* is regarded as a critical quality of language tests although this quality has not been discussed in many of the early works on testing. Bachman and Palmer (1996, p. 23) define authenticity as the “degree of correspondence of the characteristics of a given language test task to the features of a target language use (TLU) task.” The principle of *usefulness* refers to the purpose for which the test has been intended and requires tests to be developed with a specific purpose, a particular group of test takers, and a specific language use in mind. Finally, *transparency* means the availability of clear, accurate information to all stakeholders. Test procedures need to be transparent.

In particular students need to have clear information about outcomes to be evaluated; formats; weighting of items; allotment of time; marks allocation; and grading criteria. Thus, the test should not come as a surprise to test takers.

The above is an amalgam of the guiding principles of tests. Few educational systems can claim that they fulfill the principles in their entirety. Hence it is not surprising that in a developing country like Bangladesh, there are a number of constraints to good test practice. Here, the education system is often rendered dysfunctional due to poor resources, inadequate strategies for meeting declared objectives, inequalities in support systems, and short-sighted top-down language policies (Rahman 2015).

32.3.2 Summative and Formative Assessment

Although there are various types of tests defined in the literature (see Alderson et al. 1995), our focus in this study is mostly on formal "achievement tests" which are syllabus-bound and measure the learning that is expected to have taken place. It takes the form of a formal examination administered at the end of a school year or a semester, and in terms of assessment, it equates to *summative* assessment. In Bangladesh as in many countries, the achievement test is not necessarily prepared by the classroom teacher and in the case of public tests, they are externally constructed, administered, and scored by an institution or ministry department.

Summative assessment is thus conducted at the end of a program of study to assess whether and how far individuals or groups have been successful. In Bangladesh, terminal examinations are the main features of summative assessment. Most secondary schools conduct three terminal examinations. *Formative assessment,* however, is carried out during the learning process as an intervention that is designed to encourage further learning and change (Fulcher and Davidson 2007). Informal formative assessment is ongoing and the information received from such assessment is usually used as a basis for further classroom work and learner development.

Both the assessment types may be equated to Earl's (2003) *three types of assessment* in the learning scale. Summative assessment is the "assessment *of* learning" whereas formative ongoing classroom assessment aims at "assessment *for* learning" and sometimes also "assessment *as* learning" when learners are actively engaged in critically assessing themselves in an interactive form of learning.

32.3.3 Washback and Test Impact

High-stakes tests may have impact on the content taught, the methodology used and on the attitudes toward the values of educational objectives. It may also include the effect of the test on classroom practices. Due to the highly essential role of testing in students' learning, language testing professionals have directed attention to the influence of testing on teaching, learning, educational outcomes, and individuals. For many years it was asserted that language tests had a negative impact on teaching and thereby on learning/learners.

This phenomenon, known as *washback* (Alderson and Wall 1993), has been much discussed. Messick (1996, p. 243) has stated that washback is "the extent to which the test influences teachers and learners to do things they would not otherwise necessarily do." Cheng (2004) claims washback is found more in classroom practices of teachers and learners and in the content teachers choose to use in their teaching. Bailey (1996) in her seminal article which reviewed the washback concept, advocated a deeper understanding of the nature of washback, the various ways washback works, and the importance of the ways to research washback.

Washback is generally considered as being either negative or positive (Wall 1997). Negative washback is said to occur when a test content or format is based on a narrow definition of language ability, and so constrains the teaching/learning context. It can be positive when test items are aligned to the objectives of the course/curriculum and encourage teaching and learning to move toward that goal. The possibility of this sort of *positive* backwash was first discussed and promoted more than two decades ago by Bailey (1996). Cheng (2005) too has advocated changing language teaching through appropriate language testing.

The washback concept also operates within the notion of *Test Impact*. Testing researchers have clearly distinguished between the two. Test impact is defined as the effect tests may have "on individuals, policies or practices within the classroom, the school, the educational system or society as a whole" (Wall 1997, p. 291). Impact thus, is used to describe the wider influences of tests and their consequences far beyond the classroom. Bachman and Palmer (1996) presented a similar definition and viewed washback as one dimension of test impact. Cheng and Curtis (2004) outline the connection between the notion of washback and issues of curriculum alignment and examine the potential influence of washback on systemic validity when there is a mismatch between what is taught and what is tested, thus confirming the negative aspect of washback.

32.4 Methodology

In light of the discussion on the prevalent testing situation and the literature survey in the above sections, we now turn to the modality of our study. In order to address the issues raised in the research questions (see Sect. 32.1), the study used a mixed methods approach, which is detailed in the following sections.

32.4.1 Data Collection

In order to investigate these dual aims, a convenience sampling procedure based on respondent consent, availability, and access to institution was adopted to collect data. The following methods of data collection were used:

- **Document Analysis** This source of secondary data was culled from the SSC English Curriculum 2012 and the SSC English test papers of three different Examination Boards to identify and analyze examination components, test items, and mark allocation.
- **Questionnaires** with closed and open-ended questions were distributed to teachers, question-setters, markers, parents, to gauge respondents' views on the issues subsumed under the study objectives.
- **Semi-structured interviews** were conducted, sometimes as follow-up to questionnaires to get further in-depth views.
- **Focus Group Discussions (FGDs**) with education department officials and administrators and with students were undertaken.
- **Classroom Observation** was conducted in 15 schools and 15 classes were observed (with 9 female and 6 male teachers). A range of variables in terms of teacher profile was carefully pursued to ensure a distribution in gender, age, teacher training, and teaching experience. In addition, public/private schools, geographical location, single sex and co-ed settings, and variable class size ranging from 50 to 89 were also considered variables. This procedure was used primarily to triangulate the data obtained from questionnaires and interviews. As Weir and Roberts (1994) maintain, observation gives direct data about classroom events on the reality of implementation. They can be used to measure how much a particular program objective has been met and to gauge participants' expressed perceptions and beliefs.

The Respondents were the following:

- 87 English teachers (49 female, 38 male) of classes 9, 10 from 30 schools within Dhaka city and the suburbs. All teachers had a master's degree in English and had teaching experience from three to 15+ years. Only 14% had any sort of teacher training.
- 142 students (86 female, 56 male) of classes 9, 10; age group 15–17, from science, humanities, and business streams.
- Six test officials and administrators. Being too busy, they only agreed to have joint discussions which we prefer to call FGDs. Two FGDs were held, one with four members of the National Curriculum and Textbook Board (NCTB), and the other, with two members of Bangladesh Examination Unit (BEDU), Ministry of Education.
- 25 parents of students of classes 9, 10, all mothers with the exception of one father responded to questionnaires and were interviewed.
- Question-setters and Markers: 20 + 20 (male 55%, female 45%). Not all markers were question-setters and not all question-setters were markers.

32.4.2 Data Analysis

All the data obtained from the above procedures were collated and analyzed by teasing out categories (Cohen et al. 2007).

32.5 Findings

SSC English Examination Document Analysis has been carried out from the 2012 National Curriculum English Document which brought in major changes to the previous 1995 curriculum, to align itself to the fast changing needs of a globalized world. It was developed based on the National Education Policy 2010 which professed to open up a gateway to life skills for learners. The curriculum includes contents such as information and communications technology, work and life-oriented education, career education, climate change and responsibilities, adolescence and reproductive health, and women development. It also upholds the ideals, values, and inspirations of our Liberation War. In addition, changes were proposed in teaching-learning activities, and in the ways of assessment. It emphasized experiential learning, i.e., by doing, instead of rote learning, to develop personal and work skills.

The curriculum for Classes 9–10 is laid out in a matrix separately for Paper One and Paper Two. Language Outcomes, Functions and Language points and themes have been identified. Learning outcomes are expressed in terms of the four language skills. Therefore, the teaching-learning activities are supposed to be based on listening, speaking, reading, and writing where teachers' and students' activities are detailed.

32.5.1 Formal Assessment

The SSC English examination has two papers: First Paper and Second paper each comprising 100 marks. The final numerical scores are transferred into letter grades.

The First Paper comprises reading and writing sections of 50 marks each. The reading section comprises multiple choice questions, open-ended questions, information transfer items, gap-filling items, matching items, rearranging sentences to make a story and a summary.

The writing section comprises paragraph writing, story completion, informal letter writing, describing a chart/graph, and dialogue writing.

The Second Paper allocates 45 marks for grammar and 55 marks for composition. The grammar section has test items that include substitution tables, correct verb forms, narration, completing sentences, use of suffix and prefix, tag questions and connectors. The composition section comprises 55 marks and includes test items like writing a CV with a cover letter, informal/formal letters, story completion, short paragraphs, dialogue writing, graph and chart analysis, and composition on familiar topics, e.g., experiences, problems, and events.

The National Curriculum (2012, p. 85) states, "Students' learning activities will be assessed through classwork, continuous assessment, terminal/public exams using teacher prepared or centrally prepared tools. It is to note that test tools will be based on all the learning domains where necessary." How far this is implemented in practice was the aim of our study.

The stated objective of the SSC syllabus is to build communicative competence of learners but an analysis of the test items reveals there is very little fit between curriculum objectives and the test contents. The principles of communicative language testing are not followed. The communicative principle emphasizes the use of authentic language. Bachman and Palmer (1996) advocate that a test is regarded as authentic if (a) the language in the test is natural, (b) items are contextualized, (c) topics are interesting, and (d) tasks represent real-life tasks. Only a couple of test items make a feeble attempt to cover communicative tasks. Most questions test knowledge of form and content. Students' and teachers' comments also reveal that the testing system is still confined within the narrow boundaries of content-based items.

Choudhury and Holbrook (2018) carried out an in-depth impact study of 2017 SSC examination questions on assessment across three Examination Boards (out of the eight in operation nation-wide). They too reported on a number of unsatisfactory practices of the Examinations boards. Among them were:

Nature and validity of the questions: The questions were trivial in nature, were poorly constructed, and often the language used was poor. Question-setters were not meaningfully testing candidates' ability to read or write. Questions were not checked for suitability. Comprehension passages were poorly edited and the reading comprehension questions were too simple and straightforward, to the extent that candidates were able to answer questions without meaningfully reading the passage.

The marking/scoring system: There was no standardized marking scheme. It was not clear how markers scored and there were indications of impressionistic marking. Scorers had little or no training in marking, not to mention *standardized* marking.

The impact on results: The poor setting of questions was likely to unfairly penalize some candidates; thus, lower scores resulted from poorly constructed questions. On the other hand, average students were pushed up to the GP 5 range (the highest grade) due to poor question setting and a non-standardized marking scheme. Thus the test was not able to differentiate between average and more able candidates.

From the above, it is evident test designers and markers of the secondary public examination do not follow or do not understand the guiding principles of *valid* and *reliable testing*. If tests are not aligned to the objectives of the curriculum and testing principles are not applied to test design, teachers and learners will find alternative ways of coping with the tests, thus spawning practices of teaching to the test, often relying on private tutoring, guidebooks, and rote learning.

32.5.2 *School-Based Assessment of Speaking and Listening*

The document analysis further shows that the current SSC test is not in accordance with the curriculum objectives, as it tests only reading and writing but not speaking and listening. The curriculum requires the assessment of speaking and listening skills of students at classes 9–10 through a school-based assessment system (SBA), allotting 20% of the marks to speaking and listening in Paper 1 with the remaining 80% earmarked for testing reading and writing in the SSC examination. It is relevant here to point out that the SBA system in secondary schools came under criticism from the schoolteachers themselves. Begum and Farooqui (2008) in their study on SBA expressed reservations about the effectiveness of this procedure for the following reasons: first, the teachers were not trained on how to undertake this assessment in a standardized manner; second, teachers themselves were involved in private tuition, hence their scoring of students might become biased; finally, institutions may be partial toward their own students as schools are often judged by public test scores especially schools that are dependent on monthly government grants to support teacher salaries. The education ministry scrapped the idea of SBA in 2013 and despite talk of reviving it in late 2014, it has not been implemented as of yet.

32.5.3 *Findings from the Questionnaire, Interview, FGD, Class Observation Data*

The wide-ranging data collected from the various sources described earlier was analyzed in order to identify recurrent patterns or categories (Cohen et al. 2007). These are:

- Perceptions on the fit between curriculum objectives and examination practices.
- Perceptions about examinations.
- Classroom practices in relation to the test and test preparation practices.
- Question setting and marking practices.

32.5.3.1 Fit Between Curriculum Objectives and Examination Practices

Although 50% *teachers* believed that the SSC examination matched the syllabus, interview findings demonstrated that there was a lack of a clear understanding of syllabus objectives. FGD data showed *students* did not understand this concept. They referred to test items, not to curriculum objectives. *Parents* too did not have much of an idea of the objectives and only 30% were aware of the examination content.

On the other hand, *question-setters* who set Board questions claimed that there was a fit as they followed Board instructions. However, the documentary analysis of the question papers showed they were ignorant of the issues of test validity and reliability. *Examination officials and administrators* (Board and NCTB) believed that syllabus objectives matched the examinations although they were unable to defend the allegations made in the 2010 National Education Policy's section on Examination and Assessment which stated the current method mainly assessed students' rote learning. Instead, the officials talked of test specifications and its administration rather than on the need for the type of test that reflected the curriculum objectives. The Report of the Advisory Committee for the Development of English (2010) had also pointed out that weaknesses of English tests lay in the fact that question setting in public examinations was not always in accordance with the curriculum. Both these policy documents advocated reforming the examination system.

This reform appears to have taken place but more as a cosmetic change in the 2012 Curriculum document without probing deep into significant issues related to test design and validity. The *classroom observation* data also provided evidence of atomistic practices of exam-oriented teaching of grammar, vocabulary, and writing without awareness of syllabus objectives. It was apparent that there was a general confusion between syllabus content and syllabus objectives.

32.5.3.2 Perceptions About Examinations

Teachers believed that most students were motivated by good scores. At the same time 78% of teachers were aware of the negative effects of tests stating that students suffered from stress and test anxiety. Teachers themselves were anxious about preparing students for upcoming tests.

Students held a somewhat positive view of examinations. 63% stated they enjoyed studying for exams mainly due to "personal enjoyment" and "family tradition." 50% studied hard for exams because "exams help us to pay attention to studies." A majority (80%) wanted to do well in exams so that they could have bright future prospects. Nearly 30% aimed to do well in English in order to "secure good jobs." Only a small number (6%) said they wanted to do well for themselves. Finally, 20% brought up negative consequences if results were poor. Loss of money and time and loss of face were highlighted as major consequences of failure.

Although 65% of *questions-setters* were aware of the syllabus, they actually followed the guidelines set by the Board and followed trends in previous years' test papers. On the other hand, the *exam administrators* stated the majority of questions do not reflect skills needed in real life which the syllabus emphasized but they still said that the questions were based on the syllabus.

Parents were deeply aware of the significance of the SSC examination and considered it important for two reasons. Firstly, it is the first public/national examination their children face and secondly, it is a pathway to higher education. They spent huge sums of money on private tuition and often had to take their children to different coaching centers. They clearly disapproved of the private tuition culture that had taken over the country and blamed teachers for being irresponsible with their in-class teaching and for being commercial. Their children were tired, anxiety-prone, and were sometimes depressed.

One parent said the exam grades were unpredictable. “My son did well in SSC but his HSC results were very poor.” Another parent commented, “The grades have lost their value. Questions leak very easily and student ability is not measured properly.” “Young students are under a lot of pressure both physically and mentally.” They also felt the creative question format was faulty.

32.5.3.3 Classroom Practices in Relation to the Test and Test Preparation Practices

The National Curriculum and Textbook Board (2012, p. 80) states, “students’ learning activities will be assessed through classwork, continuous assessment, terminal/public exams using teacher prepared or centrally prepared tools. It is to note that test tools will be based on all the learning domains where necessary.” The research data did not appear to give any evidence of such practices.

Most *teachers* stated that they practiced items related to the two skills, reading, writing, and grammar and vocabulary more than speaking or listening in class, as the latter two skills were not included in the examination. About two/three months to the test, class teachers used past SSC examination materials for revision and further practice. When asked if they practiced any *formative* assessment in their teaching, they appeared to confuse it with *summative* assessment, saying they always set quizzes and class tests and gave scores.

The *classroom observation* data showed that often test items were practiced but as part of the syllabus. Our data demonstrated that most teachers were not well informed about the concept of informal formative assessment practices that are used in class through teacher questioning, feedback, and through student interaction and homework.

85% of *teachers* claimed they prepared their students by reviewing previous lessons, explaining difficult concepts, practicing in class, conducting model tests, and solving old question papers. Teachers perceived private tuition had a big impact on students as it helped them to prepare by providing opportunities to practice and revise. They argued that private lessons also assisted in addressing individual needs of students by identifying their weaknesses.

In terms of exam preparation, 78% of *students* said they went to private tuition classes. 30% said they revised study materials and 20% said they solved past question papers. Again 70% reported that private tuition helped them to study and prepare as these sessions made topics easier. When asked about the role of teachers in exam preparation, 70% stated that teachers played a major role both in class and at after-school private tutoring by making them practice old test papers, explaining things repeatedly, and giving suggestions on what topics and items were likely to appear in the tests.

32.5.3.4 Question Setting and Marking Practices

Exam officials and test administrators stated special directions and guidelines were issued to exam-setters and markers. In relation to the nature of the questions, they stated the majority of questions did not reflect real-life skills needed but they insisted questions were based on the syllabus.

As they worked within a set format, all t*eachers* said they used past question papers of all eight boards for setting their school or board questions. 60% of the *question-setters* said they consulted the syllabus and all claimed they followed the board guidelines. 43% of setters believed students would have previous knowledge of answers. They made special reference to "creative" questions introduced in the new syllabus. They defined "creative questions" as "testing the creativity of the learners" (23%), "allowing students to write answers in their own way" (17%), and "developing writing skills" (21%). Actually, test items that provided opportunities for meaningful output like composition writing on personal/everyday problems/events and completing stories were meant to be creative items.

As for the *students*, they were more concerned with their scores. 35% of the *students* believed institutions were to be blamed for unfair assessment. Parents (45%) complained that standardized marking was not practiced and there was inconsistent rating. Most t*eachers* and *question-setters* emphasized there was a great need for training and workshops on writing test items, and on understanding the principles of standardization of test items and scoring. They felt there should be mandatory pre-marking meetings before final scripts were checked.

32.6 Insight(s) Gained

The current study findings on secondary school testing have enabled us to draw a number of insights. Firstly, there appears to be a general confusion among planners, teachers, and assessors between syllabus (or curriculum) *objectives* and syllabus *content*. It may be mentioned again that the objective of the English syllabus is to help learners build communicative competence and develop their language skills. However, the test does not match the syllabus specifications fully and there is a lack of fit between curriculum objectives and test content. It can be argued that students are not learning to use the language and language skills are not really being developed. Some of the test items basically test knowledge of forms and there is very little focus on eliciting authentic learner language. As regards a communicative approach, authenticity, a major factor in CLT-based tests, appears to be lacking. According to Bachman and Palmer's (1996) *authenticity* criteria, the test items we analyzed rarely used natural language, were not contextualized, and often did not represent real-life tasks.

The second insight we get from our findings is that there is an absence of a clear understanding of *informal formative assessment* that teachers need to use while teaching. The teachers' classroom practices show they are not aware of the elements and benefits of informal formative assessment as a routine part of teaching and learning although the curriculum emphasizes the importance of *continuous assessment*. This formative assessment needs to be specifically related to what has been taught and the information received from such assessment is usually used as a basis for further classroom work and learner development. As reported in the findings, teachers appear to confuse it with *summative* assessment, saying they always set quizzes and class tests and awarded scores.

The third insight we gained is the crucial lack of *assessment literacy* (Taylor 2009) among secondary school teachers and testers. The most significant challenge in the field of assessment appears to be the inadequate ability of teachers to devise appropriate tests and score them accurately. Competent testsetters and scorers with knowledge about the ingredients of good tests and test specifications will have "assessment literacy." The dearth of assessment literacy among ELT professionals is an obstacle to effective testing. At present many teachers do not have any formal training or coursework in testing or assessment. Finding ways to increase the assessment literacy of our teachers and practitioners is an immediate priority.

Fourthly, the concept of *washback* plays a crucial role in most high-stakes testing. Our findings show, corroborated by Khan (2010), that English teachers' perceptions of high-stakes tests have a severe "washback" effect on teaching-learning practices as well as classroom content. The specter of tests consumes teaching/learning energies in formal education and although assessment is professed to be skills-based, memorization has become part of the learning culture and widespread private tuition on test practices acts as a helpline for students (Rahman 2015). These after-school tutoring centers, similar to *juku* (fee-charging Japanese cram schools), are run by classroom teachers where exam papers are practiced (Rohlen 1980). Hence the culture of private tuition or supplementary schooling (in small groups or larger numbers) by the very same schoolteachers or by other interventionist entrepreneurs is widespread in Bangladesh. Bray (2007) has aptly called private tuition *shadow education* since it has become a modern socio-educational macro phenomenon worldwide. Our findings reveal that even policymakers and test officials/administrators shared a strong belief that learners had little confidence on gaining much from classroom teaching of English in mainstream secondary schools and would do well if they attended private coaching. Hamid et al. (2009) maintain that private tuition is highly regarded by learners, is desired by educational consumers, and is accepted in the family culture, even among those from disadvantaged communities. Moreover, it serves as an economic boost for low-paid teachers.

Finally, with regard to test *impact*, we can trace the wider influences of the secondary school examinations and their consequences far beyond the classroom. Madaus (1988) outlined some of the distortions in the principles of testing created through *test impact*. First, the more a test is used for decision-making, the more it is likely to distort and corrupt the social processes it is intended to monitor. Second, if important decisions are related to test results, teachers will teach to the test. Third, in settings where high-stakes tests are used, past examinations will define the curriculum. Finally, a high-stakes test transfers control over the curriculum to an examination board or an agency that sets and develops the test. Most of these factors seem to prevail in the secondary testing system in Bangladesh.

32.7 Conclusion: Implications for Test Users

Based on a review of the assessment situation and our study findings and discussion, implications may be drawn regarding working toward an alignment between an English language assessment policy at the secondary level and its actual practice. The starting point needs to be the development of *assessment literac*y (Taylor 2009). Teachers are often aware of their own inability in devising and implementing communicative language test items arising from this crucial lack of knowledge in assessment. Teachers generally rely on past question papers and question banks to guide them. Principles of test design, test specifications, reliability, validity, and standardization need to be part of a teacher's learning repertoire. Standardized scoring is an integral element of assessment. In addition, teacher training/development courses need to engage teachers meaningfully in addressing the concept of the holistic approach to changes in the curriculum. This may contribute in developing in-depth knowledge about teaching and learning. This may, in turn, have an effect on teachers' perceptions, attitudes, and beliefs of their practices in the classroom.

Overall, it is necessary to create a favorable classroom environment for implementing a communicative approach to teaching English and it is equally important to assess the learning outcomes in both a formative and summative manner. As all four skills are included in the curriculum, it is justifiable to include speaking and listening as well in the assessment framework. Whether it is carried out through a school-based assessment procedure or in any other way, validity and reliability of the tests need to be maintained.

More and regulated monitoring might be useful to ensure standardized effective practices in the classroom. Students' learning needs to be evaluated regularly with appropriate feedback on the learning so as to encourage learners to look upon assessment as a positive factor in learning and developing their own language skills. Ongoing evaluation (formative assessment) is ideally placed to support learners in subject-related or generic skills development. Self-assessment and peer-assessment are also additional forms of assessments which can promote resilience, confidence, a sense of ownership and autonomy for the lifelong learner.

In conclusion, we argue for a holistic approach to curriculum reform where the aims of the curriculum feed into the design of the course materials, determine the pedagogic practices of the teachers through the communicative approach, and most importantly, align assessment procedures to the main objectives of the curriculum. In this way, assessment will promote learning. If the assessment is well planned and is derived from the specifications laid out in the curriculum, learners are "trapped" and cannot escape without learning what is intended. This would be tantamount to a *positive* and desirable backwash effect of testing.

References

Ahmed, M., Nath, S. R., Hossain, A., & Kalam, M. A. (2006). *The state of secondary education: Progress and challenges.* Education Watch 2006. Dhaka: Campaign for Popular Education (CAMPE).

Akteruzzaman, M., & Islam, R. (2017). English, education, and globalisation: A Bangladesh perspective. *IAFOR Journal of Education, 5*(1). https://doi.org/10.22492/ije.5.1.10. Accessed 12 Apr 2018.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford, England: Oxford University Press.

Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257–279.

Bangladesh National Education Policy. (2010). *Ministry of education, government of Bangladesh.* https://www.moedu.gov.bd. Accessed 10 Jan 2019.

Baseline survey of secondary school English teaching and learning. Report—Ministry of Education. (1990). Dhaka, Bangladesh: Directorate of Secondary School Education.

Begum, M., & Farooqui, S. (2008). School based assessment: Will it really change the education scenario in Bangladesh? *International Education Studies, 1*(2), 45–53.

Bray, M. (2007). *The shadow education system: Private tutoring and its implications for planners* (2nd ed.). Paris: UNESCO.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions towards their classroom teaching. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 147–170). Mahwah, NJ: Erlbaum.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study.* Cambridge: Cambridge University Press.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 3–18). Mahwah, NJ: Laurence Erlbaum & Associates.

Choudhury, R., & Holbrook, J. (2018). *Report on the impact of SSC examination questions under 3 examination boards.* Bangladesh Examinations Development Unit (BEDU) & Secondary Education Sector Investment Program (SESIP). Dhaka, Bangladesh: Ministry of Education.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (7th ed.). London: Routledge.

Coombe, C., & Hubley, N. (2009). An introduction to key assessment principles. In C. Coombe, P. Davidson, & D. Lloyd (Eds.), *The fundamentals of language assessment: A practical guide for teachers* (2nd ed., pp. 3–10). Dubai, UAE: TESOL Arabia Publications.

Cumming, A. (2009). Language assessment in education: Tests, curricula, & teaching. *Annual Review of Applied Linguistics, 29,* 90–100.

Das, S., Shaheen, R., Shrestha, P., Rahman, A., & Khan, R. (2014). Policy versus ground reality: Secondary English language assessment system in Bangladesh. *The Curriculum Journal.* https://doi.org/10.1080/09585176.2014.909323. Accessed 1 Mar 2018.

Earl, L. (2003). *Asssessment in learning: Using classroom assessment to maximise student learning.* Thousand Oaks: Corwin Press Inc.

English in Action (EIA) Research Report. (2009). *An assessment of spoken English competence among school students, teachers and adults in Bangladesh: Baseline study 1.* Dhaka, Bangladesh: English in Action.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London: Routledge.

Hamid, M. O., & Baldauf, R. B., Jr. (2008). Will CLT bail out the bogged down ELT in Bangladesh? *English Today, 24*(3), 16–24.

Hamid, M. O., & Erling, E. J. (2016). English-in-education policy and planning in Bangladesh: A critical examination. In R. Kirkpatrick (Ed.), *English language education policy in Asia* (pp. 25–48). Switzerland: Springer International Publishing.

Hamid, M. O., & Rahman, A. (2019). Language in education policy in Bangladesh. A neoliberal turn? In A. Kirkpatrick & T. Liddicoat (Eds.), *Routledge handbook on language education policy in Asia* (pp. 382–398). London: Routledge.

Hamid, M. O., Sussex, R., & Khan, A. (2009). Private tutoring in English for secondary school students in Bangladesh. *TESOL Quarterly, 43,* 281–308.

Hargreaves, A., & Fullan, M. G. (Eds.). (1992). *Understanding teacher development*. New York: Teachers College Press, Columbia University.

Hasan, M. K. (2004). A linguistic study of English language curriculum at the secondary level in Bangladesh—A communicative approach to curriculum development. *Language in India, 4.* http://www.languageinindia.com/aug2004/hasandissertation2.html. Accessed 30 Mar 2019.

Hughes, A. (2013). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.

Imam, S. R. (2005). English as a global language and the question of nation-building education in Bangladesh. *Comparative Education, 41*(4), 471–486.

Islam, S. (2018, March 23). Bad performance in math and English caused SSC pass rates to drop to lowest in nine years. *The Daily Star*. https://bdnews24.com/bangladesh/2018/05/07/bad-performance-in-math-&-english-caused-ssc-pass-rates-to-drop-to-lowest-in-nine-years. Accessed 27 Mar 2019.

Khan, R. (2010). English language assessment in Bangladesh: Developments and challenges. In Y. Moon & B. Spolsky (Eds.), *Language assessment in Asia: Local, regional or global?* (pp. 121–157). South Korea: Asia TEFL.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, About Teachers and Teaching, 65*(3), 29–46.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302. Accessed 12 Jan 2019.

National Curriculum and Textbook Board. (2012). *National curriculum 2012. English, classes VI–X.* Dhaka, Bangladesh: NCTB.

Quader, D. A. (2001). Reaction to innovation in a Language teaching project. *Journal of the Institute of Modern Languages, 12*(2), 5–20.

Rahman, A. (2007). The history and policy of English education in Bangladesh. In Y. H. Choi & B. Spolsky (Eds.), *English education in Asia: History and policies* (pp. 67–93). South Korea: Asia TEFL.

Rahman, A. (2015). Secondary English education in Bangladesh. In B. Spolsky & K. Sung (Eds.), *Secondary school English education in Asia: From policy to practice.* Routledge Critical Studies in Asian Education (pp. 85–102). New York: Routledge.

Report. Advisory Committee for the Development of English. (2010). Ministry of Education, Government of Bangladesh.

Rohlen, T. P. (1980). The *juku* phenomenon: An explanatory essay. *Journal of Japanese Studies, 6,* 207–242.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29,* 21–36.

Tsui, A., & Tollefson, J. (2007). *Language policy, culture, and identity in Asian contexts.* Hillsdale, NJ: Erlbaum.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, Volume 7. Language testing and assessment* (pp. 291–302). Dordrecht: Kluwer Academic.

Weir, C., & Roberts, J. (1994). *Evaluation in ELT*. Oxford: Blackwell.

Chapter 33
A New Model for Assessing Classroom-Based English Language Proficiency in the UAE

Slim Khemakhem

Abstract This chapter reports on the findings of a doctoral case study that investigates the correlation of the Bachelor of Education students' speaking performance on the IELTS and in a class teaching situation. The B.Ed. program at the Higher Colleges of Technology, UAE sets IELTS band 6 as a graduation requirement that students must achieve before they start their final year of studies. This is meant to ensure that graduates have the minimum English language proficiency level to teach English in UAE schools. The research study adapts a mixed-methods approach and uses a corpus of students' speaking records both on a mock IELTS speaking test and a class teaching situation. The quantitative strand of the research reveals that there is no correlation between the subjects' lexical diversity in both contexts. The qualitative strand uses conversation analysis and shows that classroom interaction is very different than test interaction. Different classroom interaction features are found to account for those differences, but the most significant one is teacher repetition of lexical items in a class situation. The implication for IELTS test users and for students in an Education Program is that interpretations of test scores can be different than the actual language proficiency level demonstrated in a teaching situation. To bridge this gap, the researcher suggests a new assessment tool that merges IELTS band descriptors with the main features of classroom interaction so that correlation between the language test task and the language use is achieved. The new instrument is called Classroom-based English Language proficiency rubric.

33.1 Introduction: Purpose and Testing Context

In the absence of an institutional assessment tool that measures the Bachelor of Education students' English language proficiency level before they graduate as teachers of English in the United Arab Emirates (UAE), the Higher Colleges of Technology (HCT) set band 6 of the IELTS test as a fundamental requirement to register for the last year of the program. A failure to provide official evidence of

S. Khemakhem (✉)
Higher Colleges of Technology, Abu Dhabi, UAE
e-mail: slim.khemakhem@hct.ac.ae

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_33

achieving the required band puts students' registration on hold and delays their graduation. The choice of an IELTS score as an indicator of potential graduates' English language proficiency in the classroom seems to be based on an assumption that language proficiency demonstrated on an internationally recognized language proficiency test would be similar to language proficiency displayed in a teaching situation.

Based on a doctoral research study (Khemakhem 2016), this chapter investigates the validity of that assumption by drawing on a comparison between students' language performance on a mock IELTS speaking test and their performance in a teaching situation. It, also, compares Education students' scores on an IELTS speaking test with their mentoring teachers' scores on a teaching practicum. Moreover, the study analyzes features of classroom interaction and examines their impact on teachers' language proficiency level in the classroom with the view of integrating them in a new assessment tool. The new tool combines the IELTS context-free framework with the classroom context-bound language proficiency.

33.2 Testing Problem Encountered

As a high-stake decision-making tool that determines whether the Bachelor of Education students can graduate as teachers of English or not, the validity of using IELTS scores is called into question, especially in relation to the correspondence between language test tasks and language use in the classroom. The findings of the study have significant implications for the use of IELTS scores as a graduation threshold not only for a teacher education program, but also for other programs. The new assessment tool will also have significant implications for validating decision-making in regard to Education students' graduation as teachers of English in the UAE.

33.3 Review of Literature

33.3.1 Language Testing

To clarify the relationship between the test task and the language use task, Bachman and Palmer (1996) emphasize the necessity of a strong correspondence between the two tasks to be able to make accurate inferences about a candidate's language ability.

> If we want to use the scores from a language test to make inferences about individuals' language ability, and possibly to make various types of decisions, we must be able to demonstrate how performance on that language test is related to language use in specific situations other than the language test itself. (Bachman and Palmer 1996, p. 10)

According to Bachman and Palmer (2010), generalizations of interpretations based on the test scores can be made not only when individual attributes are engaged in an assessment task, but also when the characteristics of the context of language use are taken into consideration in designing an assessment task. Those characteristics involve the partners in language interaction. In the case of the current study, partners in the context of the language test task (IELTS speaking test) are neither the same nor of the same language proficiency as partners in the context of the language use task (the classroom).

Fulcher (2003, p. 19) posits that context has a significant impact on language performance that needs to be taken into account when designing speaking tests. He indicates that making inferences from such tests would be more accurate because they do not describe an individual's speaking ability in general but in relation to a specific language use domain. The target language use domain, in the context of the present study, is related to the classroom where teachers' language use is focused on teaching young second language (L2) learners the basics of the English language. However, in the IELTS speaking test, with the exception of the first part where the target language use is related to personal information, it suggests different language use domains depending on the topic that the examiner would choose for the candidate. Luoma (2004, p. 30) agrees with this argument and asserts that the cognitive and the experiential aspects of the context should be considered in designing speaking tests so validity can be established, and generalizability can be applied.

33.3.2 Validity

Validity is a major factor of trust or mistrust of decisions based on a test score. Despite its well-established validity and reliability through multiple validation processes and research studies, IELTS test scores remain vulnerable to criticism when they are misused. Early work of Messick (1989) on test validity emphasizes the fact that validity is not a property of the test itself but the inferences drawn from the test scores and the decisions built on them. He postulates that the test context, the test taker's background, and their experiential history are important in test score interpretation. Therefore, he suggests that context is a significant factor in the generalizability of score interpretations. This is important for the current study as it motivates the main question, which is whether IELTS scores can be generalized to a teaching and learning context or not. Messick (1989) claims that definitions of test validity have never considered the social consequences. In his unified framework of test validity, Messick contends that the evaluation of test validity should include the evaluation of "intended or unintended social consequences of test interpretation and use" (Messick 1989, p. 84). He establishes a distinction between test interpretation that focuses on the evidential validity of the construct itself, and test interpretation that focuses on the evidential validity of test use, which should provide enough evidence of relevance of the construct to the purpose and the setting of the test use. In the current study the evidential validity of the construct is established by the use of IELTS as a valid standardized test, but the evidential validity of test interpretation and use does not seem to be established in the absence of a clear correspondence between the construct of language ability and the context of language use.

33.3.3 Language Proficiency

Bachman and Palmer (2010, p. 34) define language use as creating, interpreting, or negotiating meaning between individuals in a particular situation. They identify two kinds of interactions in language use: (1) interaction between the attributes of the language user, which includes topical knowledge, language knowledge, and affective factors, and (2) interaction between the language user and the context that can include other language users. By looking at the way IELTS scores are used to make high-stakes decisions at HCT, it is clear that the first type of interaction is taken into consideration and it forms the basis for making decisions. However, it does not seem that policy-makers pay enough attention to the second type of interaction, which involves the situation of language use, and the interaction between the Education students and their young L2 learners in the classroom.

Cummins's (2000) model of language proficiency emphasizes the impact of context in defining the language proficiency of individuals. Context disembeddedness in the speaking test situation, where contextual clues range from minimal to non-existent, does not resemble context embeddedness of the classroom where contextual clues are maximized due to the teacher's prior knowledge of the context including the lesson content, the language level of learners, and the pre-planned interaction.

Freeman et al. (2015) distinguish between general English language proficiency and context-specific proficiency. They call English language that is used in the classroom "English-for-teaching." They challenge the general belief that good teaching and learning of English language happens because of having teachers with good or high levels of English language proficiency. They argue that a good proficiency level is not enough if it is not connected with appropriate classroom practices.

In his Audience Design framework, Bell (1984) claims that differences in individuals' speech are usually attributed to differences in audience. He indicates that speakers use speech accommodation strategies that vary from one context to another depending on the type of addressees. Convergence strategies are usually applied by making linguistic choices that match the level of the addressees in a specific context to win their approval. In the context of the classroom, teachers implement convergence strategies to ensure that input is comprehensible for their students, whereas in the context of the IELTS speaking test, Education students use convergence to align their output with their examiner's expectations. In both contexts, the subjects try to display language proficiency at its highest level of convergence, but not necessarily at its highest level of proficiency.

33.3.4 Lexical Diversity

Lexical proficiency is taken as a reliable indicator of language proficiency (Daller et al. 2003; Read and Nation 2006; Yu 2009). Crossley et al. (2011) examine a number of studies that investigate different features of lexical proficiency including lexical diversity, and they conclude that lexical diversity is a reliable indicator of lexical proficiency. Jarvis (2013, p. 89) also confirms the findings of previous studies and indicates that "learners' word choices contribute to the complexity and quality of their language use."

In the present study, lexical diversity is taken as an indicator of language proficiency, and Education students' lexical diversity scores (*D*) on the mock IELTS speaking test are compared with their scores in a teaching situation to draw conclusions on the strength of their correlation.

33.3.5 Classroom Interaction

Classroom interaction is known for being unique because of its distinctive features and special context. Some of its most discussed features in the literature are teachers' control of interaction, speech modification, elicitation techniques and questions, repair, scaffolding, and repetition.

Seedhouse (2004, p. 205) postulates that the architecture of the L2 classroom interaction is distinct due to the uniqueness of the institutional context where it takes place. He indicates that the L2 classroom is "the actualization of the reflexive relationship between pedagogical focus and interactional organization." He advocates that an emic perspective in the analysis of classroom interaction is very necessary to be able to identify its properties.

Walsh (2013, p. 106) proposes a framework for the analysis of classroom interaction that combines Conversation Analysis and Corpus Linguistics (CACL). He claims that it provides a "detailed micro-analytic descriptions of spoken interaction." Like Seedhouse (2004), he believes that classroom interaction is different from any other type of conversation because it is related to the enterprise of learning a second language. He finds that Conversation Analysis does not impose a certain structure on classroom interaction but it lends itself to the properties of the context of the classroom. He posits that Corpus Linguistics complements Conversation Analysis as it examines long texts in detail with a focus on words and their combinations with disregard to their context.

33.3.5.1 Control and Monitoring of Interaction

According to Sinclair and Coulthard (1975), classroom interaction has three main components: initiation (I), response (R), and feedback (F). Teachers initiate (I) interaction by asking closed or open questions to which their students respond (R). Then, teachers give verbal or non-verbal feedback (F) on the students' responses that either takes the interaction to another level of the IRF structure involving more responses and more feedback, or it closes it to start a new IRF cycle. Therefore, teachers, according to this model, take two-thirds of the classroom interaction, and in both parts, their role is to orient interaction toward a start, a follow up, or a closure. However, on the IELTS speaking test, teachers take the other third of the interaction model (R) in which they can only respond to an initiator. Teachers have very little or no control over the way interaction goes in the speaking test due to the test format that gives priority for the examiner to initiate interaction through a question or a prompt. Myhill et al. (2006) regard teachers as orchestrators of classroom interaction, which makes their contributions to classroom interaction more prevalent and gives them control of interaction to achieve the targeted lesson objectives.

33.3.5.2 Speech Modification

Many specialists in second language acquisition (SLA) like Hatch (1978), Long (1983), and Lightbown and Spada (2006) discussed the importance of speech modification in the L2 classroom as a way to make meaning accessible for L2 learners. Teachers try to use high frequency vocabulary, and they usually try different lexical choices in order to ensure that meaning is clear for their students. They simplify sentence structure, and they reduce syntactic complexity to the most basic and familiar forms that facilitate understanding. They use body language, and sometimes translation of some lexical items into the mother tongue to make meaning clear. Walsh (2011, p. 9) identifies six speech modification strategies including confirmation checks, comprehension checks, repetition, reformulation of students' utterances, completing unfinished utterances of students, and backtracking to recall parts of the conversation. Chaudron (1988) sums up research findings in seven strategies that include slower speech rate, more frequent and longer pauses, simplified and exaggerated pronunciation, basic vocabulary, lower degree of subordination, more declarative statements, and frequent self-repetitions by teachers.

33.3.5.3 Elicitation Techniques and Use of Questions

Elicitation techniques are widely used in a classroom context. Teachers use questions abundantly to elicit required information from their students. They ask multiple questions in different structure forms and using high frequency vocabulary to encourage students' participation. Walsh (2011, p. 33) identifies elicitation as one of five purposes for display questions that teachers ask in a classroom. Chaudron (1988) indicates that previous research has found that teachers ask more display questions that have known answers than referential questions. Tsui (1992, p. 101) identifies six categories of elicitation in social conversations, but Weng in Kao et al. (2011) adds two classroom-based types which aim to check students' understanding of explained content and to check understanding of given instruction.

33.3.5.4 Repair

Error correction or "repair" plays a major role in the L2 classroom. Unlike previous features of classroom interaction where the teacher's role pertains to the initiation part, repair pertains to the feedback part. Though it is usually conducted by the teacher, repair can be performed by the same speaker or by classmates. It is achieved through four "trajectories" as per Sacks et al. (1974) classification: (1) Self-initiated self-repair, (2) Other-initiated self-repair, (3) Self-initiated other-repair, and (4) Other-initiated other-repair. Van Lier (1988) distinguishes between two types of repair, (1) "conversational repair" that focuses on meaning and (2) "didactic repair" that focuses on the form of language. Seedhouse (2004) regards the relationship between repair and the pedagogical focus of the lesson as a reflexive relationship. In total opposition with the classroom situation where a teacher does not perform "self-initiated self-repair" trajectory, in a test situation it is the only possible trajectory that a student teacher can perform.

33.3.5.5 Scaffolding

Scaffolding is a salient feature of teaching and learning, especially in a second language setting where teachers assist students with language learning and understanding of targeted skills and knowledge. Jerome Bruner (1983, p. 38) considers that adults' input in the context of child learning is formulated in a way that helps to process concepts and communicative functions. He calls it "Language Acquisition Support System" (LASS). Walsh (2013, p. 9) calls it "linguistic support" which is meant to help the learner internalize new knowledge and make use of it consciously. It is based on the concept of challenge to engage the learner and the concept of support to help understanding. He identifies three types, namely, reformulation, extension of a student's utterance, and modeling.

33.3.5.6 Repetition

Repetition constitutes a distinctive feature of classroom interaction. Many researchers in second language acquisition confirmed the salient role of repetition in the L2 classroom and those include Chaudron (1988), Cook (1994), Gass et al. (1998), and Piirainen-Marsh and Tainio (2009). There are different reasons that cause teachers' use of repeats. White and Lightbown (1984) find that teachers repeat questions many times as a way to insist on getting answers for their questions. They report that 64% of an average of four questions per minute are repeated in teachers' talk. Seedhouse (2004) indicates that teachers repeat students' wrong utterances with a rising intonation to draw their attention to errors in structure or content and to incite them to conduct self-correction. In a classroom context, repetition is a desirable feature of teacher talk as it facilitates learning and teaching, and it promotes language development. However, in a speaking test, repetition is usually associated with hesitation and redundancy, which are interpreted as lack of fluency and inability to maintain coherent flow of ideas.

33.4 Methodology

The research methodology follows a consequential explanatory mixed-methods design. The quantitative data are used to provide generalizable statistics to evaluate the strength of correspondence between the Education students' language performance on the IELTS test and in the classroom. The qualitative data provide a subsequent analysis of the factors that affect the strength of that correspondence.

The main question of the research study is the following:

> Research question: To what extent are IELTS scores valid indicators of student teachers' language proficiency in the classroom in the UAE setting?

The research sub-questions are:

> Sub research question 1: To what degree does the lexical diversity of student teachers on the IELTS speaking test look similar to their lexical diversity in the classroom?
> Sub research question 2: To what degree do the teaching practicum scores awarded by school and college mentors confirm IELTS scores?
> Sub research question 3: How does classroom interaction affect the lexical diversity of student teachers?

The answers to the first and second sub-questions are mainly provided through a quantitative data analysis. The lexical diversity of student teachers' language performance on the IELTS test and in the classroom is computed and compared to provide a statistical value for the strength of their correspondence. The analysis also compares the frequency of content words in each context to draw conclusions about the degree of resemblance between the two performances. Moreover, the quantitative analysis compares scores obtained on the IELTS speaking test with scores obtained on the teaching practicum to provide further statistical evidence of the strength of the relationship between the two main variables.

The qualitative analysis is a follow-up phase that answers the third sub-question. It analyzes the factors that affect students' performance in the classroom and determines the kind of relationship between the two performances under investigation. The analysis follows an "explanatory sequential design" (Creswell and Clark 2011, p. 81) where the qualitative phase is used as a follow-up explanatory stage that provides in-depth analysis of the statistical data.

33.4.1 Data Collection Procedures

The research data are collected from 27 Emirati students who study in the Bachelor of Education program (B.Ed.) and who got the official consent of their parents as per the cultural norms in the UAE.

33.4.2 Data Collection Instruments

33.4.2.1 IELTS Mock Speaking Test

For data collection, four different versions of IELTS speaking test were selected from different IELTS published resources. The test has ten different levels of performance that are known as bands. They start from band zero, meaning "no attempt to take the test" to level 9, which is the "expert user" level. Band 6, which is the required band for HCT Education students to graduate, is equivalent to "competent user" (IELTS 2007).

The speaking component of IELTS test has three parts. Part 1 tests the candidate's ability to answer personal questions related to where they live, their families, studies, jobs, interests, or other familiar topics. Part 2 tests the ability to talk about 1 to 2 min in response to a prompt. Part 3 tests the candidate's ability to engage in a discussion with the examiner to answer higher-order thinking questions related to the topic of part 2. The test can last between 11 and 14 min, and the assessment criteria focuses on four main areas, (1) fluency and coherence, (2) lexical resources, (3) grammatical range and accuracy, and (4) pronunciation.

33.4.2.2 Class Recordings

The same 27 subjects who took the mock IELTS speaking test were asked to record one of their classes on teaching practicum. Videotaping was not possible due to cultural reasons. Instead, students were asked to use audio-recording for any lesson of their own choice. The collected recordings varied in length, in content, and in language focus depending on the grade level.

33.4.3 Data Analysis Tools

Data analysis procedures started by transcribing the recordings of the test and the class teaching sessions, and assigning identification codes for each participant to secure anonymity.

33.4.3.1 Quantitative Analysis Tools

Lexical diversity is measured by using index *D*, which is a mathematical model created by Malvern et al. (2004) and made available as a computational software. All transcripts are converted to and coded by CHAT (Codes for Human Analysis of Transcripts) and processed by CLAN (Computerized Language Analysis) using VOC-D software provided in MacWhinney's (2010) CHILDES (Child Language Data Exchange System). To find *D*, VOC-D divides the number of types of words in a text by the number of tokens (the total number of words in a text), then it creates a probability-based formula that takes into consideration the falling curve of type-token ratio as the text gets longer and longer. At the same time, the formula creates a theoretical curve using D coefficient to find the best fit between the two curves, which is the *D* value of lexical diversity.

Following *D* computation, a statistical analysis was conducted using the Statistical Program for Social Sciences (SPSS) to provide quantitative answers for the research sub-questions 1 and 2.

A further statistical comparison was conducted using Wordsmith Tools, which is a composite program used in corpus linguistics to find out word lists and frequency, concordances, and keyword lists. It is used to provide more statistical data related to the types of lexical choices made by the subjects on the IELTS test and in the classroom in order to provide more insight into the similarities and differences between the two performances.

33.4.3.2 Qualitative Analysis Tools

The qualitative phase of analysis follows the quantitative phase with a focus on classroom interaction to identify characteristics that can explain differences between the subjects' lexical diversity on the test and in the classroom teaching. The analysis uses Walsh's (2011) combined model of Conversation Analysis and Corpus Linguistics (see Sect. 33.3.5).

33.5 Findings

33.5.1 The Quantitative Analysis

33.5.1.1 A Comparison Between the *D* Scores of IELTS and the Classroom

The computation of lexical diversity (*D*) revealed that 22 subjects out of 27 (81%) displayed higher levels of lexical diversity on the IELTS speaking test than in class teaching. On a lexical diversity scale of 100 as defined by McCarthy and Jarvis (2010, p. 383), the highest *D* score on IELTS is 86.13, whereas the highest *D* in the classroom is 71.65. In a similar pattern, the lowest score on IELTS is 50.45, whereas the lowest score in the classroom is 29.96. According to McCarthy and Jarvis (2010), a difference of 10 points in *D* scores is enough to make valid inferences. By looking at the scores of the 22 students with higher *D* values in IELTS than in the classroom, we can notice that 18 of those scores (82%) are higher by more than 10 points than the class scores. This significant statistical finding illustrates clear differences in the Education students' lexical proficiency as demonstrated on the IELTS speaking test and in the classroom. Those differences reflect a lack of correspondence between IELTS speaking test tasks and language use in the classroom.

A computation of the means of both sets of *D* scores shows that the IELTS mean is 69.05, whereas the classroom mean is 56.36, with a difference of around 12 points. This indicates once more that the discrepancies between the two performances are significantly high.

33.5.1.2 Correlations of IELTS and Classroom *D* Scores

The results generated by SPSS for the correlation of IELTS *D* scores and the class *D* scores, using Pearson product-moment correlation show a non-significant relationship between the two sets of scores with $r = 0.160$, $n = 27$, and $p > .05$ (2 tailed).

A second correlation between the raw scores of IELTS speaking test that were assigned by the examiners and the scores of the teaching practicum that were assigned by the college mentors demonstrate that there is a non-significant relationship with $r = 0.142$, $n = 27$, and $p > .05$ (2 tailed).

Both correlations give clear evidence that statistically the correspondence between the speaking test and classroom teaching is a weak one, which legitimizes the argument of the research study regarding the use of IELTS scores as indicators of language proficiency in the classroom.

33.5.1.3 Content-Word Frequency

Using Wordsmith tools (see Sect. 33.4.3.1), the top twenty content words in each context were identified and compared in terms of their frequency and their concordances. Results show that verbs are the most used words in both the IELTS and the class. However, the frequency rate in the classroom is clearly higher (70%) than on IELTS (50%). Nouns are the second in frequency with 20% only in the classroom and 30% on the test. This finding reveals that classroom talk relies heavily on using verbs to a percentage that exceeds two-thirds. Most of those verbs are instructional verbs that are commonly used by teachers like “listen,” “sit,” “write,” and “look,” etc.; IELTS, on the other hand, relies less on verbs and more on nouns than the classroom. Adjectives and adverbs represent only a small portion of the total word list. Figure 33.1 illustrates the distribution of the top twenty words in each context.

The analysis of verb lemmas (word meaning in context) in the classroom shows that verbs are used for two main purposes:

- To give instruction related to academic skills like “read,” “write,” and “listen.”
- To manage learning and teaching activities and students’ behavior like “go,” “come,” “sit,” “look,” and “finish.”

Though verbs take a smaller portion of Education students’ performance on the test compared with the classroom, the examination of word lemmas of the test performance shows a much wider variety of meanings. The differences can clearly be attributed to differences in language use in each context. The language use in the classroom is restricted to the learning and teaching process and to the management of that process as revealed by the examination of the most frequent word lemmas. However, the test scope covers a wider and more general range of topics that do not resemble in any way the classroom context.

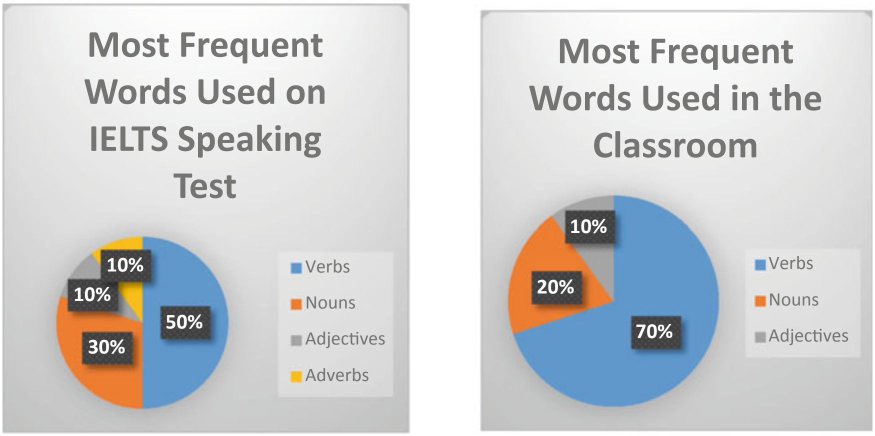


Fig. 33.1 The distribution of the top twenty words in IELTS and in the classroom (Khemakhem 2016, p. 122)

33.5.2 The Qualitative Analysis

The qualitative analysis adopts an emic perspective following Conversation Analysis as per Seedhouse's (2004) framework for the analysis of classroom interaction, and using Walsh's (2011) CACL model. The detailed examination of the transcripts shows that teachers' frequent repeats of lexical items is a major characteristic that marks teacher talk and that impacts lexical diversity values in a significant way. The different types of repeats have been identified with reference to the main classroom interaction features as discussed in Sect. 33.3.5 and to word concordances and keyword tools of Wordsmith. As a result, a new taxonomy of teacher repeats is generated, which includes six main categories and sixteen related types (Table 33.1).

The significance of repeats in teacher talk outlined in Table 33.1 gives clear evidence that language use in the classroom is distinctive and cannot be generalized to other contexts. Unlike the test context where repeating the same lexical items is taken as a limitation of lexical resources that justifies low grading, lexical repeats in the context of the classroom are taken as a virtue that justifies higher grades.

33.6 Insights Gained

In sheer disagreement with Messick's (1989) and Bachman and Palmer's (2010) definitions of valid test score interpretations and test score use for subsequent decisions, the data analysis results in the previous section unveil major concerns regarding the use of IELTS scores as indicators of the Education students' language proficiency in the classroom in the UAE. In fact, the raised concerns and the gained insights can be generalized to other similar teacher education contexts beyond the UAE where student teachers are assessed for their English language proficiency in the classroom.

The raised concerns are classified under two main categories: lack of correspondence and validity issues.

33.6.1 Lack of Correspondence

The lack of correspondence between what IELTS scores indicate and what they are used for is reflected in the following areas:

33.6.1.1 Opposing Tasks

For the test task, the subjects try to showcase their language abilities. They use low frequency words and they avoid repetition of the same lexical items to display desired lexical proficiency in order to impress the examiner and get a high score. However,

Table 33.1 A taxonomy of teacher repeats (Khemakhem 2016, p. 130)

	Repeat category	Repeat type	Repeated lexical items
1	Interaction-control repeats	In IRF exchanges	Lexical items that a teacher repeats in an initiation turn (I), especially when she does not get an immediate or an appropriate response (R)
		In group work	Lexical items that a teacher repeats to monitor interaction in group-work activities in order to keep it oriented toward the pedagogical focus of the lesson
		In repair phases	Lexical items that a teacher repeats in a repair phase as a way to encourage self-correction
2	Question repeats	Elicitation and modification repeats	Lexical items that a teacher repeats in elicitation questions or in modified forms of those questions to elicit answers
		Strict question-repeats	Lexical items in questions that teachers repeat a number of times with no modification, especially in form and accuracy contexts
		Think-time repeats	Lexical items in questions that a teacher asks more than once to give think-time to their students
3	Feedback repeats	Confirmation repeats	Lexical items in a teacher's repeats of students' correct answers/contributions as a way of confirming them
		Praise-word repeats	Lexical items that a teacher repeats when praising students for their correct answers or contributions
		Repair repeats	Lexical items that a teacher repeats to correct students' answers/contributions
4	Key-word repeats	Lesson key-word repeats	Lexical items that relate to the main focus of the lesson, and that the teacher repeats throughout the class

(continued)

Table 33.1 (continued)

	Repeat category	Repeat type	Repeated lexical items
		Activity key-word repeats	Lexical items that relate to a specific activity and that the teacher repeats while the activity is being conducted
		Story key-word repeats	Lexical items that a teacher repeats as part of rhyming lines in a story
5	Approach-related repeats	Language-drill repeats	Lexical items that a teacher repeats to model the pronunciation/spelling of new vocabulary and/or target grammar structures
		Scaffolding repeats	Lexical items that a teacher repeats in scaffolding turns
6	Procedural repeats	Instruction-clarification repeats	Lexical items that a teacher repeats to clarify the procedure of carrying out an activity in a successful way, or lexical items that are used to clarify instructions for an activity when students show that they do not understand the procedure clearly
		Classroom management repeats	Lexical items that a teacher repeats to control students' attention, movement/behavior, and to start and finish activities

for the classroom language use, students try to use high frequency words and use repetition of key lexical items more often to facilitate learning and to demonstrate effective teaching skills.

33.6.1.2 Different Characteristics of Test Takers and Language Users

There is a lack of correspondence between the characteristics of Education students as test takers and their characteristics as language users in the classroom. As test takers, they respond to the examiner's questions or prompts while trying to display the highest level of complexity and diversity and while taking into consideration the status of the examiner as an expert language user. However, as language users in the classroom, they simplify their language by limiting their lexical choices to familiar vocabulary, using simple sentence structures, and deploying repetition continuously to encourage students' participation and enhance language learning.

33.6.1.3 Different Contexts and Different Cognitive Demands

Lack of contextual clues in the test task situation solicit higher-order cognitive abilities to be able to provide coherent, appropriate, and accurate answers while showcasing diversified lexical and grammatical knowledge. However, abundance of contextual clues in the classroom, due to prior knowledge of the curriculum and the students' level of language proficiency, requires skillful use of a limited range of lexical items and grammatical patterns to facilitate learning. Therefore, teachers' cognitive abilities are engaged in filtering out low frequency words and complex language structures for better learning outcomes.

Field (2011) considers that the cognitive validity of a speaking task is established when candidates use the same mental processes in real life. This is clearly not established in the way IELTS speaking scores are used, especially the discrepancies found between the Education students' scores on the speaking IELTS test and their scores on the teaching practicum.

33.6.1.4 Opposing Accommodation Strategies

Statistical findings show clear discrepancies in lexical diversity scores due to differences in addressees in each context. The audience design in the classroom is configured according to the properties of the L2 classroom interaction. Language use is controlled by the proficiency level of young L2 learners, which can be at a very beginning level. Teachers use accommodation strategies to simplify their language and to make it comprehensible for their addressees in order for it to be conducive to learning. On the test, however, the configuration is controlled by the language level of the examiner, which is at a higher level of proficiency than the test taker. The accommodation strategy in a test situation works in a different direction compared with that of the classroom. On the test, Education students try to converge with their examiner's level of proficiency following an ascending path. It starts at their own level of proficiency and tries to reach the highest point closer to the examiner's level in order to secure the highest possible score. By contrast, in the classroom Education students' accommodation strategies follow a descending path. Their convergence starts from their own language level down to the furthest point closer to the proficiency level of their students.

33.6.2 Validity Issues

As a result of the previously discussed discrepancies, a number of validity issues arise when using IELTS scores as indicators of the Education students' language proficiency in the classroom. However, it is worth noting and emphasizing that the validity issues discussed in this section do not put into question IELTS as a widely recognized valid test of English language proficiency through multiple research studies and validation procedures, but the way its scores are used to measure what the test is not designed for, which is the ability to teach English language for young ESL learners.

33.6.2.1 Inappropriate Inferences

Inferences based on the required IELTS score, support the assumption that students who get band 6 or more are competent language users and by definition will be good teachers of English in UAE schools. However, the quantitative and the qualitative data analyses in the previous section demonstrate that it is a misconception because preferred performance on the language test task does not resemble required performance in the classroom. Therefore, inferences from test scores are not valid assumptions about Education students' language proficiency in the classroom. The discrepancies found between students' grades on the teaching practicum and their scores on the IELTS speaking test show that inferences based on the test scores are inaccurate. Teaching practicum mentors grade their student teachers' language proficiency in terms of its appropriateness for the classroom context. This includes their ability to grade down their language proficiency level to the level of their learners using high frequency words, simple sentence structures, and appropriate use of repetitions. Demonstrating opposite practices brings their scores down as it is regarded as a failure to use appropriate accommodation strategies in order to facilitate learning.

33.6.2.2 Generalizability Issues

Test score generalizability depends on the similarity between the context of the test and that of language use situation (Messick 1989). Context validity of IELTS scores for the graduation of Education students is compromised because of a number of contradictions between the two contexts of students' performances. These include, but are not restricted to, the setting, the purpose of the two tasks, audience design, features of interaction, linguistic challenges, and so on. Where the context of the test requires showcasing language ability at its highest level of complexity and diversity, the context of the classroom requires displaying it at the most appropriate level of simplicity and familiarity for the learners to ensure effective teaching and learning. Therefore, taking the test context as a representative of the classroom context is an erroneous assumption that makes generalizations of the test scores to the teaching and learning context invalid.

33.6.2.3 Misinformed Consequences

With reference to Messick's (1989) consequential basis for test interpretation and use, the value implications of the IELTS scores misinform decision-making at HCT in regard to Education students' English language proficiency for teaching and learning.

In fact, IELTS band 6, which is taken as a reference band has been proven to be an invalid indicator, in the current research study. The lack of correspondence between the test task and classroom language use suggests that test scores do not provide accurate information on Education students' language proficiency in the classroom, and consequently they represent an invalid basis for subsequent decisions. IELTS test scores, in this sense, misinform decision-making and lead to invalid consequences that relate to Education students' future careers.

33.7 Conclusion: Implications for Test Users

To bridge the gap between the intended aim of using IELTS scores for the context of the Education program and the current flawed use, a new assessment tool that measures Classroom-Based English Language Proficiency (C-BELP) is proposed. It is a tool that assesses Education students' English language proficiency as they conduct their teaching and learning activities in the classroom. It adapts IELTS assessment criteria while considering the characteristics of classroom interaction. In fact, IELTS band descriptors are translated into classroom-based English language proficiency descriptors using the findings and insights gained from the actual research study.

Four levels of classroom-based English language proficiency are identified in correspondence with the four years of study in the Bachelor of Education program and with IELTS bands 5 to 8:

Level 1 corresponds to IELTS English language proficiency band 5, which is the program entry requirement.
Level 2 corresponds to year two, which is equivalent to IELTS band 6.
Level 3 corresponds to year three, which is equivalent to IELTS band 7.
Level 4 corresponds to year four, which is equivalent to IELTS band 8.

Like IELTS, the assessment criteria are based on four main language proficiency skills, namely, fluency and coherence, lexical resources, grammatical range and accuracy, and pronunciation. However, the descriptors for each level are adapted to the context of the classroom and the requirements of teaching and learning (Table 33.2).

The assessment rubric is designed for college mentors to assess Education students' English language proficiency in the classroom context. The use of the rubric will help to overcome validity issues identified in previous sections.

The C-BELP rubric is suggested to be used solely for the assessment of required English language proficiency in the classroom, but cannot supersede IELTS or any other standardized English language proficiency test that measures academic or general language proficiency. Ideally, the Education program at HCT should adopt this new model (C-BELP) to make informed and valid decisions at the end of year three in regard to student teachers' competence in using English for teaching. Achieving level 3 of C-BELP is an appropriate indicator that Education students have

Table 33.2 An assessment rubric for classroom-based english language proficiency (C-BELP) (Khemakhem 2016, p. 195)

Level	Fluency and coherence	Lexical resources	Grammatical range and accuracy	Pronunciation
4	• Speaks fluently while adjusting the pace to the level of the learners • Speaks coherently all the time • Uses repetition adequately	• Uses an adequate range of vocabulary items • Uses appropriate vocabulary items for the level of the learners • Uses high frequency words all the time	• Uses an adequate range of simple and complex structures • Produces accurate sentences and word forms all the time • Uses complex sentences only when they are appropriate for the levels of the learners	• Uses a wide range of pronunciation features • Makes speech clear enough to all learners • Models correct pronunciation through clear articulation of sounds all the time
3	• Speaks fluently while trying to adjust pace to the level of the learners • Speaks coherently most of the time • Uses repetition adequately	• Uses a range of vocabulary items that is mostly appropriate for the learners • Uses appropriate vocabulary items for the level of the learners most of the time • Uses high frequency words most of the time	• Uses a range of simple and complex structures that are mostly appropriate for the learners' level • Produces accurate sentences and word forms most of the time • Uses complex sentences that are generally appropriate for the levels of the students	• Uses a reasonable range of pronunciation features • Makes speech clear most of the time • Model correct pronunciation through clear articulation of sounds most of the time
2	• Shows some hesitancy and some difficulty to adjust the pace to the level of the learners • Speaks fluently while showing some difficulty to adjust pace to the level of the learners • Speaks coherently most of the time • Does not use repetition adequately	• Uses a range of vocabulary items that is sometimes above the learners' level • Uses appropriate vocabulary items for the level of the learners on irregular basis • Uses some low frequency words	• Uses a range of simple and complex structures that is sometimes above the learners' level • Produces some inaccurate sentences and word forms • Uses some complex sentences that can confuse learners and hinder comprehension	• Uses a limited range of pronunciation features • Shows difficulties to make speech clear for the learners • Shows difficulties to model correct pronunciation of some sounds

(continued)

Table 33.2 (continued)

Level	Fluency and coherence	Lexical resources	Grammatical range and accuracy	Pronunciation
1	• Shows frequent hesitancy that affects fluency and message clarity • Speaks fluently but fails to adjust pace to the levels of the learners • Produces incoherent utterances • Does not use repetition	• Uses a wide range of vocabulary items • Uses low frequency words • Uses inappropriate vocabulary for the level of the learners	• Uses a wide range of simple and complex structures • Produces frequent inaccuracies in sentence and word forms • Uses complex sentences most of the time	• Uses a very limited range of pronunciation features • Shows difficulties to make speech clear for the learners • Fails to model correct pronunciation of sounds

attained the minimum required level of English language proficiency for the classroom that allows them to access year four and graduate. However, for the academic work that the year-three students submit for the different assessments of the program courses, the required IELTS band 6 can be incorporated into the Education assessment rubrics. As applied for C-BELP, the IELTS bands 5, 6, 7, and 8 criteria can be merged with Education assessment criteria and descriptors to generate a new assessment rubric for submitted academic work. It can be used along with the C-BELP to validate the Education Program assessment practices, inferences, and consequences for both English for academic purposes and English for teaching.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bell, A. (1984). Language style as audience design. *Language in Society, 13*(2), 145–204.

Bruner, J. (1983). *Child's talk*. Oxford: Oxford University Press.

Chaudron, C. (1988). *Second language classroom*. Cambridge: Cambridge University Press.

Cook, G. (1994). Repetition and learning by heart: An aspect of intimate discourse, and its implications. *ELT Journal, 48*(2), 133–141.

Creswell, J. W., & Clark, V. P. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28*(4), 561–580.

Cummins, J. (2000). *Language, power and pedagogy*. Clevedon: Multilingual Matters.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197–222.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Studies in language testing 30: Examining speaking* (pp. 65–111). Cambridge: Cambridge University Press.

Freeman, D., Katz, A., Gomez, P. G., & Burns, A. (2015). English-for-teaching: Rethinking teacher proficiency in the classroom. *ELT Journal, 69*(2), 129–139.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Education.

Gass, S. M., Mackey, A., & Pica, T. (1998). The role of input and interaction in second language acquisition: Introduction to the special issue. *Modern Language Journal, 82*(3), 299–307.

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition*. Rowley, MA: Newbury House.

IELTS. (2007). *Official IELTS practice materials*. Cambridge: University of Cambridge ESOL Examination.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(1), 87–106.

Kao, S. M., Carkin, G., & Hsu, L. F. (2011). Questioning techniques for promoting language learning with students of limited L2 oral proficiency in a drama oriented language classroom. *The Journal of Applied Theatre and Performance, 6*(4), 489–515.

Khemakhem, S. (2016). *Investigating the predictive validity of IELTS for a teacher education program in UAE*. Unpublished doctoral thesis, University of the West of England, Bristol, England. http://eprints.uwe.ac.uk/28078. Accessed 10 Sept 2018.

Lightbown, P. M., & Spada, N. (2006). *How languages are learned* (3rd ed.). Oxford: Oxford University Press.

Long, M. H. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics, 4*(2), 126–141.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

MacWhinney, B. (2010). *The CHILDES project*. http://childes.psy.cmu.edu/manuals/CHAT.pdf. Accessed 15 Oct 2010.

Malvern, D., Richards, B. J., Chipere, N., & Dúran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behaviour Research Methods, 42*(2), 381–392.

Messick, S. (1989). Validity. In R. L. Lynn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Myhill, D., Jones, S., & Hopper, R. (2006). *Talking, listening, learning: Effective talk in the primary classroom*. Berkshire: Open UNIVERSITY PRESS.

Piirainen-Marsh, A., & Tainio, L. (2009). Other-repetition as a resource for participation in the activity of playing a video game. *The Modern Language Journal, 93*(2), 153–169.

Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports, 6*, 207–231.

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). The interactional architecture of the language classroom: A conversation analysis perspective. *Language Learning, 54*(S1), 1–262.

Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an analysis of discourse*. Oxford: Oxford University Press.

Tsui, A. (1992). A functional description of questions. In M. Coulthard (Ed.), *Advances in spoken discourse analysis*. London: Routledge.

van Lier, L. (1988). What's wrong with classroom talk? *Prospects, 3*(3), 267–283.

Walsh, S. (2011). *Exploring classroom discourse*. Oxon: Routledge.

Walsh, S. (2013). *Classroom discourse and teacher development*. Edinburgh: Edinburgh Press.

White, J., & Lightbown, P. M. (1984). Asking and answering in ESL classes. *Canadian Modern Language Review, 40*(2), 228–244.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 1–24. https://doi.org/10.1093/applin/amp024.

Chapter 34
Assessing Teacher Discourse in a Pre-Service Spoken English Proficiency Test in Malta

Odette Vassallo, Daniel Xerri, and Larissa Jonk

Abstract This chapter discusses how the inadequacies of a general spoken English proficiency component for pre-service teachers were addressed by means of the design and implementation of a spoken proficiency test that incorporates teacher discourse as one of its assessment criteria. The assessment of teacher discourse is shown to be an appropriate means of addressing the operational needs of pre-service teachers of English. The chapter explains the rationale behind the test and the way it was designed. It also considers some of the changes and effects observed upon the implementation of the test in the English language teaching sector in Malta.

34.1 Introduction: Purpose and Testing Context

The Spoken English Proficiency Test for Teachers (SEPTT) was launched in Malta in 2017 in response to requests made by the country's English Language Teaching (ELT) industry. Malta is an archipelago in the Mediterranean Sea that gained independence in 1964. Prior to that it was a colony of the British Empire for 164 years. Its colonial heritage has left an impact on the country's linguistic landscape. English is one of Malta's two official languages, the other being Maltese. The vast majority of Maltese citizens are bilingual and both languages are taught in general education from a very young age. Most citizens consider themselves proficient in English (National Statistics Office [NSO] 2014, 2018a).

Capitalizing on the English language proficiency of the Maltese population, business owners started opening private language schools in the 1960s. The number of language schools in the country currently amounts to 38. These schools cater exclusively for the needs of around 87,000 foreign students who visit Malta to learn English every year (NSO 2018b). All of these schools are regulated by legislation that first came into effect in 1996 and was subsequently updated in 2015. This legislation has the purpose of ensuring high standards in the language teaching industry. Hence, besides the fact that all language teaching operations require a license, teachers and

O. Vassallo (✉) · D. Xerri · L. Jonk
University of Malta, Msida, Malta
e-mail: odette.vassallo@um.edu.mt

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_34

other personnel working in schools need to meet minimum requirements in terms of academic qualifications, age, and other criteria in order to be granted a permit to work in the sector.

In the case of teachers, permit applicants need to provide evidence of having satisfactorily completed a pre-service course on language teaching methodology and of having attained an appropriate level of language proficiency. The number of teachers presently working in private language schools amounts to more than 1,200, of which only 10% work on a full-time basis (NSO 2018b). These teachers are expected to have a high level of proficiency in English. Before the implementation of SEPTT, proficiency was mainly assessed by means of the Test for English Language Teachers (TELT), which incorporated a spoken component. Primarily, TELT operated as a language awareness test but also contained a general proficiency component designed by the ELT Council, the industry regulator.

This chapter describes the problem that led to the design of SEPTT and how this test aimed to address it by assessing candidates' spoken English language proficiency for the ELT context. The chapter discusses how the incorporation of teacher discourse as a key criterion in SEPTT is the main means by which this contextualized language proficiency is assessed.

34.2 Testing Problem Encountered

In spite of the existence of TELT's spoken component, in 2014 school owners, experienced teachers, and other stakeholders started raising concerns about the inadequate level of spoken English of newly employed teachers (Chetcuti 2014). While these concerns were reported in the media as having to do solely with the proficiency of foreign language teachers working in Malta, the ELT Council's consultation with various language testing experts found that the main problem consisted of the inadequacy of TELT's spoken component as a means of assessing the proficiency of both Maltese and foreign teachers. The spoken component was a means of assessing general spoken proficiency but as demonstrated by the literature this might not be sufficient for the operational needs of individuals working in a language teaching environment (Freeman et al. 2015; Van Canh and Renandya 2017). In fact, recommendations were made to include a "language improvement component in methodology courses" (Van Canh and Renandya 2017, p. 79), given that there is a need to incorporate "classroom" English proficiency. A thorough analysis of a series of classroom observations held at private language schools in Malta, showed that both newly qualified teachers and seasoned practitioners lacked awareness of the type of discourse that fits the classroom, which we refer to as "teacher discourse." Consequently, they were often unclear in their explanations, instructions, feedback, etc., and failed to offer a good language model for their learners. However, before elaborating on the concept of teacher discourse, we shall present our working definition of discourse.

Discourse, when interpreted broadly, refers to language in context (Markee 2015). It signals what is beyond the word or sentence by involving the context and encompassing shared knowledge, assumptions, intended meaning, and actual interpretation. Classroom discourse—also typically referred to as "teacher talk"—is associated with the classroom interaction that occurs in a social and institutional context with set norms that inform us on what we can do or say in a classroom. Classroom activity is a "socially constructed and negotiated activity" (Christie 2002, p. 161), which involves contrived interaction with a specific goal that is framed by the teacher and students to allow teaching and learning to occur. A narrower view of classroom discourse refers to the teacher–student interaction involving specific features of language, such as conversational frames and teacher directives where the teacher is perceived to give instructions and students are the recipients.

Most research on classroom discourse investigates the interactional processes in the language classroom and is concerned with both the discourse within and beyond the classroom and the social implications of this. Recently, the focus shifted to how language as a model for the learner is inextricably connected to the specialized classroom language.

A study conducted by Skinner (2016) reflects on feedback gathered from MA TESOL supervisors about trainees who seemed to have a limited understanding of teacher talk, since the latter often focussed on their own personal understanding of their use of classroom language instead of how appropriate their talk was vis-à-vis their pedagogical aim. Skinner (2016) analyzed the understanding of a number of ESL trainees, and her findings revealed that trainees' understanding of effective teacher talk varied. In reaction to these findings, Skinner (2016) proposes that "teacher talk should be recognized as a threshold concept and made explicit in teacher education curriculum" (p. 152).

Similarly, Walsh (2011) argues that few teacher education programs devote time to "developing understanding of interactional processes and the relationship between ways in which language is used to establish, develop and promote understanding" (p. 3). Walsh (2011) advocates for a strand in teacher education that includes classroom interaction in order to sensitize pre-service teachers to its importance and encourage a deeper understanding of the contexts they will be teaching in.

Walsh (2011) introduces the concept of Classroom Interactional Competence (CIC) to "promote understanding and facilitate professional development" (p. 1). He explains that in the second language classroom, students access knowledge, develop skills, negotiate meaning, and seek to clarify understanding through language in interaction. In a bid to promote this level of awareness in teacher education, Walsh designed the Self-Evaluation of Teacher Talk (SETT) framework which aims to foster teacher development through classroom interaction.

Four features of classroom discourse are highlighted by Walsh (2011): (1) control of the interaction; (2) speech modification; (3) elicitation; (4) repair. The first feature describes the teachers in a position of power in the classroom as they have control over such patterns of communication as turn taking and topic selection, among others. Spoken language is the second feature that teachers control when they adhere to the typical patterns of slowing down their speech, being louder, employing deliberate pausing and emphasis, as well as modeling their language to avoid the danger of learners getting lost while navigating classroom discourse. The latter is similar to the "grading language" as presented by Thornbury and Watkins (2007, p. 16). Teachers are expected to be in possession of a range of linguistic resources to facilitate comprehension and assist the learning process. Therefore, they are expected to be fully aware of their idiolect and reflect on whether it could benefit or hinder learners' understanding, and thus, adjust or adapt. Teachers should be able to use appropriate transitional and discourse markers to keep the classroom discourse whole, and they are expected to adopt a range of strategies in their discourse repertoire, including but not limited to repetition, reformulation, and backtracking. Through elicitation—the third feature—teachers are able to control and monitor discourse by asking questions such as "display questions," which offer a number of functions in the form of elicitation of responses, checking understanding, guiding learners, promoting involvement, and concept checking. The final feature refers to how teachers handle errors in the classroom and the choices they make in how and when they foreground these errors or delegate part of the responsibility to the learners (Walsh 2011, p. 19). Through an extensive explanation of these four main discourse features, Walsh (2011) highlights that understanding the discourse of the classroom is crucial because discourse is taught through the discourse of teachers.

Equally significant is Seedhouse's (2005) notion of "goal-oriented institutional discourse" (p. 171) when discussing Conversational Analysis (CA) and language learning. Teacher talk is a type of institutional interaction with an institutional goal in an institutional setting. Seedhouse (2005) explains that "CA presents competence as variable and co-constructed by participants in interaction" (p. 172).

In defining classroom discourse and teacher talk, we have established that a teacher's choice of language in the classroom has a clear pedagogical purpose that goes beyond the general language proficiency of a teacher. The specialized language which teachers are expected to adopt for effective communication and to facilitate learning is what makes classroom interaction effective. Thus, in designing the SEPTT test construct, we were mindful of this responsibility that a pre-service teacher should be cognisant of when engaging in classroom discourse, that is "teacher discourse." SEPTT serves as an exit test for teaching methodology courses and complements TELT. The tasks are designed to simulate a classroom context and candidates are encouraged to engage in classroom interaction. A detailed description of the teacher discourse criterion and the test material design are provided in the next section.

34.3 Solution/Resolution of the Problem

Given the emphasis placed on language teaching quality by the legal notice regulating the ELT industry (Government of Malta 2015), the ELT Council engaged in a consultation process with different stakeholders and decided to commission the Centre for English Language Proficiency (CELP) at the University of Malta to design a new test whose purpose would be that of assessing candidates' spoken English language proficiency for the ELT context. The design and implementation phases involved the participation of stakeholder representatives so as to ensure that this homegrown test would meet the needs of teachers, school owners, and other key roles within the ELT industry in Malta. SEPTT was made a legal requirement for all new teaching permit applicants, irrespective of their first language, nationality, or qualifications. The spoken component in TELT was removed and the latter continues to act solely as a language awareness test.

SEPTT is a 15-minute test that employs an examiner-to-candidate format (CELP 2017). This three-part test opens with an introductory interview in which candidates are asked about their interests, plans, and training with respect to ELT. By means of a two-way exchange, candidates answer a set number of questions aimed at assessing their ability to talk about familiar topics related to ELT. Part 2 consists of a long turn based on a prompt that candidates would have been asked to examine prior to entering the test room. The prompts used in this part of the test focus on particular aspects of an English language lesson, such as classroom management, the communication of content, and the setting up of lesson activities. The prompts help candidates to use language for the purposes of presenting, defining, developing, and exploring information related to specific language teaching scenarios. The prompts consist of detailed rubrics, as well as printed and visual elements. Part 3 acts as a conversation between the examiner and the candidate in which a scenario related to the prompt used in the previous part of the test is explored in further detail. Candidates are provided with a rubric and some time to examine it, after which they are asked a number of questions aimed at assessing their ability to give instructions and respond to the indicated scenario.

To enhance rater reliability, SEPTT is entirely scripted and every test is audio recorded. Since examiners act as interlocutors while also timing the three parts, initiating interaction, and rating candidates' performance, training plays a fundamental role. Examiners are not only trained to closely adhere to test procedures, but they are also periodically evaluated on their ability to use and interpret the analytic rating scale purposely developed for SEPTT.

The rating scale consists of five assessment criteria and twenty descriptors that correspond to four bands. The highest band is 4 and the lowest is 1. Band 3 is the boundary between operational and pre-operational candidates. Candidates wishing to apply for a teaching permit need to have a minimum of Band 3. The assessment criteria are: teacher discourse, coherence and cohesion, pronunciation, grammar, and vocabulary. While candidates are assigned a band for each one of these five criteria, the overall band for a candidate's performance in the test is determined by the lowest band they attain for any specific criterion. This is meant to ensure that candidates possess a satisfactory language level in all five criteria.

Teacher discourse is the criterion that makes SEPTT a useful tool in determining whether candidates have the desired level of spoken English proficiency needed for them to operate in a language teaching context. The test replicates classroom tasks and routines in order to simulate teacher discourse for candidates. Teacher discourse is a significant game-changing criterion in SEPTT because it sets this high-stakes language test apart from other general English proficiency tests. This is because it adopts an ESP-driven approach to language proficiency. SEPTT tests the language teachers are expected to adopt in classroom interaction. The test materials are modeled on teachers' use of English in classroom discourse. Basing the test construct, and more specifically the criterion teacher discourse, on Freeman et al.'s (2015) English-for-Teaching, the test content is designed to authentically represent tasks conducted by teachers in the language classroom. Embedded in the teacher discourse criterion is Freeman et al.'s (2015) notion that proficiencies are always situated in specific contexts and bound by a particular social practice. This ESP-driven approach is further substantiated by Van Canh and Renandya (2017) who stress that apart from being highly proficient in general English, ELT practitioners should also be "adept at using the language to create conducive learning environments" (p. 79).

SEPTT takes into consideration candidates' target language use context and refrains from testing their knowledge of pedagogy. It exploits that knowledge, which is acquired by pre-service teachers during teaching methodology courses, to elicit teacher discourse based on the activities determined by the test materials. The key elements in the teacher discourse criterion assess candidates' ability to speak at length on teaching-related topics, and equally assess candidates' range of discourse functions appropriate to the teaching context, including explaining, presenting information, giving instructions, and summarizing. SEPTT replicates classroom tasks and routines in order for candidates to engage in classroom discourse. These tasks reflect Freeman et al.'s (2015) grouping of similar tasks and routines into three functional areas: managing the classroom; understanding and communicating lesson content; and assessing students and giving them feedback. They are also linked to Walsh's (2011) four features of classroom discourse. The emphasis is on the use of a teacher-specific register that is key to pre-service education.

34.4 Insights Gained

As part of a feedback loop that exists through the ELT Council, reviewers of pre-service English language teaching methodology courses in Malta have reported some changes and SEPTT is now featuring in these courses. In teacher education programs, ab initio teachers are made to think about what tasks they would use, and about how and why they would adapt tasks or use different material altogether. Prior to SEPTT, there was less of a focus for participants to talk about this during teacher education courses. Thus, it is likely that SEPTT test material has encouraged teacher educators preparing candidates for the exam to practice thinking aloud and talking about what they would do with teaching materials and how they would use them.

Since the introduction of SEPTT, participants on pre-service teacher education courses are made to give more attention to the actual language they would use in class; for example, the language used when giving instructions and checking whether learners have understood instructions. Teachers are made to focus on breaking down language to ensure that there is repetition and reformulation in order to ensure clarification.

Observations made by SEPTT examiners reflect on the downside to candidates being "prepared" for the test in pre-service courses. Some candidates seem to have adopted patterns of formulaic language and their delivery may be perceived as scripted. Candidates who have received such training know that giving instructions and checking instructions may feature in the test. Therefore, task modification may be necessary to broaden the range of actual language use candidates have to consider.

Contrary to the language school owners, some of the teacher educators responsible for pre-service courses offered some initial resistance to the launch of SEPTT. However, through informal interviews, they have recently acknowledged and commented on the test's positive washback effect as it has encouraged pre-service teachers to focus more on their own language and the level of appropriacy in relation to pedagogical aims. Teachers' attention has shifted to teaching concepts and the actual language they would use in class, which has enriched teacher education courses.

Concurrently, SEPTT examiners have reported that when the test was first introduced, candidates struggled with using language appropriate to classroom discourse during the individual turn in the second part of the test. However, with teacher educators giving it more importance in pre-service methodology courses, SEPTT examiners who have been examining from the launch of the test have observed that candidates are generally better prepared to talk about the material for the required length of time in Part 2 of the test now that it has been in place for over a year.

34.5 Conclusion: Implications for Test Users

In retrospect, there was clearly a need to incorporate a shift in the speaking focus to teacher discourse. With SEPTT candidates demonstrating that generally they are more able to speak at length about a teaching context by demonstrating an awareness of the type of classroom interaction expected for learning to be effective, this lacuna has been addressed.

SEPTT was designed to create an ESP-oriented context within which to test pre-service teacher's language. As a result, it has also reinforced teacher education courses by adding a speaking focus to the topic of teaching. This has helped to enhance the value of training and has allowed for better prepared trainees because they are also made to articulate their thoughts on teaching as well as practice "actual" classroom language.

The value that SEPTT gives to teacher discourse means that pre-service teachers and other test users are provided with a more authentic measure of the spoken proficiency required by professionals operating in a language teaching environment. Teacher discourse is sometimes overlooked during pre-service teacher education, and trainees are expected to develop this competence while on the job. By foregrounding the importance of teacher discourse, SEPTT ensures that teacher education focuses on the specific linguistic needs of teachers when they interact with learners in the classroom. Preliminary findings demonstrate that since SEPTT's introduction the main washback effect has been that pre-service teacher educators are focusing on enhancing the quality of teacher talk as an essential part of classroom discourse (Vassallo et al. 2017). It is postulated that this will be of benefit to learners because they are provided with better language role models. Teachers are also likely to benefit given that the test is enabling them to operate more effectively in the classroom with regard to their use of the language required for teaching purposes. Employers, school owners, and the rest of the language teaching industry can also rest assured that the high and equitable standard set by SEPTT will continue to fuel the growth of a sector whose competitive advantage resides in offering a quality service to the thousands of learners who visit Malta annually. SEPTT illustrates how the incorporation of teacher discourse as a criterion in the assessment of pre-service teachers can help to improve the quality of language education.

The way forward is two-pronged as it will consist of an evidence-based internal modification of the test, and an outward reflection with a focus on a positive washback effect. In line with our inward-looking changes, we shall seek to generate different test materials that give opportunities of simulated classroom interaction for candidates to avoid perceived effects of test targeted preparation (see Farnsworth 2013) and thus encourage a deeper understanding of teacher discourse. Compounding this, a spoken corpus is being constructed and this will allow us to analyse teacher discourse in more detail. As an outward-looking measure, we shall engage in a reflection on how to make teacher discourse resonate beyond the test. One of the main desirable washback effects is for pre-service teachers to understand that they should use language that facilitates the learning process, and a means by which this could be achieved is through recording and self-reflection during teaching practice. Perhaps this would allow teacher educators to focus more on encouraging trainees to talk about teacher topics, such as what teachers would do in the classroom in given situations, how they would use or adapt material, and why they would choose to incorporate certain tasks or use certain material. The recording would follow such reflection on classroom discourse.

Another aspect that pre-service teachers may benefit from is reference to classroom discourse that goes beyond the choice of language targeting proficiency. ELT professionals encounter multiple nationalities in private language schools, which implies that they are immersed in a multicultural context. Thus, a broader understanding of classroom discourse could be introduced in pre-service training programs to learn to address the diverse cross-cultural communication and those cultural patterns that influence students' learning patterns. According to Rymes (2016), active discourse inquiry improves student academic achievement. Such classroom discourse analysis allows them to focus on the "communicative repertoire" that echoes the students' communicative diversity.

References

CELP. (2017). *Spoken English proficiency test for teachers (SEPTT)*. Msida: University of Malta. https://eltcouncil.gov.mt/en/Documents/SEPTT/SEPTT%20Manual.pdf. Accessed 13 July, 2019.

Chetcuti, K. (2014). Language school jobs 'go to overseas tutors'. *Times of Malta*. https://www.timesofmalta.com/articles/view/20140520/local/Language-school-jobs-go-to-overseas-tutors-.519695. Accessed 13 July, 2019.

Christie, F. (2002). *Classroom discourse analysis: A functional perspective*. London: Continuum.

Farnsworth, T. (2013). Effects of targeted test preparation on scores of two tests of oral English as a second language. *TESOL Quarterly, 47*(1), 148–156. https://doi.org/10.1002/tesq.75.

Freeman, D., Katz, A., Garcia Gomez, P., & Burns, A. (2015). English-for-teaching: Rethinking teacher proficiency in the classroom. *ELT Journal, 69*(2), 129–139. https://doi.org/10.1093/elt/ccu074.

Government of Malta. (2015). *L.N. 221 of 2015: English Language Teaching Council regulations, 2015*. Malta: Government of Malta. https://eltcouncil.gov.mt/en/5th%20Conference%20Pictures/ELT%20Council%20Regulations.pdf. Accessed 13 July, 2019.

Markee, N. (Ed.). (2015). *The handbook of classroom discourse and interactional*. Chichester: Wiley Blackwell.

NSO. (2014). *Census of population and housing 2011: Final report*. Valletta: NSO. https://nso.gov.mt/en/publicatons/Publications_by_Unit/Documents/01_Methodology_and_Research/Census2011_FinalReport.pdf. Accessed 13 July, 2019.

NSO. (2018a). *Adult education survey: 2016*. Valletta: NSO. https://nso.gov.mt/en/publicatons/Publications_by_Unit/Documents/C4_Education_and_Information_Society_Statistics/AES_publication.pdf. Accessed 13 July, 2019.

NSO. (2018b). *Teaching English as a foreign language: 2017*. Valletta: NSO. https://nso.gov.mt/en/News_Releases/View_by_Unit/Unit_C4/Documents/News2018_042.pdf. Accessed 13 July, 2019.

Rymes, B. (2016). *Classroom discourse analysis: A tool for critical reflection* (2nd ed.). London: Routledge.

Seedhouse, P. (2005). Conversation analysis and language learning. *Language Teaching, 38,* 165–187. https://doi.org/10.1017/S0261444805003010.

Skinner, B. (2016). Effective teacher talk: A threshold concept in TESOL. *ELT Journal, 71*(2), 150–159. https://doi.org/10.1093/elt/ccwo62.

Thornbury, S., & Watkins, P. (2007). *The CELTA course: Trainee book*. Cambridge: Cambridge University Press.

Van Canh, L., & Renandya, W. A. (2017). Teachers' English proficiency and classroom language use: A conversation analysis study. *RELC Journal, 48*(1), 67–81. https://doi.org/10.1177/0033688217690935.

Vassallo, O., Xerri, D., & Grech, S. (2017). Testing pre-service teachers' spoken English proficiency. In E. Gutiérrez Eugenio (Ed.), *Learning and assessment: Making the connections* (pp. 39–45). Cambridge: ALTE.

Walsh, S. (2011). *Exploring classroom discourse: Language in action*. London: Routledge.

Chapter 35
High-Stakes Test Preparation in Iran: The Interplay of Pedagogy, Test Content, and Context

Shahrzad Saif

Abstract Previous work on high-stakes test preparation (henceforth TP) in English as a Foreign Language (EFL) contexts is sparse and has mostly focused on teachers' perceptions of test influence on the content and outcome of preparation courses linked to them. Certain studies, however, show that context-specific elements, such as stakeholders' perceptions and social/political realities of the setting, equally influence TP. This implies that a high-stakes test used in different contexts could potentially lead to different TP practices. Adopting a qualitative approach, this study investigates the nature of language instruction in an International English Language Testing System (IELTS) preparation center in Iran, a context where high-stakes TP is widely practiced but whose nature is rarely studied. Research questions address the nature of TP practices and how it relates to the test content in this context, as well as the stakeholders' perceptions of the test and their effects on the choice of instructional activities, content, methods, and strategies in TP courses. Data were gathered through questionnaires, interviews, observations, and focus-group interviews in ten-week-long IELTS preparation courses offered in a major TP center. A total of 56 test takers, 6 teachers, and 3 test center administrators participated in the study. The results, analyzed qualitatively and triangulated through cross-verification, point to the test center and its culture shaping the orientation of TP courses. Whereas the focus of TP is found to be on the test demands, instructional practices go beyond test-inspired activities, reflecting certain contextual factors such as students' goals and needs, teachers' experience, belief in second language (L2) learning, and stakeholders' awareness of learners' needs.

S. Saif (✉)
Université Laval, Québec City, QC, Canada
e-mail: shahrzad.saif@lli.ulaval.ca

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_35

35.1 Introduction: Purpose and Testing Context

As part of a larger multi-phase, multi-context investigation into preparation for high-stakes English language tests, this study explores the nature of test preparation (TP) practices in the EFL context of Iran. The study investigates teachers' instructional practices, their relationship to the test content, learners' experiences, and a range of context-specific factors shaping the stakeholders' (administrators, teachers, students) perceptions in a specific TP center.

English is taught as a foreign language in Iran, rarely used for everyday communication. However, learning English is widely popular in Iran; it is taught at all levels in private schools and after the elementary level in public schools. University students take English as a mandatory subject and regularly use English textbooks or online materials. Young Iranians are motivated to learn English to succeed in the highly competitive university entrance examination. The recent decade, however, has witnessed an unprecedented rise in enrollments for private English schools due to the need to pass one of the high-stakes standardized English language tests, such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS), to qualify for immigration to English-speaking countries, or for admission to foreign universities. Among the standardized tests, IELTS is by far the most widely used in Iran, mainly due to the political tensions with the USA, where the TOEFL is developed, not to mention that the IELTS score is increasingly accepted by academic institutions and governments around the world for educational or immigration purposes. Iran is currently among the top 25 countries with the highest number of test takers; IELTS Annual Review (British Council 2003) ranked Iran as the 7th for the general version and the 16th for the academic version. Since then, IELTS test centers have opened up in major Iranian cities, making it considerably easier for test takers to write the test. This surge in candidacy has naturally led to a growing number of for-profit TP centers offering IELTS preparation courses across the country. Given the important consequences of IELTS scores for candidates, and the life-altering nature of the decisions made based on them, it is important to gain insight into the factors that directly or indirectly influence the instructional activities in this context.

The specific context of this study is an authorized TP center with branches in major Iranian cities.

35.2 Testing Problem Encountered

Early studies of preparation for educational high-stakes tests were conducted out of concern that measurement-driven instruction could adversely affect the equity and fairness of the decisions and inferences made about people based on their test performance (Madaus 1988; Mehrens and Kaminski 1989; Messick 1982). In language testing research, high-stakes TP is often studied as part of the investigation of *test* *washback* (Alderson and Hamp-Lyons 1996; Cheng 2005; Green 2007; Hayes and Read 2004; Matoush and Fu 2012; Saif 2012; Wall and Horák 2011, among others), generally defined as the effects of testing on language teaching and learning. Results have overwhelmingly shown that high-stakes tests—whose scores are the basis for important decisions being made—influence teaching and learning practices, and have real or perceived consequences for test takers. However, previous research has also shown that the relationship between high-stakes tests and the teaching/learning activities linked to them is complex, by no means direct or predictable (Messick 1996; Saif 2006; Wall and Alderson 1993, among others). There are factors, in addition to the content and stakes of the test, that influence instruction in TP courses (Yu et al. 2017). Context-specific elements ranging from teachers' beliefs and their role as agents of change (Chappell et al. 2015; Irving and Mullock 2006; Mickan and Motteram 2008), to political, educational, and social realities of the context where the test is used (Cheng 2005; Muñoz and Álvarez 2010; Yu et al. 2017) have all been identified as factors influencing TP practices. Yet, with the exception of a few studies conducted in English as a Second Language (ESL) contexts (Alderson and Hamp-Lyons 1996; Mickan and Motteram 2008), descriptive studies of the instructional activities in high-stakes language TP courses are lacking, as are investigations of multiple stakeholders' perceptions of the content and nature of TP in EFL contexts like Iran.

The following research questions are addressed in this study:

Research question 1: What are the stakeholders' (administrators, teachers, students) perceptions of the IELTS in Iran?

Research question 2: How do these perceptions affect instructional practices (choice of teaching activities, content, methods, strategies) in TP courses?

Research question 3: How does the content of instruction relate to the test content?

35.3 Review of Literature

Systematic investigation into high-stakes TP courses dates back four decades (Madaus 1988; Mehrens and Kaminski 1989; Messick 1982; Miyasaka 2000; Smith 1991, among others). Messick (1982) defines TP or *coaching* as an attempt aimed at improving test scores. Messick hypothesizes that, depending on the content and nature of the preparation activities, TP could be both beneficial and harmful to the validity of the test scores. Madaus (1988), however, labels TP as *measurement-driven instruction* and discusses its consequences in terms of five principles. He portrays the power of tests as a *perceptual phenomenon* (p. 35) shaping the test consequences for teaching and learning. Like Messick (1982), Madaus highlights the impact of test questions on the content of preparation courses and cautions against item type, rather than the skill or objective, driving instruction. Also focusing on TP content, Mehrens and Kaminski (1989) argue that the use of materials, built around the actual test, in TP classrooms jeopardizes the generalizability of the score interpretations.

In language testing, amid the increasing demand for high-stakes TP worldwide, research into how teachers and test takers prepare for high-stakes tests, and the effects of TP practices on test takers' performance, has gained prominence. This investigation is often conducted in the context of washback studies, and like the educational research discussed above, has focused on the content of the TP courses. Three major studies initially conceptualized the mechanism through which washback operates by specifying the factors in the educational context that interact with the test and with each other.

Alderson and Wall (1993) characterize washback by proposing 15 hypotheses, ranging from the most general to the most specific. They argue that any research into washback should specify the nature and the expected effects of the test, and consider the context where the test is used and the decisions made based on its scores. Hughes (1993), however, distinguishes between three different bases for investigating test washback: the *participants* (students, teachers, administrators, materials developers/publishers), the *process* (actions taken by the participants which may eventually lead to learning), and the *product* (outcome and quality of learning). Hughes maintains that by affecting the perceptions of the *participants*, a test can potentially affect the *process* and the *product* of learning and thus promote the intended effects. This position, of course, implies that the stakeholders' perceptions in a given context could very well remain unaffected by the test and therefore impede the intended test influence. He highlights a number of factors, such as test stakes, teachers' desire for student success, familiarity with the test's content, availability of resources, and teachers' qualifications, that could directly interfere with test washback. Hughes's three major categories are also represented in Bailey's model of washback (1996) that distinguishes between test influence on the test takers and other stakeholders (teachers, administrators, curriculum developers). She refers to the former as *washback to the learners* and the latter as *washback to the program.* She, too, highlights context as an important factor in investigating washback.

Even though empirical research in this area over the past two decades has resulted in new expanded models of washback that include and/or specify several other contributing factors (Green 2007; Saif 2006; Shih 2009, among others), the three areas—participants, process, product—identified by Hughes (1993) and highlighted by Bailey (1996) remain the core areas investigated by washback studies (Cheng et al. 2004; Hayes and Read 2004; Matoush an Fu 2012; Mickan and Motteram 2008; Saif 2012; Wall and Horák 2011) in the setting of TP courses. They also entail the aspects of TP elaborated by Messick (1982), Madaus (1988), and Mehrens and Kaminski (1989).

This study focuses on the *participants* and the *process* as conceptualized by Hughes. In particular, it explores the perceptions of the participants (administrators, teachers, students), the nature and content of the instructional activities, and the context-specific factors (Alderson and Hamp-Lyons 1996) influencing the participants' perceptions and classroom behavior.

35.4 Methodology

This study adopts a case study approach involving qualitative data gathered from multiple sources of information (test center administrators, ESL teachers, students) within a major IELTS TP center. Yin (2014) defines case study as "an empirical inquiry that investigates a contemporary phenomenon (the "case") in depth and within its real-world context" (p. 16). Using a *triangulation of sources* method (Patton 2002), the study compares qualitative data collected through a variety of instruments (observations, interviews, focus groups, questionnaire) to gain insight into the TP practices at this center in Iran.

35.4.1 Participants

A total of 56 students enrolled in two IELTS preparation courses, 6 ESL teachers, and 3 test center administrators (Chief Executive Officer [CEO], General Director, Academic Advisor) participated in the study. Table 35.1 summarizes the demographic characteristics of the participants.

35.4.2 Instruments and Procedure

Information about the context of the study, the stakeholders' perceptions of the test, and the TP process was gathered through five different instruments (Appendices 1–5): Student questionnaire, teacher and administrator interviews, classroom observations, and student focus-group interviews. The data were collected before, during, and after the 10-week-long IELTS preparation courses in the following order.

35.4.2.1 Before the Course

Test center administrators' feedback was sought in one-to-one interviews before the course. Semi-structured interviews (Appendix 1), conducted in Farsi with occasional use of English, were entirely audio-recorded and subsequently transcribed verbatim.

Table 35.1 Participants' demographic characteristics

	N	Gender		Age range			Teaching experience (years)		Highest degree completed	Years of English	L1
		M	F	20s	30s	40s	EFL	IELTS			
Students	56	24 43% NI*** = 8 (14%)	24 43%	38 68%	18 32%	0	NA*	NA*	HS** (5%) BA (36%) MA (36%) PhD (7%) NI (16%)	2–5 (46%), 6–10 (20%), 10 + (4%) NI (30%)	Farsi
Teachers	6	4	2	0	2	4	8–15	4–6	5 MA 1 BA	NA*	Farsi
Administrators	3	3	0	0	1	2	6–20	0–8	1 PhD 1 BA 1 MBA	NA*	Farsi

*Not applicable
**High school diploma
***No information provided

35.4.2.2 During the Course

In the early weeks, teachers were interviewed (Appendix 2) and a questionnaire was administered to students (Appendix 3). The questionnaire and the interview were designed to capture the participants' views with respect to the aspects of TP frequently discussed in previous studies, specifically the issues raised by Alderson and Hamp-Lyons (1996), Hughes (1993), Madaus (1988), Mehrens and Kaminski (1989), and Messick (1982). The student questionnaire was developed in English and translated into Farsi. The translated version was reviewed by two native Farsi-speaking teachers and three graduate students for clarity and accuracy. To ensure the respondents completely understood the questions, the Farsi version was administered to the students. Like the administrators', the teachers' interviews were conducted in Farsi (and in English where appropriate) and were audio-recorded for subsequent analysis.

During the 10-week course, data were also gathered at specific times (weeks 1, 5, 9) through audio-recorded classroom observations for a total of 18 hours. The observation protocol (Saif et al. 2019, Appendix 4) was an adapted version of COLT (Spada and Fromlich 1995).

35.4.2.3 After the Course

Upon the course completion, using convenience sampling, two focus groups of 8–10 students were formed with the students who in their questionnaires had said "yes" to a follow-up 75-minute interview. The interviews, conducted in Farsi and audio-recorded, focused mainly on their reaction to the TP course and the degree to which their expectations were met (Appendix 5). They were also asked to comment on their study habits and choice of supplementary practice materials.

35.4.3 Data Analysis

Data analysis was conducted in accordance with Patton's (2002) triangulation strategies and Lewis and Ritchie's (2003) guidelines for qualitative analysis of data. To ensure correct identification, categorization, and naming of the phenomena identified through various instruments, the two research assistants who had interacted with the participants during the data collection carried out the analyses. Observation data were compiled and labeled using the components of the coding sheet and the observer's notes. In cases of missing or ambiguous information, recorded audios were consulted.

The analysis of the questionnaire and interview data was guided by the research questions. In keeping with the study design (Lewis and Ritchie 2003; Patton 2002), for each group of participants, the information from their interview or questionnaire served as the point of departure for presenting the results, which were then cross-checked against—and presented side by side with—the corresponding information from other stakeholders' and observation accounts (Appendices 6, 7). The findings were categorized and labeled in terms of the threads common to the multiple data sources: The nature of instruction (activities, content, material, method), the relationship between the teaching content and the test, stakeholders' perceptions of the test and their effects on TP activities, and the contextual factors shaping stakeholders' perceptions. The findings are presented with respect to these elements for each group of stakeholders.

35.5 Findings

35.5.1 Test Center: Context and Administration

The test center where the study takes place is a well-known private institute offering GE and TP courses to adult EFL learners. To better understand the context where IELTS preparation takes place, the dynamics between different players at the center, and its possible impact on the content of TP courses, one-to-one interviews were conducted with the CEO (MBA), the General Director (BA, English Literature), and the Academic Consultant (PhD, TESL).

The respondents, who have years of experience teaching GE courses, expressed high confidence in their professional qualifications and underlined the direct relationship between their expertise and their role in the center's day-to-day operations. The nature and extent of their familiarity with the IELTS and TP courses differ with their responsibilities. The CEO has never taught TP courses but is well informed about the test's stakes, its global standing, and the students' desire for training for the test. The director of language courses and the academic consultant, however, have taught TP courses and possess intimate knowledge of the IELTS content, tasks, and constructs.

They identified the society's interest in language learning and the increasing demand for IELTS preparation as the foremost motivation behind establishing the center and referred to the ongoing competition between private language institutes as an incentive for offering and expanding TP services. However, they all noted that the center's main mission is to promote L2 learning in the society and help the students achieve their educational goals.

According to the general director, the center adopts rigorous placement practices to ensure homogeneity in TP courses. The officials even offer personalized advice, and, despite the center being for-profit, go as far as referring the students to other private institutes if their specific needs cannot be addressed here. The academic consultant underlined the center's preoccupation with improving students' *language abilities,* even though certain among them only care about passing the test. The CEO credited the center's progressive approach, its preoccupation with the students' needs, and the competent instructors for its success and excellent reputation. The center, according to the director and academic consultant, has gained publicity through word of mouth because of getting results and high academic quality. This sentiment was echoed by the students who, in their questionnaires and focus-group interviews, mentioned *highly qualified teachers*, followed by *high success rate*, *use of original IELTS materials* and *well-organized courses*, as main reasons why they chose this center.

As for *how* the management achieves and maintains academic quality, the respondents detailed the center's priorities and decision-making processes as summarized in Table 35.2.

The rigorous hiring and evaluation process for teachers explains why students, in their questionnaires, and interviews conducted after the training, commented positively about teachers' competence.

Finally, the management team was concerned with enforcing and maintaining high standards of teaching in TP courses in the face of shortages and learners' high, and at times, unrealistic expectations. According to the CEO, maintaining the quality of service and teaching is the center's highest priority because the students pay for their training and should be kept satisfied. The center's director and academic consultant, however, expressed frustration over keeping the balance between what the students want (passing IELTS) and what they really need (language training), which, according to the center's director, could create a gap between center's academic goals and what really happens in TP classrooms.

35.5.2 Teachers and Features of TP Instruction

All TP instructors ($N = 6$) participated in the study (see Table 35.1 for their demographic characteristics). Interview results—analyzed and categorized in terms of teachers' qualifications, test familiarity, teaching methodology, course content, instructional activities, evaluation method(s), and the strategies they promoted in class—were cross-checked with information from students' questionnaire/group interviews and observations that targeted these same elements (Appendix 6).

The teachers expressed satisfaction with the learning outcomes of their courses. They considered students' motivation as an important factor for success and characterized them as focused and goal-oriented, investing much time, energy, and money to complete the program. According to teachers, compared with GE courses, TP costs considerably more and has a lower dropout rate. Teachers' characterization

Table 35.2 Administrators' accounts of the decision-making processes

TP instruction	Source		
	CEO	General director	Academic consultant
TP content	• Determined based on students' needs	• Uniform materials; commercially available IELTS material • Supplementary materials by teachers based on students' needs	• For lower-level students, language abilities rather than direct IELTS strategies promoted • TP activities geared to "language learning" in low proficiency students
Teachers' role	• Most important players in center • Regularly consulted • Center keeps the teachers motivated and up-to-date	• 100% involved; no decision made individually	• Teachers consulted but administration makes final decision as teachers may change
Teachers' hiring	• Selected based on proficiency level, expertise, teaching skills, willingness to participate in workshops/professional development sessions	• SAME comments as CEO • Teaching demonstration (IELTS preparation courses)	• Same as CEO • Familiarity with IELTS commercial materials; job interview, teaching demonstration
Choice of material	• Original IELTS material (first piloted, adopted if proven effective); teachers consulted for supplementary materials	• Acts as liaison between teachers/students and management; reviews different commercially available TP materials	• Administrators and teachers discuss options before any final decision
Quality control	• Course and material evaluations by students and teachers, teacher feedback, teacher evaluation by students, class observations by the director, individual reports/comments by students	• Same as CEO • 100% control over TP activities	• Same as CEO • Administration views TP quality crucial to attracting clientele

(continued)

Table 35.2 (continued)

TP instruction	Source		
	CEO	General director	Academic consultant
Student in TP courses	• Not all TP students take IELTS soon after the course; center considers students' needs/goals (passing IELTS, improving proficiency); students planning to take IELTS motivated, benefit from practice tests	• TP students more advanced, evaluated more often • Frequent IELTS-based quizzes/more training for beginners	• Placement interviews; those planning to take IELTS have high expectations even if their GE level is inadequate for TP • TP students older than non-TP students, work experience

of learners, also confirmed by observations, explains the center administrators' previously mentioned preoccupation with high academic standards in TP courses.

The teachers also credited their approach to teaching language abilities and the variety of "test-based" activities for the success of the TP courses. Four teachers teaching both TP and GE courses commented that their approaches in the two courses were different. They listed "teaching micro-abilities" (grammar, vocabulary, morphology), "pair work based on topics of interest," "text analysis," "note taking" as examples of activities in GE courses. For TP, however, they focused on test-taking techniques and activities that were "exam-based" such as pair work practicing tasks similar to test tasks, and writing activities inspired by the test's themes and topics. During a 3-hour observation session in week 5, for example, of 17 teaching activities, 16 were IELTS related with 13 direct references to the test and 4 timed in-class assessments that closely simulated IELTS situations; these echoed the activities the students repeatedly ranked as most important.

Teachers further mentioned their classroom practices were influenced by students' "need to improve their language proficiency," particularly the spoken language, to which the students do not have exposure outside class. Observations supported this claim. Nevertheless, the students in their questionnaires/interviews expressed concern over the lack of sufficient oral practice in class. Observations showed that TP teaching was communicatively oriented with task-based activities and strategy training. In their interviews, the teachers had already hinted at the reason behind adopting a task-based approach for TP by commenting that "cramming" worked well for non-IELTS courses but not for TP courses. The observations, however, showed that, despite their claim to communicative teaching, the teachers, when necessary, resorted to other methods, like explicit grammar teaching, to get the point across. Their methodology was, therefore, eclectic (deductive explanation of grammar, repetition drills, communicative). As for the materials, teacher interviews and class observations confirmed the administration's claim to the use of uniform, 100% test-related materials for TP.

35.5.3 Students' Perceptions

The data from the student questionnaires and focus-group interviews (Appendix 7) point to an informed, highly motivated student population in TP courses. Compared with those enrolled in non-IELTS courses, the TP clientele are older with higher education levels (79% have at least a B.A.). The results show that the top reasons behind enrolment in TP courses are "study abroad" followed by "immigration" although 37% of the respondents only mentioned "learning English" as their primary motive. Focus-group participants further confirmed that motivation to "pass the test" and "master English" drives their learning behavior in TP courses. This is why, they said, as part of their out-of-class learning, they use IELTS tutorial/practice websites and writing samples. They unanimously agreed that TP courses helped improve their GE proficiency, whereas the reverse does not hold true. This was supported by 63% of the questionnaire respondents who had previously taken GE courses.

The students consistently highlighted "test-taking techniques" and "proficiency in four skills" as abilities they expect to improve by TP. They ranked "speaking" followed by "writing" as the most important abilities, and ranked "oral communication activities" as the single most important activity they want to be practiced in class. Thinking about their current TP courses, all learners (100%) approved of teachers' use of L2, although 36% among them would support the use of Farsi, when necessary, to ensure comprehension. Observations, however, confirm that teachers stuck to L2 use at all times. Also, in focus-group interviews, conducted after the course, the students expressed satisfaction with teachers' methodology, an outcome that mostly conformed to students' expectations expressed before the course through questionnaires. As for potential learning outcomes, the students (91%) perceived the type of exposure and training they received in IELTS preparation courses as different from that of GE courses. They (86%) expressed that the learning effects of TP go beyond passing the test and improve the general proficiency.

35.6 Insights Gained

Regarding the stakeholders' perceptions of the test (RQ1), what emerges from the results reported above is a rigorous TP practice involving a whole host of activities, strategies, materials, resources, and individuals in this particular context. The findings reveal a concerted effort by all parties involved (students, teachers, center administrators) to achieve *success* in IELTS.

The TP student participants were mature, self-described as highly motivated, and perceived the test score as consequential to their futures. Most of them, who need the IELTS score for admission to a foreign university, choose to prepare for and take the test in Iran rather than abroad because of the cost attached to it, given the elevated exchange rates in Iran. Administrators' perceptions of the test and its importance are, therefore, shaped both directly, through their own familiarity with the test tasks, content, and international standing, and indirectly, through awareness of the students' perceived goals and need to prepare for the test. Students' *goal-orientedness* is clearly the focal point of the center's TP activities and the driving force behind the center's culture of disciplined and impersonal approach to TP. Despite the center being for-profit, the evidence shows a strong research-based approach to TP and an administration sensitive to clients' needs. The center's commitment to hiring highly qualified teachers with intimate knowledge of the IELTS, in part in reaction to the students' expectations, has resulted in a homogeneous group of professionals who work in harmony and play a positive and important role in the center's decision-making. It is, therefore, not surprising that teachers have similar perceptions of the test, and an in-depth knowledge of its content. Their teaching practices are mainly inspired by the test. This, combined with the students' positive attitude toward L2 and motivation to pass IELTS, could translate to successful TP in this particular context with potential positive consequences for improved proficiency.

As for the nature of instruction in TP courses (RQ2) and its relationship to the test content (RQ3), the results indicate a variety of class activities practicing the four abilities measured by the test, and heavily influenced by the test content and format. Classroom observation notes from Week 1 show the teacher announced that the course was skills-based, and before introducing the skill-focused activities (reading, listening) provided an orientation to the corresponding component of the test. TP activities (as reported by teachers and supported by observations) are diverse and range from test-taking techniques, timed-assessments, practice tests, strategy teaching, and skill-focused activities to teaching and practicing individual grammatical items, all the while using the IELTS materials. Classroom observations present a snapshot of how in real time the teachers and students engage in such activities and how much time and weight is allocated to each activity. Class activities were chosen by the teacher, were teacher-led, and involved detailed modeling and explanations of the tasks. In-class assignments and homework were mostly individual work, always test directed and based on the skills practiced in class. The students participated in class activities, rarely missed a class, and completed the assigned tasks with interest. Observations also showed the teachers gave equal attention to practicing the skills the test measures. The students, however, disagreed; they expected teachers to prioritize *speaking* in class since they have no out-of-class exposure to English in Iran.

Respecting teachers' methodology, the data as a whole reveals a complex pattern. The teachers reported "communicative" as their dominant methodology when practising different tasks measured by the test. They, however, characterized their approach as *eclectic,* influenced not only by the test content, but also by students' learning objectives. The observations strongly support this assertion, revealing that the activities were mostly communicatively oriented, interactional where necessary, and filled with direct strategy training and advice on test-taking techniques for each skill. The transcripts from a Week 1, 180-minute session, for example, document 15 episodes, 11 of which are test-related activities during which the teacher made direct references to the IELTS content (format, topic, text length, time constraint) and promoted test-taking techniques. Note episode 12 from Week 1 (class 2).

Teacher: *Do you know what the speaking section of the test is like? how much time you'll have? how to get a good score?*

-He played a video (Thompson Exam Essentials) about IELTS Speaking and checked if the students understood it; asked them questions about the key information in the video.

Teacher: *Open your books, page 139, the speaking section. We'll do a chart on familiar topics (food, hometown, hobbies, etc.). Brainstorm for 2 minutes, then write 2 "wh" questions for each topic.*

- He circulated in class and gave them hints (use present tense) and referred to forms commonly used in the IELTS. He explicitly explained a grammatical/usage point (*which* vs. *what)* to the whole class as he noticed it was a common problem.

Teacher: *Now form groups of two and ask each other the questions you've prepared. You've 8 minutes for this task. Remember what you heard in the video; fluency is very important for this part of the test. You need to pay attention to your pronunciation, intonation, choice of words, grammar, ...*

-He circulated in class, listened to students' exchanges and took notes.
-After the task, he gave feedback: underlined the problems with intonation, pronunciation, grammar, and word choice, commented on students' use of facial expressions and eye contact (or lack thereof); asked them to clarify their ideas.
-He asked the students to redo the activity in light of his comments.

Observations also point to teachers' repeated use of deductive explanation of grammar and repetition drills where necessary:

-The teacher talked about "emphatic structures." He put it on the board and asked the students to copy it
What+subject+verb+to be+clause/phrase→What I enjoyed was the spectacular view of the mountain. (Week 5, episode 2)

Note that the students, in their focus-group interviews, expressed satisfaction with the teachers' methodology.

Collectively, the findings portray the IELTS preparation as a collaborative effort in this center. They point to a positive dynamic between the stakeholders, and an in-depth understanding, by teachers and administrators, of students' goals and educational needs. The data clearly show that this awareness shapes the administrators' decision-making, and in turn, the teachers' pedagogy and methodology. At the same time, the findings reveal that satisfying students' *perceived* needs is not done at the expense of the quality of instruction. Note, for example, the difference between teachers/administrator's and students' views regarding the students' preparedness for TP. According to the general manager and teachers, many students who apply for TP courses are discouraged from taking the course if they do not meet the required proficiency level for TP. Also, the language of instruction in class is L2 in spite of the difficulty it poses for certain students.

35.7 Conclusion: Implications for Test Users

Certain conclusions are drawn based on this study: In general, the findings characterize TP in this EFL context as a complex activity that involves multiple actors working in harmony. It requires awareness of the high-stakes test's content and format, material selection, course planning, choice of appropriate teaching methodologies, activities, strategies, and familiarity with students' goals, needs, and interests.

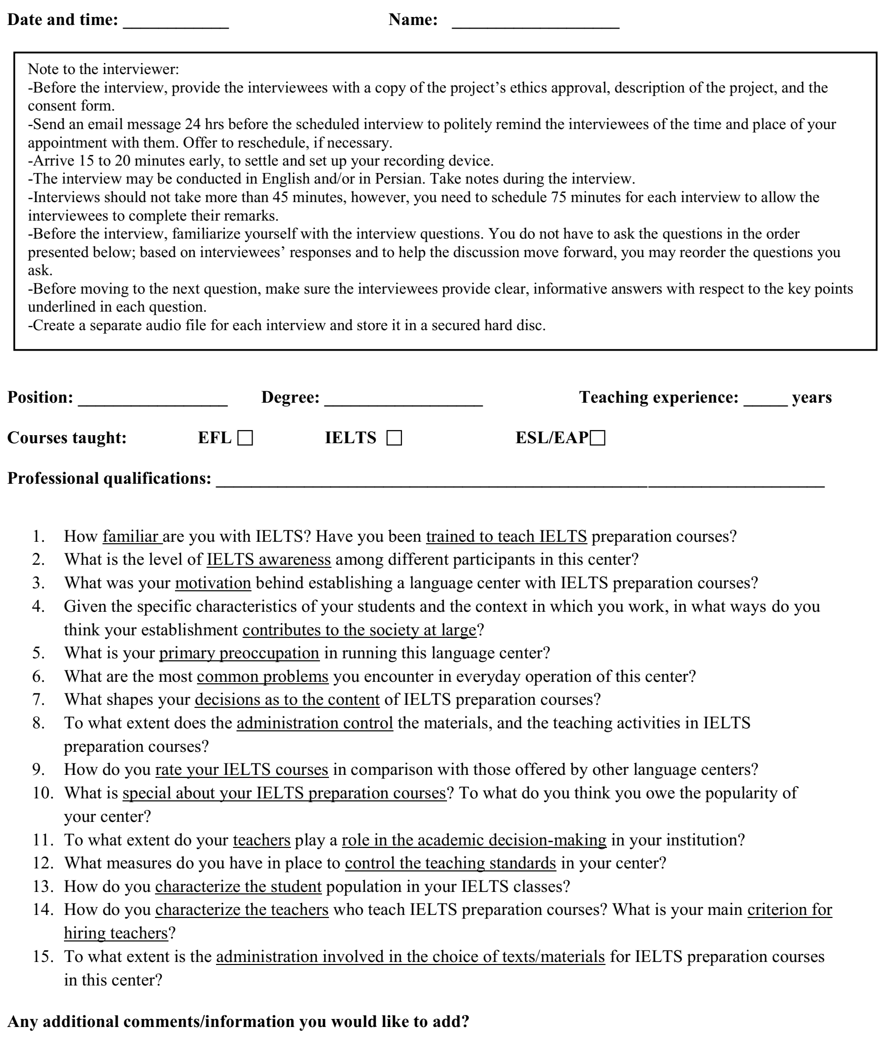
The instructional practices in this context strongly emphasize the improvement of the constructs measured by the IELTS, a focus also shared by TP courses offered in ESL contexts (Hayes and Read 2004; Mickan and Motteram 2008). Additionally, the contextual factors (teachers' experience, belief in English language learning, stakeholders' awareness of learners' needs and objectives) play a considerable role in shaping the orientation of TP in this context. As a result, learners receive rigorous training geared to not only high-stakes test-taking techniques, but more importantly, the development of language abilities measured by the test. These findings are somewhat different from the TP experience in ESL contexts, which focuses on a narrow range of constructs measured by the test—and not target language use (Smith 1991). Hawkey, in his 2006 impact study, concludes that one of the features of TP courses are "learners who are motivated, but sometimes to the extent of wanting, even demanding, a narrower IELTS focus than their teacher would otherwise tend to offer" (p. 112). The current study has directly examined this problem and, therefore, has important implications for other EFL contexts where learners' perceived needs, or as Hawkey (2006) puts it, *wants* or *demands* are shown to adversely affect TP effectiveness.

In addition, teachers and test-oriented materials are found to be fundamental to TP in this context where teachers' classroom approach is impersonal. Course materials, class activities, evaluations, and out-of-class practices are strictly modeled after the IELTS content and format. Teachers adopt a variety of methods, ranging from deductive explanation of forms, to a communicative approach, for teaching the point at hand. Moreover, unlike certain other foreign language contexts (Badger and Yan 2012), teachers in this context use L2 for instruction without making references to the local culture. Conversely, references to L2 culture, that could potentially help students with their performance on test tasks, are common in this specific context.

Acknowledgements This work was supported by the Social Sciences and Humanities Research Council, Canada, under the Partnership Grant number 890-2010-0030. Heartfelt thanks go to Dibagaran Language Institute's dedicated management, instructors, and students whose continued collaboration made this research possible.

Appendix 1

Administrator Interview



Date and time: ____________ **Name:** ___________________

Note to the interviewer:
-Before the interview, provide the interviewees with a copy of the project's ethics approval, description of the project, and the consent form.
-Send an email message 24 hrs before the scheduled interview to politely remind the interviewees of the time and place of your appointment with them. Offer to reschedule, if necessary.
-Arrive 15 to 20 minutes early, to settle and set up your recording device.
-The interview may be conducted in English and/or in Persian. Take notes during the interview.
-Interviews should not take more than 45 minutes, however, you need to schedule 75 minutes for each interview to allow the interviewees to complete their remarks.
-Before the interview, familiarize yourself with the interview questions. You do not have to ask the questions in the order presented below; based on interviewees' responses and to help the discussion move forward, you may reorder the questions you ask.
-Before moving to the next question, make sure the interviewees provide clear, informative answers with respect to the key points underlined in each question.
-Create a separate audio file for each interview and store it in a secured hard disc.

Position: _________________ **Degree:** __________________ **Teaching experience:** _____ **years**

Courses taught: **EFL** □ **IELTS** □ **ESL/EAP**□

Professional qualifications: ____

1. How familiar are you with IELTS? Have you been trained to teach IELTS preparation courses?
2. What is the level of IELTS awareness among different participants in this center?
3. What was your motivation behind establishing a language center with IELTS preparation courses?
4. Given the specific characteristics of your students and the context in which you work, in what ways do you think your establishment contributes to the society at large?
5. What is your primary preoccupation in running this language center?
6. What are the most common problems you encounter in everyday operation of this center?
7. What shapes your decisions as to the content of IELTS preparation courses?
8. To what extent does the administration control the materials, and the teaching activities in IELTS preparation courses?
9. How do you rate your IELTS courses in comparison with those offered by other language centers?
10. What is special about your IELTS preparation courses? To what do you think you owe the popularity of your center?
11. To what extent do your teachers play a role in the academic decision-making in your institution?
12. What measures do you have in place to control the teaching standards in your center?
13. How do you characterize the student population in your IELTS classes?
14. How do you characterize the teachers who teach IELTS preparation courses? What is your main criterion for hiring teachers?
15. To what extent is the administration involved in the choice of texts/materials for IELTS preparation courses in this center?

Any additional comments/information you would like to add?

Appendix 2

Teacher Interview

Date and time: ____________ **Name:** ___________________

Note to the interviewer:
-Before the interview, provide the interviewees with a copy of the project's ethics approval, description of the project, and the consent form.
-Send an email message 24 hrs before the scheduled interview to politely remind the interviewees of the time and place of your appointment with them. Offer to reschedule, if necessary.
-Arrive 15 to 20 minutes early, to settle and set up your recording device.
-The interview may be conducted in English and/or in Persian. Take notes during the interview.
-Interview time should be between 60-75 minutes, however, you need to schedule 90 minutes for each interview to allow the interviewees to complete their remarks.
-Before the interview, familiarize yourself with the interview questions. You do not have to ask the questions in the order presented below; based on interviewees' responses and to help the discussion move forward, you may reorder the questions you ask.
-Before moving to the next question, make sure the interviewees provide clear, informative answers with respect to the key points underlined in each question.
-Create a separate audio file for each interview and store it in a secured hard disc.

Degree: __________________ **Teaching experience:** _____ **years**

Courses taught: EFL □ ___ **years** **IELTS** □ ___ **years** **ESL**□____ **years**

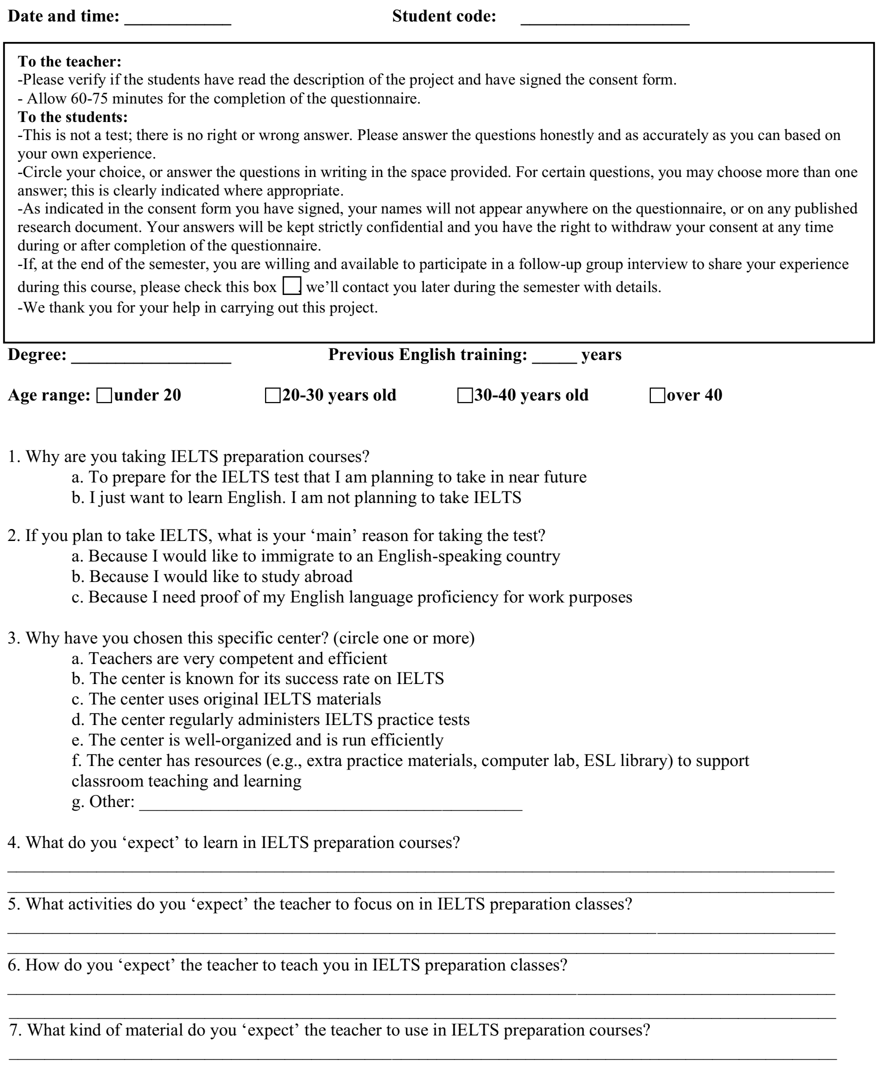
Professional qualifications: ____

1. How familiar are you with the IELTS? Have you been trained to teach IELTS preparation courses?
2. What is your personal opinion about IELTS? Do you consider it a useful test for admission purposes?
3. Given the specific characteristics of your students and the context in which you work, what is your number one IELTS preparation advice for your students?
4. What do you think is/are the most important skill(s) the students should prepare for in this context?
5. How do you prepare your students for IELTS? What guides your teaching methodology in IELTS preparation classes?
6. Do you consider your methodology in IELTS courses similar to what you do in your non-IELTS/GE classes? Why/why not?
7. Do you believe the students who take GE courses could still pass IELTS without taking preparation courses?
8. In your opinion, do IELTS preparation courses improve students' proficiency in English?
9. Are the materials you use in your IELTS preparation classes pre-determined (by the center, for example)?
   -If not, what would you choose as course materials?
   -If yes, do you cover them thoroughly? If you do not cover the material thoroughly, on what sections do you put more emphasis in class? What is your rationale for choosing these sections?
10. How do you choose your class activities? How well do you think they help prepare the students for the test?
11. What do you assign as out-of-class assignments/homework to students in your IELTS preparation courses?
12. In your IELTS courses, do you give your students tips on how to study and what to focus on? Examples?
13. What study tips do you give to your students in your non-IELTS GE courses? Examples?
14. Describe the language center you are working in, its organization, decision-making process, the dynamics/relationships (between the students, the teachers and the students, the center and the students).
15. If you had a choice, what would you prefer to teach; IELTS preparation courses or GE courses? Why?
16. How do you characterize the students in IELTS preparation courses? Are they any different from your students in non-IELTS/GE courses? How?

Any additional comments/information you would like to add?

Appendix 3

Student Questionnaire (English Version)



Date and time: ____________ **Student code: ___________________**

To the teacher:
-Please verify if the students have read the description of the project and have signed the consent form.
- Allow 60-75 minutes for the completion of the questionnaire.
To the students:
-This is not a test; there is no right or wrong answer. Please answer the questions honestly and as accurately as you can based on your own experience.
-Circle your choice, or answer the questions in writing in the space provided. For certain questions, you may choose more than one answer; this is clearly indicated where appropriate.
-As indicated in the consent form you have signed, your names will not appear anywhere on the questionnaire, or on any published research document. Your answers will be kept strictly confidential and you have the right to withdraw your consent at any time during or after completion of the questionnaire.
-If, at the end of the semester, you are willing and available to participate in a follow-up group interview to share your experience during this course, please check this box □ we'll contact you later during the semester with details.
-We thank you for your help in carrying out this project.

Degree: __________________ **Previous English training: _____ years**

Age range: □under 20 □20-30 years old □30-40 years old □over 40

1. Why are you taking IELTS preparation courses?
   a. To prepare for the IELTS test that I am planning to take in near future
   b. I just want to learn English. I am not planning to take IELTS

2. If you plan to take IELTS, what is your 'main' reason for taking the test?
   a. Because I would like to immigrate to an English-speaking country
   b. Because I would like to study abroad
   c. Because I need proof of my English language proficiency for work purposes

3. Why have you chosen this specific center? (circle one or more)
   a. Teachers are very competent and efficient
   b. The center is known for its success rate on IELTS
   c. The center uses original IELTS materials
   d. The center regularly administers IELTS practice tests
   e. The center is well-organized and is run efficiently
   f. The center has resources (e.g., extra practice materials, computer lab, ESL library) to support classroom teaching and learning
   g. Other: ____

4. What do you 'expect' to learn in IELTS preparation courses?
____
____

5. What activities do you 'expect' the teacher to focus on in IELTS preparation classes?
____
____

6. How do you 'expect' the teacher to teach you in IELTS preparation classes?
____
____

7. What kind of material do you 'expect' the teacher to use in IELTS preparation courses?
____

8. Are you taking English courses or involved in language learning activities other than IELTS preparation courses?
a. YES b. NO

8.1 If you answered 'YES' to question 8, circle one or more of the following options?
a. I use internet a lot b. I watch English TV channels
c. I read in English a lot d. I take a General English course
e. I take an English for Academic Purposes course.
f. Other: ____

9. What kind of materials do you use outside of IELTS preparation courses? (circle one or more)
a. IELTS-related materials/textbooks
b. Non-IELTS materials
c. Other: ____

10. What language do you use in IELTS preparation classes for communicating with your teacher/peers?
a. Mostly English b. Mostly Farsi
Why? ____

11. What language do you 'expect' the teacher to use in IELTS preparation courses?
a. English b. Farsi c. Both
Why? ____

12. Name one class activity you consider as the 'most important' activity for success on IELTS?
____

13. So far, in your opinion, has the preparation course you are taking been useful in improving your chances of success at IETLS? a. YES b. NO

13.1 If yes, what has/have been the most useful aspect(s) of your IELTS preparation course?
____
____

13.4 If not, what would you change about the present IELTS preparation course you are taking?
____
____

14. Do you think your language learning practices in preparation for IELTS are different from those in General English or other non-IELTS courses you have taken previously? a. YES b. NO

15. Do you 'expect' your general English proficiency to improve as a result of preparation for the IELTS?
a. YES b. NO

16. If your sole purpose were to improve your English language proficiency (and not passing IELTS), what kind of course would you take?
a. A course in English for Academic Purposes
b. A course in General English
c. An IELTS preparation course
d. Other: ________________________

Please feel free to share with us any additional comments or suggestions you have about IELTS preparation or learning English in general in this specific context.

Appendix 4

OBSERVATION CODING SHEET
(Saif et al., 2019. Reprinted by permission from the publisher, Taylor & Francis Ltd.)

Centre: **Class/Week:**
Observation Date: **Coding Date:**
Teacher: **Total Class Time:**

Textbook== >
Time:

Time	Classroom Activity(ies)		Participants' roles			Course Content										Course Materials				Reference(s) to IELTS content, parts, topics, length, item format, test-taking strategies	Notes for each activity
	Activity #	Test directed? (YES/NO)	Teacher-led	Students' group work	Students work individually	Language abilities practiced								teaching / learning activity (ies)	In-class assessment? (mock exams; timed-tests?)	IELTS		Non-IELTS			
						R	W	S	L	gr	voc	pro	Strategy use			In-class	assigned as homework	In-class	assigned as homework		

Appendix 5

Student Group Interview

Date and time: ____________ **Group number: _____**

Note to the interviewer:
-Before the interview, provide the interviewees with a copy of the project's ethics approval, description of the project, and the consent form.
-Send an email message 24 hrs before the scheduled interview to politely remind the group of the time and place of your appointment with them.
-Arrive 15 to 20 minutes early, to settle and set up your recording device.
-Conduct the interview in the form of a roundtable, in Farsi. The students may answer in English and/or in Farsi. Take notes during the interview.
-Interview time should be between 75-90 minutes, however, you need to schedule 120 minutes to allow for everyone in the group to contribute to the discussion and express their opinion and/or react to others' responses.
-Before the group interview, familiarize yourself with the interview questions. You do not have to ask the questions in the order presented below; based on interviewees' responses and to help the discussion move forward, you may reorder the questions you ask.
-Create a separate audio file for each group interview and store it in a secured hard disc.

1. What do you think about the IELTS, its format?
2. How important is it to you to succeed on the IELTS? How consequential do you think a preparation course is to your success on IELTS?
3. Thinking about the IELTS course you just completed:
   a. what do you like the most about it?
   b. what class activities you participated in are the most crucial to your success on the IELTS?
   c. how do you characterize the relationship between the teaching activities and the IELTS content?
   d. how do you characterize the materials?
   e. what component of the IELTS, if any, do you think your teacher should have prioritized in class?

4. Do you think taking a General English course could help with your performance on the IELTS? In what ways?
5. Do you believe if you take General English courses long enough, you could still succeed on the IELTS without taking any preparation course? Elaborate.
6. Do you prepare for IELTS the same way you prepare for the tests in non-IELTS classes? Why/why not?
7. What guides your learning behavior in the IELTS preparation course? How do you choose your out-of-class learning activities to further practice what you learn in class?
8. In your opinion, do IELTS preparation courses improve your proficiency in English as well? Elaborate.
9. Did you get tips from your teacher as to how to study and what to focus on in your IELTS courses/practices? Do you think these tips help you perform better on the test? Give examples.
10. Describe the language center you are studying in, the dynamics, its organization, the relationships (between the students, the teachers and the students, the center and the students). What do you like/dislike about it?
11. Presently, how do you rate your English language proficiency? If you do not obtain the desired score on the IELTS, do you think it is because of your inadequate English language proficiency level? Explain.
12. How different do you think IELTS preparation courses are from other non-IELTS courses you have so far taken?

Any additional comment or suggestion about IELTS preparation or learning English in general in this specific context?

Appendix 6

TEACHER INTERVIEW RESULTS (cross-checked against parallel data)

Teachers’ perceptions of	Source		
	Teacher interviews (#of respondents)	Class observations	Students’ questionnaire and focus-group interviews
Qualifications to teach TP	• COMMAND of English (6); ability to transfer knowledge (4); rapport with students (2), classroom management (2); teaching experience (5); familiarity with test (6); yield results (3)		Q: “qualified teachers” (51%) main reason for choosing this center FGI: ‘competent teachers’
The test (IELTS)	• Appropriate for admission (3): measures GE and four abilities • Inappropriate for admission (3): doesn’t measure language, subjective scoring of oral and written abilities, listening section lacks context	• Clear, direct, frequent references to test content/tasks while teaching; some deductive grammar (teachers digress to teach common grammatical problems) • Exact IELTS tasks/questions as class activity or homework	FGI: IELTS extremely important to future plans/immigration; stressful; at least a 6.5 score needed; not necessarily measuring language abilities
What the students should focus on	• Improving GE (5); strategy use and practice (2); defining their language learning goals (2); use multiple learning resources (1)	• Brought test instructions to students’ attention; form-focused approach for grammar; correct pronunciation; accuracy feedback after practicing each task	Q: ranked “four skills” (39%) followed by “speaking” (34%) as skills the teachers should focus on
Important abilities to promote	• Familiarity with test method/content (1); four abilities (3); reading (2)	• Four skills practiced and promoted in class; grammar and vocabulary mostly taught as part of reading and listening comprehension activities	Q: ranked “test-taking techniques” (45%) followed by “four abilities” (27%) for TP class; “speaking” most important for *passing the test* (29%) FGI: proficiency alone not enough for success, test-taking techniques and familiarity with test format crucial

(continued)

(continued)

Teachers’ perceptions of	Source		
	Teacher interviews (#of respondents)	Class observations	Students’ questionnaire and focus-group interviews
Own teaching method	• Integration of four skills (2), communication strategies (1); needs-based (1); GE (5), test-based teaching/test-taking strategies (3); individualized training (1)	• Communicative method; occasional deductive teaching of grammar; L2 at all times; students’ mistakes explicitly explained	Q: teachers’ current approach is helpful; current teaching methods used
Teaching TP vs. GE courses	• Language abilities fundamental to both (1); different approach for TP because of life-changing consequences (2); strategy training for TP vs. cramming of rules/words in GE (2); test-taking techniques (1); TP is test-based (1); frequent practice tests for TP (1)	• TP activities mostly test-oriented; practice four language skills • No observation data from GE courses	Q: different courses (91%); TP courses also improve GE proficiency (86%) vs. GE courses helps pass IELTS (32%) FGI: very different in nature; cramming doesn’t work for IELTS; TP process and test-taking skill matters
TP and language proficiency	• TP courses also improve proficiency (6); teacher’s methodology, teaching activities and materials play a role too (3)	• Classroom activities represent four skills; simulate test format; • Recommended out-of-class activities mostly help with fluency and communicative ability	FGI: TP helps proficiency; GE courses don’t help IELTS success
Teaching materials	• Pre-determined materials (6); teachers choose course activities based on students’ level/needs (4); supplementary materials based on students’ needs (2)	• IELTS official materials; students’ writing used as practice materials (common problems discussed; students justify use of forms/vocabulary/ideas)	Q: 55% use IELTS-related supplementary materials FGI: materials for “speaking” useful but not sufficient; more audio/video materials needed

(continued)

(continued)

Teachers' perceptions of	Source		
	Teacher interviews (#of respondents)	Class observations	Students' questionnaire and focus-group interviews
Teaching activities	• Test-oriented activities (5); timed practice-tests (1); GE practice (2); give results (3)	• Class activities reflect test tasks and real-life; writing activities reflect test format; in-depth analysis of students' performance (what to do/not to do in real test)	Q: ranked "test-oriented activities (12.5%)"; followed by "keep doing what they are doing" (11%) FGI: teaching activities directly related to test
Recommended out-of-class activities	• Practice tests (2); practice four skills (2); reading all the time (1); formulaic phrase bank (1); use internet resources (2); recommendations tailored to individual needs (1)	• Homework assigned based on IELTS workbook; promote reading "as much as possible" in English	Q: involved in out-of-class activities (35%) ranked "internet" (52%), "TV" (45%), "reading" (25%) as top choices FGI: online IELTS practice material; music and movies; reading all sorts of materials online; group conversations
Strategies to promote	• Reading strategies (5); test-taking tips (2); communication strategies similar to those in GE (2)	• Both communication and test-taking strategies; skill related strategies frequently taught	FGI: teachers' tips (formulaic patterns, listening/reading strategies, recording ourselves) help perform better
TP students	• Motivated to pass the test (4); just pass the test not learn English (1); not as fun as GE students (1); hard-working, serious, rarely give up or drop out (2)	• Participate in class activities; rarely missing the course; follow instructions; volunteer their written work/oral responses for analysis in class; not everyone gets a chance to speak because of time limit; take notes consistently	FGI: our learning guided by need to succeed, personal and collective motivation; teachers' energy, center's culture and force

(continued)

(continued)

Teachers’ perceptions of	Source		
	Teacher interviews (#of respondents)	Class observations	Students’ questionnaire and focus-group interviews
Test center	• Quality-oriented (1); high standards/training for teachers/common goal (2); dynamic and professional environment (3); everything under control (1); pre-selected materials (1); students’ feedback/needs valued (1); friendly relationships (2)		Q: ranked center strengths as “qualified teachers” (51%); “original IELTS materials” (34%); “practice IELTS tests” (29%); “well-organized” (20%); “success rate” (16%) FGI: positive experience: relationship with teachers/staff; experienced teachers; organized, but short class time
Comments/observations	• Would not stick to specific materials/activities/methods and would choose what necessary to help students (1); students’ needs should come first (1); materials should match those used in other countries (1); have taken the test myself (1); teachers should be very familiar with the test (2)		• Four skills should be focused on/practiced in depth; no exposure to English in Iran so oral abilities especially important; separate classes for different skills; should focus on students’ weaknesses/problems and needs

Appendix 7

STUDENT QUESTIONNAIRE AND FOCUS-GROUP RESULTS (cross-checked against parallel data)

Students' perceptions	Source		
	Student questionnaire	Student focus-group Interviews	Observations
Why take TP courses/IELTS	• Education abroad (70%); work abroad (41%); immigration (21%); learn English (37%)	• Education abroad; immigration • Improving proficiency (TP helps GE development better than other courses) • IELTS extremely important to future plans; stressful; immigration depends on it; need a minimum 6.5 score; test doesn't necessarily measure language abilities so training is necessary	
Why this center	• Qualified teachers (51%); original IELTS materials (34%); practice IELTS tests (29%); well-organized (20%); success rate (16%)	• Positive experience: relationship with teachers/staff; experienced teachers; organized, but short class time • Referral from others	

(continued)

(continued)

Students' perceptions	Source		
	Student questionnaire	Student focus-group Interviews	Observations
Own motives	• Education abroad (70%); work abroad (41%); immigration (21%); learn English (37%)	• Students' learning behavior guided by need to succeed, personal and collective motivation; teachers' energy, center's culture and force	• Students participate in all class activities; rarely miss the course; follow instructions; volunteer their written work/oral responses for class practice but not everyone gets a chance to speak because of time limit; take notes consistently
Expect the TP to focus on	• "Test-taking techniques" (45%) followed by "four abilities" (27%); "speaking" (25%); "vocabulary" (16%); "listening" (12.5%); "writing" (12.5%)	• Proficiency alone not enough for success, test-taking techniques and familiarity with test format crucial • Integration of skills in class not productive; students in Iran learn a lot of grammar and vocabulary but cannot speak in English; TP should focus on skills students lack	• Teachers occasionally focus on and explicitly teach grammatical points highlighted in IELTS material • Class instruction covers four skills; none of the classes observed focused on just one skill
Expected teacher activities	• Four skills (39%); speaking (34%); writing (30%); reading (12.5)	• 'Speaking' and 'writing' most important skills to prepare for • Out-of-class oral practice helps but students need to practice speaking with teacher supervision to make sure they use correct socio-cultural references	• Plenty of interactional activities related to four skills but no prolonged focus on any particular skill • Activities mostly test-related

(continued)

(continued)

Students' perceptions	Source		
	Student questionnaire	Student focus-group Interviews	Observations
Choice of out-of-class activities	• 35% involved in out-of-class activities like "internet" (52%), "TV" (45%), and "reading" (25%)	• Online IELTS practice material; music and movies; reading all sorts of materials online; group conversations	• Homework assigned based on IELTS workbook; promote reading "as much as possible" in English • Reading comprehension assignments • Writing and grammar assignments directly related to test
TP activities and the test	• Four kills tested by IELTS should be practiced in class • Course should address students' weaknesses/problems/language needs	• Class activities related to the test • Four skills should be focused on/practiced in depth; • Oral section of the test especially difficult for students; no exposure to English in Iran, separate classes entirely devoted to oral practice and a lot of feedback needed	• Clear, direct, frequent references to parts of the test while teaching; some grammar discussions not test-directed (teachers digress when they notice a grammatical point is a common problem for students); teachers use exact IELTS tasks/questions as class activity or homework; constantly provide test-taking tips
Preferred teacher method	• Teachers' current approach is helpful; focus on most recent teaching methods; use of technology	• Teachers' methodology helps with test preparation; but class periods are too short to fully practice skills	• Mostly communicative method; deductive teaching of grammar at times; always English in class • Real-time feedback on students written/oral production in class; students' mistakes explicitly explained

(continued)

(continued)

Students' perceptions	Source		
	Student questionnaire	Student focus-group Interviews	Observations
Preferred TP materials	• IELTS original materials for all skills/sub-skills • IELTS-related supplementary materials (55%)	• "Speaking" material useful but not sufficient; more authentic materials audio/video/movies needed	• IELTS official materials used; students' writing used for class practice (common problems discussed; students justify their use of forms/vocabulary/ideas)
Other ways of preparation for IELTS	• Among those involved in out-of-class activities (35%): "internet" (52%), "TV" (45%), "free reading" (25%)	• Online IELTS practice material; music and movies; reading various texts online; group conversations	• Homework assigned based on IELTS workbook; promote reading 'as much as possible' in English
Use of L2 in/out-of-class	• Out-of-class L2 use (86%) boosts proficiency; L2 by teacher (64%) helps mastery of English; both L1 and L2 in class (36%) ensures comprehension	• Teachers encourage students to think in English; students need to know more about L2 culture • Need teachers' feedback on language use	• Students and teachers use L2 in class • Some sporadic use of L1 by students; teacher reacts in L2
Most important for test success	• Speaking (29%) important for success but writing the most useful aspect of TP courses (34%) • 25% would improve the "speaking" content of the course	• Students need guidance and practice for speaking and writing abilities • Students need more real-life material than IELTS texts provide (need to understand L2 culture), videos and movies very helpful but more materials needed	• Most homework based on IELTS materials and texts

(continued)

(continued)

Students' perceptions	Source		
	Student questionnaire	Student focus-group Interviews	Observations
IELTS vs. GE courses	• Learning practices are different (91%); 63% previously taken GE courses • TP courses also improve GE (86%) vs. 32% who think GE courses help pass IELTS	• Very different in nature; cramming doesn't work for IELTS; TP process and test-taking skills matter • TP helps proficiency but GE courses don't help pass IELTS	• Most TP activities test-oriented and practice four skills • No observation data from GE courses
Strategy training	• Test-taking techniques a priority (45%)	• Teachers' tips (formulaic patterns, listening/reading strategies, recording ourselves, etc.) help perform better	• Both communication and test-taking strategies; skill-related strategies frequently highlighted while teaching

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280–297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14,* 115–129.

Badger, R., & Yan, X. (2012). To what extent is communicative language teaching a feature of IELTS classes in China? *IELTS Research Reports, 3,* 169–212.

Bailey, K. M. (1996). Working for washback: A review of washback concepts in language testing. *Language Testing, 13*(3), 256–279.

British Council. (2003). *IELTS Annual Review*. http://www.nlc.cl/ielts/IELTSAnnualReview2003_v1.pdf. Accessed 15 Aug 2019.

Chappell, P., Bodis, A., & Jackson, H. (2015). The impact of teacher cognition and classroom practices on IELTS test preparation courses in the Australian ELICOS sector. *IELTS Research Reports, 6,* 1–61.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study.* Cambridge: Cambridge University Press.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods.* Mahwah, NJ: Lawrence Erlbaum.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education (Studies in Language Testing 25).* Cambridge: Cambridge University Press.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000.* Cambridge: UCLES/Cambridge University Press.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 97–112). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Hughes, A. (1993). *Backwash and TOEFL 2000*. Unpublished manuscript, University of Reading.

Irving, A., & Mullock, B. (2006). Learning to teach the Cambridge CAE: A case study. *Prospect, 21*(2), 82–116.

Lewis, J., & Ritchie, J. (2003). Generalizing from qualitative research. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice: A guide for social science students and researchers* (pp. 263–286). London: Sage.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29–46.

Matoush, M. M., & Fu, D. (2012). Tests of English language as significant thresholds for college-bound Chinese and the washback of test-preparation. *Changing English, 19*(1), 111–121.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*(1), 14–22.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17,* 67–91.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256.

Mickan, P., & Motteram, J. (2008). An ethnographic study of classroom instruction in an IELTS preparation program. *IELTS Research Reports, 8,* 1–26.

Miyasaka, J. R. (2000). *A framework for evaluating the validity of test preparation practices.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Muñoz, A. P., & Álvarez, M. E. (2010). Washback of an oral assessment in the EFL classroom. *Language Testing, 27*(1), 33–49.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Journal of Language Testing, 23*(1), 1–34.

Saif, S. (2012). The washback of a task-based test of spoken language on the development of international teaching assistants' strategic competence. In G. Gorsuch (Ed.), *Working theories for teaching assistant and international teaching assistant development* (pp. 575–608). Stillwater, OK: New Forums Press.

Saif, S., Ma, J., May, L., & Cheng, L. (2019). Complexity of test preparation across three contexts: Case studies from Australia, Iran, and China. *Assessment in Education: Principles, Policy & Practice*. https://doi.org/10.1080/0969594x.2019.1700211.

Shih, C. M. (2009). How tests change teaching: A model for reference. *English Teaching: Practice and Critique, 8*(2), 188–206.

Smith, M. L. (1991). Meanings of test preparation. *American Educational Research Journal, 28,* 521–542.

Spada, N., & Fromlich, M. (1995). *COLT—Communicative orientation of language teaching observation scheme: Coding conventions and applications*. Sydney: NCELTR.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41–69.

Wall, D., & Horák, T. (2011). *The impact of changes in the TOEFL exam on teaching in a sample of countries in Europe: Phase 3, the role of the coursebook, phase 4, describing change* (TOEFL iBT Research Report No. TOEFLiBT-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02277.x. Accessed 15 Aug 2019.

Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: Sage.

Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., et al. (2017). Preparing for the speaking tasks of the TOEFL iBT test: An investigation of the journeys of Chinese test takers. *TOEFL iBT Research Report, 28,* 1–59.

Chapter 36
Development of a Profile-Based Writing Scale: How Collaboration with Teachers Enhanced Assessment Practice in a Post-Admission ESL Writing Program at a USA University

Xun Yan, Ha Ram Kim, and John Kotnarowski

Abstract This study reports on the revision of a rating scale for an ESL writing placement test and demonstrates how collaboration with teachers can both enhance the assessment practice within an ESL writing program and raise teachers' awareness about assessment literacy. Following a data-driven approach to scale development, teachers participated in a three-stage revision process, where they were asked to (1) reflect on the range of writing performances in ESL courses, (2) evaluate sample essays from the test and revise descriptors for the new scale, and (3) pilot-rate new essays using the new scale. During the first stage, both teachers and testers recognized that test takers display different strengths and weaknesses in argument development and lexico-grammar. However, when evaluating sample essays, teachers seemed to weigh argument development more heavily, whereas testers placed a higher value on lexico-grammatical accuracy. Additionally, when rating argument development, some teachers relied heavily on surface/structural rhetorical features rather than essay content. These contrasts resulted in conflicting ratings on certain essay profiles. Through several rounds of discussion, these differences were eventually mitigated by creating separate criteria for argument development and lexico-grammar. The revised scale strikes a better balance between argument development and lexico-grammar, more accurately covering the range of writing performances among test takers. The revision process standardized the conceptualization and operationalization of writing quality, shifting teachers' focus from surface rhetorical features to essay content. In return, collaboration with teachers enhanced testers' understanding of the local instructional contexts. Teachers' involvement promoted collaborative assessment-related dialogues and practices within the ESL program, strengthening the alignment across curriculum, instruction, and assessment.

X. Yan (✉) · J. Kotnarowski
University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: xyanacademic@gmail.com

H. R. Kim
University of California, Irvine, USA

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_36

36.1 Introduction: Purpose and Testing Context

In the last decade, many university campuses across North America have witnessed a significant increase in the number of incoming international students (Institute of International Education 2017). With this “surge” in the international student population, the need/demand for English for Academic Purposes (EAP) courses has increased, and concerns have been raised as to the amount of support that should be given to these students, particularly in relation to English language support in Writing, Pronunciation, and International Teaching Assistant Training (Chung 2014). Consequently, it has become an important task for language programs to find placement measures that accurately assess students’ English proficiency. This study reports on the revision of a rating scale for an ESL writing placement test at the University of Illinois at Urbana-Champaign (UIUC) to provide placement and diagnostic information about students’ writing ability. By involving ESL writing teachers in the re-scaling process, this study demonstrates how collaboration with teachers can both enhance the assessment practice within an ESL writing program and raise teachers’ awareness about assessment literacy.

The ESL Service Courses at UIUC provide language support for over 10,500 international students across campus in order to facilitate student success in an English academic environment and communicate in a way that conforms to the conventions of academic discourse. UIUC does not have pre-matriculation or bridge courses, and the largest unit in ESL at UIUC is the Academic Writing Program, which offers courses at both undergraduate and graduate levels. Many of the ESL courses are taught by both full-time lecturers and teacher candidates in master’s or doctoral programs at UIUC with backgrounds in Linguistics, Teaching English as a Second Language (TESL), or Education.

The English Placement Test (EPT) is administered to newly admitted international students who do not meet the English requirement as mandated by the university. The EPT is offered throughout the year, but the majority of the test administrations are offered during summer and the week before the start of Fall and Spring semesters. The EPT consists of two parts: a written and an oral exam. In the written part of the EPT (which is the focus of this study), students write an argumentative essay following academic writing conventions on a given topic based on a reading passage and a short lecture. Based on the results of the EPT, students are placed into the appropriate ESL course(s). Undergraduate students are placed into either a two-semester course sequence (ESL 111, 112) or a single-semester course (ESL 115) to fulfill their language requirement. Similarly, graduate students can be placed into either a two-semester course sequence (ESL 511, 512) or a single-semester course (ESL 515). Graduate students in certain fields may also choose to complete equivalent coursework from the “English for Specific Courses” business tracks (ESL 521, 522). ESL courses at UIUC tend to focus on rhetorical features, with the lower level (ESL 111/ESL 511) focusing on general structure and conventions of academic writing and the higher level (ESL 115/ESL 511) focusing on more fine-grained rhetorical features (see Table 36.1).

Table 36.1 ESL writing courses offered at the University of Illinois at Urbana-Champaign

	Two-semester	One-semester	ESL exemption
UG	ESL 111, 112	ESL 115	RHET 105: Mainstream first-year composition courses
G-General	ESL 511, 512	ESL 515	No ESL courses required
G-Business	ESL 521, 522	ESL 522	No ESL courses required

36.2 Testing Problems Encountered

UIUC is not alone in terms of the assessment challenges faced by the local ESL program. These challenges arise as a result of the tension between a restricted range of English proficiency levels among the admitted ESL students and the need to make more fine-grained distinction and diagnosis of the students' writing skills. We explicate the assessment challenges for language assessments embedded in local ESL programs below.

36.2.1 Restricted Range of Proficiency

Many large universities in the US have set minimum cut-off test scores to screen the applicants by their English proficiency (e.g., TOEFL or IELTS total scores). At the same time, schools also often require students who scored lower than a (required) certain test score on the school's individual placement test to receive additional language support by taking courses in their ESL program. For instance, UIUC requires international undergraduate students who score below 103 in total, and below 25 on either speaking or writing sections of the TOEFL iBT (below 7.5 in total and below 7 on either speaking or writing sections of the IELTS) to take the EPT (https://linguistics.illinois.edu/languages/english-placement-test); graduate students whose total TOEFL score is below 103 (or total IELTS score below 7.5) are also required to take the EPT. Purdue University, another Midwestern university with a large international population, also requires undergraduate students who scored 100 or less on TOEFL iBT, or less than 7.5 on the IELTS to take the institutional English for Academic Purposes placement test. Therefore, the students who end up taking the ESL courses are within a restricted range of proficiency with little variance in their performance (roughly with an overall score of 80–100 on TOEFL), and this provides unique challenges for local ESL programs.

Many local programs rely on or adopt existing rating scales from general proficiency tests like TOEFL, which target a fuller range of proficiency. This may be largely due to budgetary reasons, lack of expertise, and/or the amount of effort and time involved in creating new rating scales. Even though the use of existing rating scales has many advantages and has been widely researched and well validated (e.g., Chapelle et al. 2010), these scales are not necessarily sensitive enough to capture the nuance of writing performances among the students within the restricted range (Cho and Bridgeman 2012; Bridgeman et al. 2015; Ginther and Yan 2018). As most university-level ESL students fall within the high intermediate to low-advanced proficiency group, such scales tend to have limited use due to this restricted range of proficiency.

36.2.2 Lack of Diagnostic Functions of Writing Placement Tests

Another challenge for post-admission language tests is the need to provide not only placement decisions but also diagnostic information for language instructors about students' actual language/writing performance and skills. Oftentimes, post-admission language tests focus on a holistic score to provide placement decisions, but they are not specifically designed to provide specific diagnoses of students' writer profiles. Here, we define profile as a description of the strengths and weaknesses in students' writing performance. As a result, fine-grained diagnosis of students' language ability is forced to take place in classrooms, and teachers need to collect students' writing samples for diagnostic purposes. However, there has been a lack of attention or interest in diagnostic tests, especially with the dominance of high-stakes testing (e.g., Alderson 2005); local tests embedded in language programs are learning-oriented and would benefit much from combining placement and diagnostic purposes (Purpura 2004). These tests not only provide placement decisions but also demonstrate students' strengths and weaknesses, especially with the increase of incoming students with proficiency levels that fall within the aforementioned restricted range of proficiency. In addition, the nature of language proficiency changes across speakers of different proficiency levels (e.g., Alderson 1991; Kunnan 1992; Oltman et al. 1988), and different "profiles" of writers exist even within the same proficiency group with varying strength and weakness in relevant sub-skills. For instance, within the intermediate-advanced proficiency range, some students have strong receptive skills (reading and listening), while showing contrasting weakness in productive skills (speaking and listening) (Bridgeman et al. 2015; Ginther and Yan 2018). Jarvis et al. (2003) have also shown that multiple profiles exist even among highly rated compositions, despite some identifiable common traits in their writing quality. These profiles can have different implications for instruction. Therefore, in order to fully capture the subtle profile differences in students' performances, rating scales for post-admission language tests need to explore the possibility of identifying learner profiles, which can offer both placement and diagnosis of ESL students. This paper reports on the revision process for a rating scale for an ESL writing placement test, as an effort to rate both writing proficiency and articulate profile differences among ESL students. By involving both teachers and testers, we explore how collaboration between teachers and testers influences their evaluation of writing performances of ESL students at a large US university. Specifically, we address the following three research questions.

> Research question 1: What kind of writing profiles are represented in the ESL writing courses and test?
> Research question 2: How do teachers and testers evaluate essays across proficiency levels and profiles? Are there any differences in how they rate essays?
> Research question 3: Can the differences in rating between teachers and testers, if any, be mitigated during a scale development process?

36.3 Review of the Literature

36.3.1 Different Approaches to Scale Development

In order to develop a rating scale that can capture profile differences in students' performances, it is necessary to discuss how rating scales are developed. Several approaches to the construction of rating scales have been proposed in the literature. Two major approaches to scale development, as classified by Fulcher et al. (2011), are the measurement-based and performance-based data-driven approaches. The measurement-based approach is probably the oldest and most commonly used method of constructing a scale, and it starts by identifying the common features to be evaluated and/or descriptors at varying levels of proficiency. This approach relies on existing scales or the intuitions of the people who are perceived to be "experts" in teaching or assessment of the subject. Once identified, level descriptors are placed into a single scale based on the estimates of their difficulty. No real performance analysis is required at this stage, but a post hoc measurement method (e.g., Rasch analysis) is used afterwards to test and ensure the reliability of the descriptors and the validity of score inferences (e.g., Banerjee et al. 2015; Fulcher et al. 2011). Because of the subjective nature of developing rating criteria and descriptors, such a priori developed scales have been criticized for being less specific, imprecise, and thus resulting in inconsistent ratings across raters (Knoch 2009). Many scholars have pointed out that the language used in these scales is often relativistic, abstract, and impressionistic, which allows for subjective interpretations of the features differentiating between bands or proficiency levels (Brindley 1998; Mickan 2003; Upshur and Turner 1995). Due to this weak link between the scale (meaning) and performance (score) of intuitively developed scales (Fulcher et al. 2011; Pollitt and Murray 1996), there have been some concerns that raters might not be able to successfully make fine-grained distinctions of different traits across levels and lose important diagnostic information (Knoch 2009).

Unlike the measurement-based approach that relies on intuitively derived, predetermined features, and post hoc measurements to ensure reliability and/or validity, a performance-based, data-driven approach starts from collecting and analyzing the actual performance samples. The analyses of performance data result in the identification of key features or traits that can distinguish performances between different proficiency levels. The number of levels in a scale is also empirically established using discriminant analysis, and the features identified in an earlier analysis are used to describe each level in the scale (Fulcher 1993, 1996, 2003; Fulcher et al. 2011). Even though this method allows for a close analysis of actual performance samples and strengthens the link between scale and actual performance, it is not without criticisms; researchers have noted that the data-driven approach to scale development can be time-consuming, and it produces analytic descriptors—often linguistic constructs—that human raters might find difficult to use in real-time rating (e.g., Fulcher 2003; Fulcher et al. 2011; Upshur and Turner 1995; Banerjee et al. 2015). Addressing the issues above, an ideal scale development model should take advantage of both approaches by involving both experienced teachers and testers in the scale development process. Teachers and testers can collaborate to (1) collect and analyze actual performance samples, (2) identify traits or features that can help identify different performance profiles, and (3) develop key descriptors that can be easily operationalized by human raters and provide diagnostic information about the examinees. Furthermore, in local contexts where ESL students are within a restricted range of proficiency, a performance data-based scale would be a better approach between the two when developing a profile-based scale. While most measurement-based scales focus on examining the full range of proficiency levels, performance data-based scales start by identifying the key features of each performance level observed in actual writing samples. This process allows testers to capture and identify the subtle differences in a limited proficiency range.

36.3.2 Collaboration with Teachers in Test Development: Challenges and Benefits

In language programs, there are two groups of people who are more likely to be familiar with the test performance data: (1) testers who participate as well as coordinate the rating and (2) instructors who teach the courses and participate in rating as an additional duty. Therefore, in our approach to developing a new profile-based scale, we have involved both groups. Although teachers' assessment literacy has been questioned in the previous literature (Mertler 2009; Popham 2001; Stiggins 1999; White 2009), thereby raising concerns about their qualification for test development, it is common practice to involve teachers in different stages of test development (e.g., Fulcher et al. 2011; Gudrun and Aberg-Bengtsson 2012).

There are both advantages and disadvantages of involving teachers in the test development process. Teachers are strong at language pedagogy and are familiar with both the teaching and assessment contexts in their local institutions; however, they tend to be weak in theoretical and technical knowledge related to assessment. In writing assessment, Crusan et al. (2016) revealed that teachers had mixed feelings about scoring and writing assessment and professed their lack of assessment literacy (p. 50). In particular, about 80% of teachers in the survey responded that they were unsure about how to design scoring rubrics. They do not exactly know how to develop a rating scale even though they, in fact, have been using various types of rating rubrics in their classrooms. So, while they are forced to use rating rubrics in their classrooms as a form of "best practice," they are not confident in their ability to develop rating scales themselves. Despite these challenges, there are also benefits of involving teachers more in the test development process. One of the biggest strengths and contributions that teachers can provide is their knowledge of both their students (language learners) and the local context. As such, they are able to identify the students' profiles (strengths and weakness) and reflect this knowledge in assessment practices (Purpura 2004). The experience of being involved in various stages of test development enables them to better connect teaching and assessment (Yan et al. 2018), and also provides an opportunity to better understand their students within the restricted range. Next, involving teachers in actual assessment practices can help enhance their assessment literacy. The need to involve teachers in assessment practices to increase their assessment literacy has been addressed by many scholars (e.g., Crusan 2010; Crusan et al. 2016; Lee 2010; Weigle 2007; Xu and Brown 2016; Yan et al. 2018). Weigle (2007) emphasized that teachers need to acquire skills including developing, administering, and scoring writing tasks, as well as the ability to recognize components of a good paper or good writing. Yan et al. (2018) also suggested that teachers have and can further develop good intuitions about assessment concepts and principles by participating in various assessment practices, and these iterative, accumulated assessment experiences would enable them to feel more "empowered" as teacher-assessors, an insight that aligns with several other studies (Crusan et al. 2016; Lee 2010; Xu and Brown 2016).

Therefore, despite some concerns about teachers' "lack" of assessment literacy, there is evidence in support of the advantages of adopting a collaborative approach to developing assessment materials, such as a rating scale. This collaboration provides opportunities for teachers to increase their assessment literacy and better understand their students' language profiles. It can also be beneficial to testers, as teachers can contribute knowledge about their students and develop descriptors aligned with the assessment language used in writing classrooms. The collaboration in assessment practices thus helps to bridge the gap between the two groups and increases the usability and interpretability of the rating scale. As such, the communication between teachers and testers should be promoted in order for successful collaboration in assessment practices (Baker and Riches 2018; Jin 2010).

36.4 Methodology

The revision of the rating scale involved three groups who contributed to different stages of the revision process: testers, experienced teachers, and apprentice teacher-raters. Testers included the faculty supervisor of the EPT and a graduate research

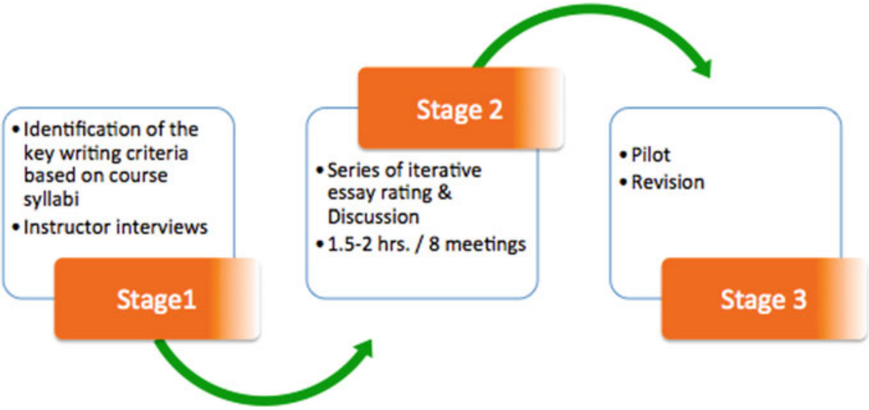

Stage1
• Identification of the key writing criteria based on course syllabi
• Instructor interviews
Stage 2
• Series of iterative essay rating & Discussion
• 1.5-2 hrs. / 8 meetings
Stage 3
• Pilot
• Revision

Fig. 36.1 Stages of essay scale revision

assistant who had experience both as a testing coordinator and instructor in the ESL program. The experienced teacher group consisted of three teaching assistants (TAs) and two lecturers who had taught in the ESL program for more than 3 years. The third group, apprentice teacher-raters, consisted of both lecturers and graduate teaching assistants in the ESL program with varying levels of teaching experience ranging from 1 to 2.5 years. Teacher-raters include the other 8 lecturers in the ESL program. The three TAs participated in Stage 1 of the revision process (instructor interviews); the two testers and two experienced lecturers participated in Stage 2 (iterative rating and discussion sessions and scale development); and all apprentice and experienced teacher-raters participated in Stage 3 (scale piloting).

Figure 36.1 illustrates the scale revision process for the EPT. Following a data-driven scale development approach (Fulcher et al. 2011), the revision process was guided by three purposes: (1) to identify the levels of writing proficiency among the ESL students at UIUC; (2) to identify the number of writing skill profiles among the ESL students at UIUC; and (3) to explore the possibility of developing a rating scale that integrates the placement options (proficiency levels) as well as the writing profiles. The path to achieve these three purposes took place across three distinct stages. The first stage consisted of an analysis of existing materials (with respect to previous iterations of the placement test and the syllabi of the courses students taking the test could potentially be placed into) as well as interviews of instructors. The second stage was dedicated to an intensive series of evaluation-discussion sessions that took place across two months. As a result of these efforts, a new (revised) rating scale was developed. Finally, during the third stage, the new scale was piloted and refined during a series of iterative rater training and calibration meetings with experienced instructors. More detailed information about each of these stages is included below.

36.4.1 Stage 1: Analysis of Existing Materials and Instructor Interviews

The purpose of Stage 1 was to identify the key writing criteria and range of writing performance profiles through an analysis of the existing course syllabi review and interviews with experienced instructors. The previous iterations of the EPT scale had been created without explicit input from instructors. In fact, in the most recent version of the test prior to the revision, there were only two possible placement levels (e.g., low vs. high levels), which was too limited to provide useful diagnostic information about the students' writing performance. By interviewing experienced instructors, the revision team was able to get a better idea of the full range of possible performance levels in the writing courses. Additionally, by performing a comparative analysis of writing course syllabi, the revised scale was grounded in the writing program curricula in the hopes of more accurately aligning with the key writing criteria and student outcomes proposed for the writing courses. In doing so, the revised scale was designed to be more "user friendly" for the ESL writing instructors when rating essays for the EPT. Finally, by incorporating the findings from the instructor interviews and analysis of existing materials, it was hoped that the new scale would help the placement test inform writing instruction by giving instructors a concise summary of the incoming students strengths and areas for growth before they entered the class.

36.4.2 Stage 2: Developing the New Scale

After identifying key writing criteria and writing performance profiles, the next step in the process was a series of iterative essay rating and discussion sessions. There were eight such meetings that took place across three months with each session lasting approximately two hours. During each session, each individual rater would read and rate 20 EPT essays randomly selected from previous semesters after which, the group would reconvene and discuss their ratings and rating process. Following the conclusion of the rating and discussion sessions, a profile-based rating scale was produced that reflected the decision-making process employed by the experienced raters, integrated the key evaluation criteria, and posed guiding questions to help raters assign writing profiles.

36.4.3 Stage 3: Refining the Scale and Training for Its Use

Finally, after its development, the first iteration of the scale was piloted by a group of 8 lecturers within the ESL Writing program. Over the course of one semester, the pilot instructors rated 55 essays in a series of 4 batches, with each batch containing 10–15 essays. In addition to the scale, the instructors were also provided with a glossary of terms to encourage consistent interpretation of the descriptors. Instructors were asked to rate the essays and provide comments about their ratings as well as feedback about the scale through a *Qualtrics* online rating and feedback form. After submitting their ratings, the instructors would attend a one hour rater training session during which one of the test administrators would lead the group through a discussion of five benchmark essays (one for each performance profile) to raise awareness of features common to each profile and help instructors better understand how to distinguish between the profiles. Finally, at the end of the semester, the feedback from raters was compiled and changes to the original scale were made in preparation for its use as part of a rater certificate program beginning in the subsequent semester.

36.5 Findings

36.5.1 Stage 1: Lexico-Grammar vs. Argumentation: Mismatch Between Curricular Emphasis and Actual Writing Performance and Needs

An examination of the course syllabi revealed that the ESL writing courses at UIUC were heavily focused on rhetorical features (e.g., paragraph structure, writing convention), whereas less attention/focus was given to lexico-grammar (e.g., word choice, syntactic complexity, and accuracy). Over the course of the semester, students are instructed on a range of topics such as constructing thesis statements, paragraph structure and argument development, understanding how audience and purpose can shape a piece of writing, choosing and evaluating sources, and avoiding plagiarism. There are two levels of ESL courses at UIUC, at both the graduate and undergraduate levels, corresponding to the two placement levels on the old EPT scale. The main focus of the lower-level course (ESL 111/ESL 511) is on introducing students to American academic writing at the paragraph level—its basic structure, development, and patterns of organization. The upper-level course (ESL 115/ESL 515) focuses more on the practices of research-based writing for American academic audiences, such as developing a research question, searching library databases, creating an annotated bibliography, synthesizing sources, and drafting and revising research papers. General principles of academic writing, such as awareness of audience and purpose, coherence and unity, clear thesis statements, paragraph structure, and formal academic style, are also discussed in this course. However, from the interviews with experienced teachers, it became apparent that a binary scale, separating students into higher and lower proficiency levels, was overly simplistic. In reality, there seemed to be different writing ability profiles within each placement/proficiency level. For example, all instructors mentioned that some students presented strong lexico-grammar control but required more support with respect to source use and organization. Conversely, instructors noted the presence of students with a good command of argument development and logic who demonstrated difficulty in expressing themselves clearly due to a lack of linguistic resources. Moreover, the need to improve rhetorical features does not necessarily differentiate writers across educational levels (e.g., undergraduate vs. graduate students); nor is the need for grammar necessarily a function of overall language or writing proficiency. The mix of profiles and needs can occur even in advanced-level writing courses. The analysis of course syllabi and interview data with experienced instructors revealed a mismatch between the curricular foci and actual performance profiles. This mismatch creates a problem of priority for teachers during instruction as well as during the rating process for the EPT. That is, when giving feedback to students, although teachers might identify salient lexico-grammatical issues, because of the structure of the syllabus, they tend to spend less time on lexico-grammar. Because of this mismatch, there had been a great degree of variability and inconsistency among the teacher-raters in terms of how they operationalize writing proficiency and how they weigh different textual features such as grammar, rhetorical structure, citation, and plagiarism. Some raters even stopped using the rubric when rating EPT essays and relied on their impressions instead. This mismatch provided the rationale for revising the EPT rating scale to better reflect the range of writing performance and instructional needs for different profiles.

36.5.2 Stage 2: Content vs. Structural Features: Conflicting Ratings Between Teachers and Testers

During the second stage of the revision process, two experienced teachers and testers rated and discussed four sets of 20 essays, in an attempt to agree on a common set of criteria to describe the levels and profiles of writing proficiency represented in the test performance. The rationale for the sessions was to identify textual features used to differentiate essay performance and then incorporate these features into the new scale as revised descriptors.

As the discussions progressed, all raters placed essays into three levels (“high,” “mid,” and “low”) without prior agreement. During the post-rating discussion, all raters were found to consistently employ three criteria for rating the essay: source use, argument development, and lexico-grammar. In terms of agreement, the teachers and testers tended to agree on essays of exceptional quality. These essays tended to demonstrate a good command of lexico-grammar and effective argumentation. However, there was notable disagreement on how raters weighed the different criteria; that is, the testers prioritized essays that demonstrated stronger lexico-grammar; however, the teachers overlooked lexico-grammatical errors and instead prioritized essays that demonstrated writing conventions similar to those featured by the existing curriculum (e.g., essays with well-organized paragraphs of balanced length, in-text citations). This contrast resulted in conflicting rankings on certain essay profiles. For example, an essay received conflicting rankings from the two groups. This essay has excellent lexico-grammar but was missing a clear stance in the introduction paragraph; this essay was ranked among the middle of the essays by the teachers, but was ranked among the top third by the testers. More interestingly, teachers and testers also showed differences in the rating of argumentation. When rating the effectiveness of argumentation, some teachers relied heavily on surface/structural rhetorical features rather than focusing on the essay content. A typical profile that elicited this contrast is referred to as "TOEFL template essays" by one of the teachers. This type of essay tends to have an elaborate introduction paragraph with sophisticated lexico-grammar; however, the content in the paragraph does not necessarily align well with the topic of the prompt, and the complexity of lexico-grammar tends to drop noticeably in the body paragraphs. In addition, with respect to paragraph organization, this essay profile tends to have clear topic sentences and evidence from the sources in each paragraph, but the supporting details in each paragraph tend to be list-like, irrelevant, or not closely linked to the topic sentence. Through rounds of discussion, it became clear to teachers because of two main reasons. First, the course syllabi are organized by more structural aspects of academic writing, so it is easier for teachers to simply focus on the structural aspects of academic writing (e.g., focusing more on paragraph structure instead of effectiveness of argumentation). Second, the old EPT scale had as many as eight criteria, which prevented the raters from focusing on content under timed conditions. As it was less time-consuming, most raters focused on the surface features of the writing. While the "TOEFL template"-type essay appeared to be an easy agreement among raters after rounds of discussion, the teachers and testers could not convince each other on the relative importance of lexico-grammar and argumentation. However, through rounds of discussion, each group became more aware of the importance of both lexico-grammar and argumentation and started to weigh the two categories more equally.

36.5.3 Stage 3: Solution: Emergence of a Profile-Based Rating Scale

The equal weights between lexico-grammar and argumentation induced a total of five writer profiles across four proficiency levels. These profiles were incorporated in the initial draft of the new scale, visually represented in Fig. 36.2. These five profiles not only align with the experienced teachers' perception of writer profiles in the classroom during the interviews, but also covered all the "conflicting" essay profiles that emerged from the rater discussion stage. For example, the C profile is dedicated to the "TOEFL template-like" essays, with acceptable lexico-grammar but rather list-like argumentation. However, among the five profiles, it was difficult to differentiate the proficiency levels between B1 and B2, due to the relative strength of lexico-grammar vs. argumentation. Therefore, we classify both profiles under the same writing proficiency level. Taken together, the five essay profiles represent four levels of overall writing proficiency, and these proficiency levels rendered three course

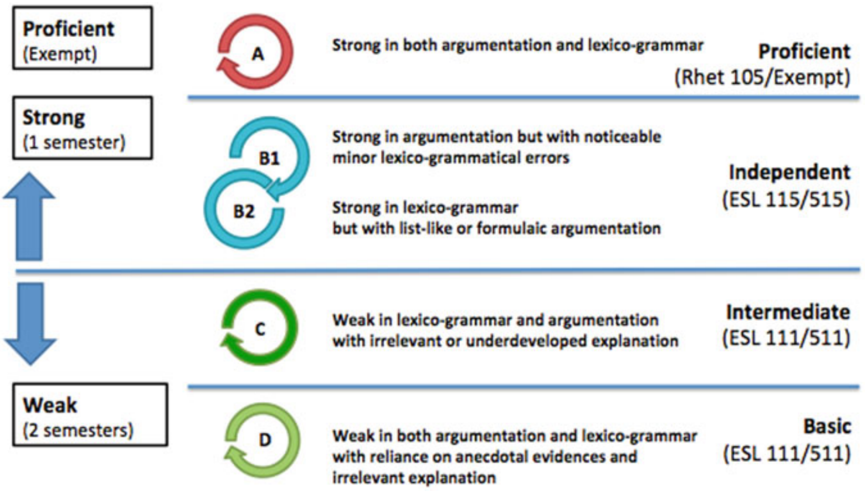

Proficient
(Exempt)
Strong
(1 semester)
Weak
(2 semesters)
A
Strong in both argumentation and lexico-grammar
Proficient
(Rhet 105/Exempt)
B1
B2
Strong in argumentation but with noticeable minor lexico-grammatical errors
Strong in lexico-grammar but with list-like or formulaic argumentation
Independent
(ESL 115/515)
C
Weak in lexico-grammar and argumentation with irrelevant or underdeveloped explanation
Intermediate
(ESL 111/511)
D
Weak in both argumentation and lexico-grammar with reliance on anecdotal evidences and irrelevant explanation
Basic
(ESL 111/511)

Fig. 36.2 Graphic representation of the initial profile-based rating scale

placement options: one-semester mainstream first-year composition (RHET 105: A), one-semester ESL writing course (ESL 115/515: B1 and B2), and two-semester ESL writing course (ESL 111/511: C1, C2, D) (see Table 36.1).

The initial draft of the new EPT scale was piloted and refined iteratively with all teachers ($N = 9$) through the course of a semester in Fall 2016. Rater training and calibration was provided for each iteration. As the scale was refined, we noticed that some descriptors in earlier versions of the scale were not the same as the language regularly employed by the writing teachers in the program for essay evaluation and feedback. Based on rater feedback, the descriptors were modified to align with the assessment language teachers use to describe/evaluate essay performance in the classroom. For example, the term "Premise" was changed to "Topic Sentence," to refer to the main point made within a paragraph. After several rounds of refinement, and perhaps due to these changes in language, teachers appeared to perform better in the rating assignment.

Overall rater agreement for each round is summarized in Fig. 36.3. Over time, raters improved in terms of exact profile agreement (black bar, increased from 33% to 55%), placement agreement (black and gray bars combined, increased from 41% to 92%), and adjacent agreement (black, gray, and white bars combined, increased from 58% to 100%). Agreement figures for individual raters also improved over time (see Fig. 36.4). These results suggest that raters started to adapt to the new rating scale and could use them to rate essays fairly effectively.

Nonetheless, through rounds of rater training and calibration sessions, we also noticed that the initial five profiles that had emerged from the rater discussions were too specific and narrow to cover all possible essay profiles. There was a need to expand the C-level writer profiles. A sixth essay profile emerged at the C level, that is, essays with acceptable argumentation but noticeable lexico-grammatical errors that tend to

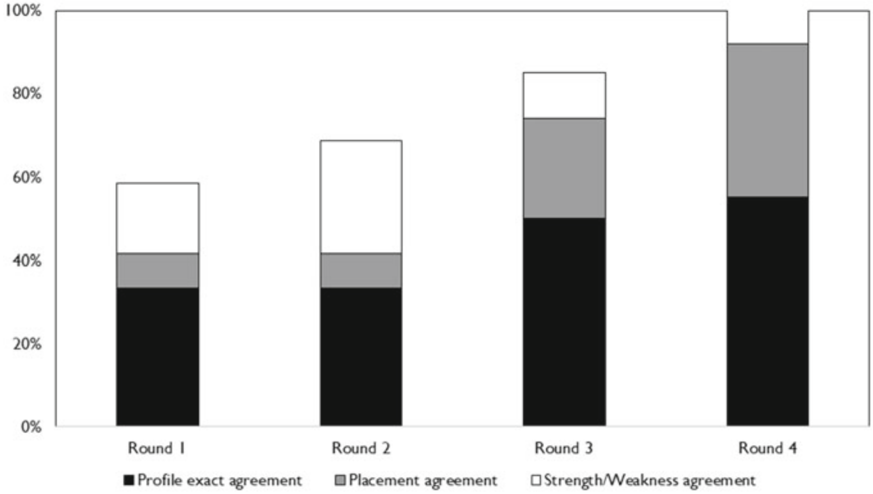


Fig. 36.3 Overall rater agreement on the new EPT scale over time

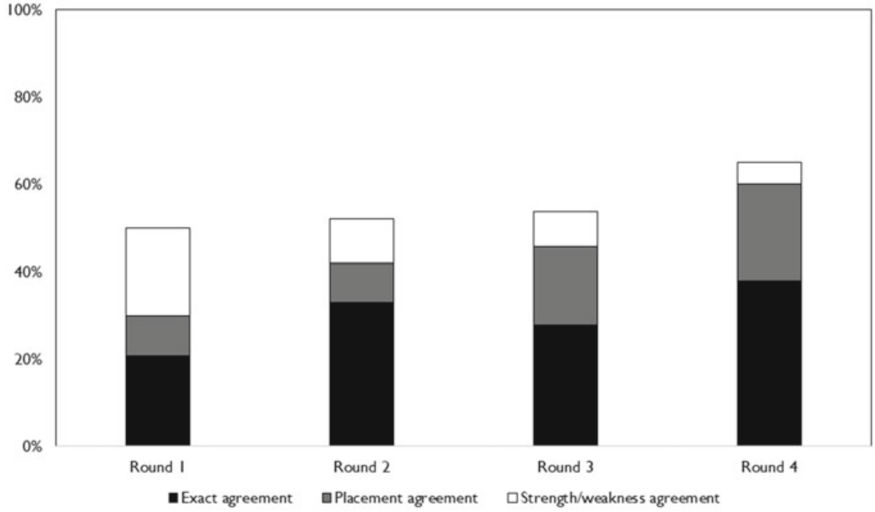


Fig. 36.4 Individual rater agreement on the new EPT scale over time

cause processing difficulty on the raters or other related comprehensibility issues. In addition, some essays that raters placed under the original C profile did not necessarily resemble a TOEFL template (e.g., list-like argumentation), but were rather vague and unclear in their argumentation. While reflecting upon these discussions, we realized that the six profiles can be better classified as different combinations of relative strengths and weaknesses in lexico-grammar and argumentation. These

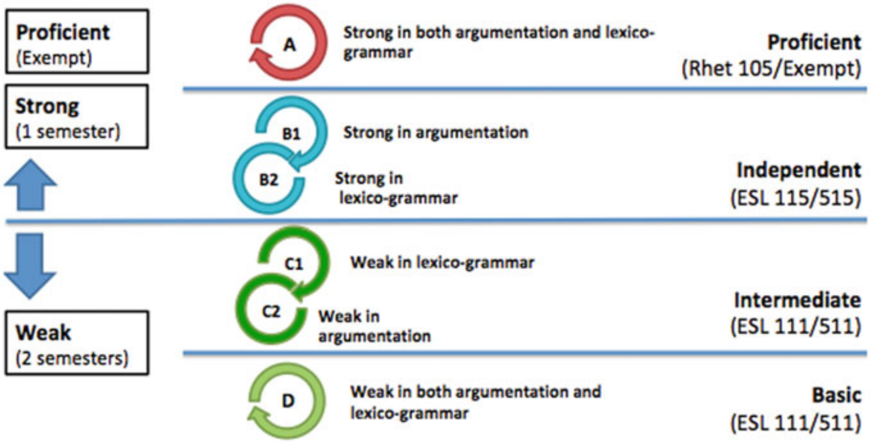


Fig. 36.5 5 graphic representation of the revised profile-based rating scale

classifications were still able to cover all the typical profiles we observed in the test performances and also resolved the issue of scale descriptors being too specific and particular. Based on these findings, we revised our scale (see Fig. 36.5 for the visual representation of the revised scale).

The revised scale is a hybrid profile-based decision tree that provides both profile descriptors and placement recommendations. This scale strikes a balance between argument development and lexico-grammar, better covering the range of writing performance among test takers.

36.6 Insights Gained

In this study, we examined how teachers and testers' evaluation of writing performances changes during a scale revision project within an ESL writing program. At the beginning of the revision process, teachers and testers showed differing emphases on the scoring of lexico-grammar and argumentation. However, through iterative discussion and calibration, teachers and testers became more aligned in their scoring process and collaboratively developed a profile-based rating scale to better represent the range of writing performances and instructional needs within the ESL program. Through the findings of this study, useful insight has been gained into the usefulness of a profile-based writing scale in university settings and teachers' development of language assessment literacy through collaboration with language testers.

36.6.1 The Usefulness of a Profile-Based Scale: Combining Placement and Diagnostic Assessment

There are several advantages of a profiled-based rating scale. The most important benefit of the profile-based scale is that it reflects the instructional needs of different writer profiles within the ESL program. Although the new scale requires additional time for essay scoring, it can provide diagnostic information about the students. The diagnostic information can be utilized in the classroom, informing writing instructors about their students' mastery on knowledge and skills relevant to academic writing. From a broader perspective, the profile-based scale addresses the longstanding debate on the importance of lexico-grammar vs. argumentation. To some extent, it can be reasonably argued that successful argumentation requires writers to have a good command of lexico-grammar; however, essays with complex and accurate lexico-grammar do not necessarily entail effective argumentation. In instructional practice, it is not uncommon that first-year composition courses designed for ESL students in US universities tend to focus on teaching rhetorical conventions. In our context, we observed a similar phenomenon, where many teachers say, "*We do not focus on language; we only teach argumentation, not language.*" However, assuming that first-year ESL students have little need of improving lexico-grammar in writing or overlooking this need will likely result in a disservice to ESL students. In our context, because of the de-emphasis on lexico-grammar, it was difficult for writing instructors to maintain a balance between lexico-grammar and argumentation during the scoring process. This scale prompts writing teachers as well as program directors to strike a balance between the two and be more attentive to student needs.

36.6.2 Impact of Tester-Teacher Collaboration: Scale Descriptors as a Lingua Franca for Writing Assessment

The revision process provides important implications for tester-teacher collaboration and teachers' development of assessment literacy. As demonstrated in this study, through iterative rating and recalibration sessions, teacher-raters became more aligned to the scale and with one another. Discussions among teachers and testers helped standardize the conceptualization and operationalization of writing performance features in both assessment and instructional contexts. This process developed within language teachers' awareness of assessment literacy and attention to sound assessment practice. More importantly, the involvement of teachers in scale development and revision promoted collaborative dialogue and practice in writing assessment between testing specialists and writing instructors. In our context, the development of scale descriptors also created a *lingua franca* of writing assessment within the ESL program, strengthening the alignment between the placement test and the writing program curriculum.

36.7 Conclusion: Implications for Teachers and Testers

This chapter reports on the revision process of a rating scale for an ESL writing placement test, where language teachers and testers collaborated to create a profile-based rating scale that provides both placement recommendations and diagnostic information. The revision process standardized the conceptualization and operationalization of writing quality, shifting teachers' focus from surface rhetorical features to essay content and comprehensibility. More importantly, the involvement of teachers in scale development promoted collaborative, assessment-related dialogues and practices between testers and teachers, enhancing the assessment literacy and practice in the ESL writing program.

References

Alderson, J. C. (1991). Language testing in the 90's: How far have we come? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing.* Singapore: SEAMEO Regional Language Center.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Baker, B. A., & Riches, C. (2018). The development of EFL examinations in Haiti: Collaboration and language assessment literacy development. *Language Testing, 35*(4), 557–581.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26,* 5–19.

Bridgeman, B., Cho, Y., & DiPietro, S. (2015). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing, 33,* 307–318.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112–140). Cambridge: Cambridge University Press.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing, 29,* 421–442.

Chung, S. J. (2014). *Quality management and stakeholder accountability: Computerization of a mandate-driven placement test*. Doctoral dissertation, University of Illinois at Urbana-Champaign.

Crusan, D. (2010). *Assessment in the second language writing classroom.* Ann Arbor, MI: University of Michigan Press.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28,* 43–56.

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language.* Unpublished PhD thesis, University of Lancaster, UK.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Fulcher, G. (2003). *Testing second language speaking.* London: Pearson Longman.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing, 35*(2), 271–295.

Gudrun, E., & Aberg-Bengtsson, L. (2012). A collaborative approach to national test development. In D. Tsagari & C. Ildiko (Eds.), *Collaboration in language testing and assessment: Language testing and evaluation* (pp. 93–108). Frankfurt: Peter Lang.

Institute of International Education. (2017). *Open doors: Report on international educational exchange*. New York, NY: Institute of International Education.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377–403.

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing, 27*(4), 555–584.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

Kunnan, A. J. (1992). An investigation of a criterion-referenced testing using Gtheory, and factor and cluster analysis. *Language Testing, 9,* 30–49.

Lee, I. (2010). Writing teacher education and teacher learning: Testimonies of four EFL teachers. *Journal of Second Language Writing, 19*(3), 143–157.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(2), 101–113.

Mickan, P. (2003). *'What's your score?' An investigation into language descriptors for rating written performance*. Canberra: IELTS Australia.

Oltman, P. K., Stricker, L. J., & Barrows, R. S. (1988). *Native language, English proficiency and the structure of the Test of English as a Foreign Language* (TOEFL Research Rep. No. 27, ETS Research Rep. No. 88–26). Princeton, NJ: Educational Testing Service.

Pollitt, A., & Murray, N. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Language testing 3: Performance, testing, cognition and assessment* (pp. 74–91). Cambridge: Cambridge University Press.

Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Purpura, J. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Stiggins, R. J. (1999). Are you assessment literate? *High School Magazine, 6*(5), 20–23.

University of Illinois at Urbana-Champaign. (n.d.-b). *English placement test.* Retrieved September 28, 2018, from https://linguistics.illinois.edu/languages/english-placement-test-ept.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12.

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194–209.

White, E. (2009). Are you assessment literate? Some fundamental questions regarding effective classroom-based assessment. *OnCUE Journal, 3*(1), 3–25.

Xu, Y., & Brown, G. T. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58,* 149–162.

Yan, X., Zhang, C., & Fan, J. (2018). "Assessment knowledge is important, but…": How contextual and experiential factors mediate assessment practice and training needs of language teachers. *System, 74,* 158–168.

Part VI
Closing Thoughts

Chapter 37
Reflecting on *Challenges in Language Testing Around the World*

Betty Lanteigne, Christine Coombe, and James Dean Brown

37.1 Learning from Challenges

Given the focus and motivation of this volume, we thought it would be interesting to search "learning from mistakes" in Google Scholar and found 22,500 responses. On the first page of those responses (10 articles), the fields represented were one each from the architecture, risk analysis, and climate research fields, two from the management field, and *five from the medical field*. Thus, learning from mistakes seems to be a valid approach to improving how things are done in a variety of fields. Since half of those references were from the medical field, we followed up and found that a large movement has developed in USA hospitals based on learning from mistakes. Among other things, they have learned that developing and using simple checklists designed to avoid mistakes can considerably cut hospital deaths (e.g., see Allen et al. 2018, and many other articles and books). If medical doctors can recognize their mistakes and usefully learn from them, surely language teaching professionals can, too.

When we narrowed our Google Scholar search to include "learning from mistakes" and "language teaching," we got only 535 responses, and the first page of 10 articles were all about students learning from their mistakes. Is it the case that only our students make mistakes and learn from them, as the most prominent results on Google Scholar would seem to suggest? If so, that still does not mean that language teachers do not make mistakes or even that they do not learn from their mistakes. Nor does it mean that there are no articles on the topic. After all, we cited Brown (2010, 2012,

B. Lanteigne (✉)
LCC International University, Klaipeda, Lithuania
e-mail: blanteigne@lcc.lt

C. Coombe
Higher Colleges of Technology, Dubai, UAE

J. D. Brown
University of Hawai'i at Mānoa, Honolulu, HI, USA

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_37

2014) in Chapter 1 indicating that at least one language teaching professional found that he could indeed learn from the mistakes he made in language testing, curriculum design, and research methodology. However, as far as we know, there is not much reflection in our literature on this issue. In a sense, all of this supports the need for a volume like this, one motivated to emphasize such challenges, problems, and negative issues, as well as how others have dealt with them. We hope that learning from the experiences of others will thus help us avoid blindly making those same mistakes over and over and over again.

37.2 Connections Throughout the Volume

Generally, each part of the volume has a main theme. All of the chapters in Part I are connected to each other in that they deal with problems that arise in test use policy with regard to interpreting scores, negative effects of tests, or misuse of tests. The main theme of Part II is about how we can learn from thinking about testing world languages; that is, the chapters address tests of languages other than British, Australasian, and North American inner circle varieties of English. The six chapters in Part III focus on learning from program-level language tests and the need for more enhanced levels of assessment literacy for all stakeholders in the assessment process. Chapters in Part IV discuss tests of reading, writing, speaking, and/or listening skills. And the main theme of Part V relates to learning from tests, teachers, and language assessment literacy. Though these five parts are focused on distinctly different themes, that does not mean that the chapters in each part are exclusively about that main theme.

Other connections across the five parts that are not covered by the main themes are also important. For example, the chapters in Parts II and V also tend to stress the importance of assessment literacy. In addition, the chapters across all five parts of the volume represent contexts from all around the world and are either experience-based or data-based. Aside from those in Part II, the chapters in this volume are about language testing in general or English language tests, especially aspects relating to high-stakes assessment like marking, language assessment literacy policies, and practices. And of course, all chapters deal in one way or another with real-world challenges that arise in language assessment and how language teaching professionals of all sorts have dealt with such challenges.

37.3 Insights Gained and Implications for Test Users

In putting together this volume about the challenges of language testing in real contexts around the world, four key takeaways have emerged: (1) the realization that all types of research make valuable contributions to the knowledge base of our profession, (2) the importance of collaboration and its contribution to effective language testing, (3) real-world language use and its relevance to the challenges faced by language teaching and testing professionals worldwide, and (4) the need for increased or enhanced language assessment literacy skills and knowledge for all key stakeholders in the educational context.

Research in a general sense is considered to be a systematic and comprehensive process that involves the study of a particular topic, subject, or phenomenon to gain more in depth knowledge about it. In applied linguistics, Brown (2004, p. 478) describes research as "any systematic and principled inquiry." Thus, as Abraham Maslow (1966, pp. 129, 133) put it:

> There are some who will insist that "scientific" knowledge is and must be clear, unequivocally defined, unmistakable, demonstrable, repeatable, communicable, logical, rational, verbalizable, conscious… But what shall we say, then, about the first stages of knowledge, the precursors of these final forms, the beginnings that each of us can easily enough experience himself. It is both useful and correct to consider as falling within that definition of knowledge all "protoknowledge," so long as its probability of being correct is greater than chance. Knowledge is then seen as more reliable or less reliable but still knowledge so long as its probability is greater than chance.

The first key takeaway mentioned above was that many types of research make valuable contributions to our knowledge base. Indeed, many types of applied linguistics research are represented in this volume from data-based studies using quantitative, qualitative, or mixed-methods methodological frameworks, to more experienced-based action research projects and initiatives. A major takeaway of this volume is the value that all different types of research have for the enhancement of knowledge in our profession and its importance to our development as language teaching/testing professionals. In fact, development and research go hand in hand. The process of doing research and investigating different questions in our field using a variety of different research methodologies and data collection techniques helps us build knowledge in and about the field and also helps us develop our own learning abilities as individuals. The empirical results emanating from the studies in this volume increase our awareness about recent issues and advancements around the world in our field and also act as a means of communication for language teaching professionals.

The importance of collaboration is another key takeaway from this volume, as collaboration helps people learn from each other, and learning from one another and our own challenges is a principal goal of this edited collection. Collaboration is essentially the practice of sharing knowledge and ideas to achieve a common goal. This can mean asking others for feedback and/or their opinions, sharing knowledge, and finding out how other language teaching professionals approach an issue or a problem and/or deal with challenges in their educational contexts. In areas of the world where oftentimes academics are reluctant to collaborate, researchers need to think of collaboration as a way of enhancing ideas and increasing creativity and productivity. Through the academic collaborations and partnerships that are evidenced in 23 chapters in this volume, authors have been able to complement each other's work and evolve ideas much further than on their own. By also looking for partnerships and collaborating externally, language teaching professionals are able to innovate much more quickly and find solutions to problems that they are currently facing. Another related recommendation resulting from this volume is the need for language testers to collaborate with other key stakeholders like teachers, administrators, curriculum developers, parents in some contexts, and even beyond the educational community to members of the work force involved in the language use being tested.

Another key takeaway that emerged from the chapters in this volume is the importance of real-world language use and how it is related to language testing. McNamara and Roever (2006) raise two points pertinent to real-world language use: Real-world communication in a language being tested goes beyond the constraints of the language test tasks, and real-world communication in a language involves integration of multiple aspects of language ability. Bachman and Palmer (1996) indicated that *tasks of real-world target language use* should be the basis for both language instruction tasks and language test tasks. They (in 2010) defined *language use* as "the creation or interpretation of intended meanings in discourse by an individual, or … the dynamic and interactive negotiation of intended meanings between two or more individuals in a particular situation" (p. 34). Such real-world language use involves people communicating their meanings in actual communicative settings situated in social contexts within larger domains of daily life. The extent to which real-world language use is reflected in language test tasks is an essential component of developing an argument justifying use of the developed assessment (Bachman and Palmer 2010). Ultimately, tests of language ability should seek to measure test taker performance of tasks representative of language use in the targeted domain of real-world communication. In this volume 13 chapters relate to the issue of language tests being used to evaluate *real-world* language *use.*

Perhaps the most crucial takeaway from the research represented in this volume is the importance of language teachers and call for them to broaden their knowledge of Language Assessment Literacy (LAL). It is equally important that other key stakeholders possess LAL as they, too, are influenced by and often face consequences due to language tests. In general terms, LAL refers to the knowledge, skills, and principles of language testing and assessment. Assessing student work and language proficiency is one of the most important responsibilities of language teachers, because the quality of classroom teaching is closely associated with the quality of the assessments we employ. As such, it is essential for teachers to possess knowledge of and about the various language tests and assessments that they are often called on to use in their classrooms. Language test takers, as well, are another stakeholder group that require LAL to learn about and understand how they are assessed and how best to prepare for the various assessments that they will experience. Similarly, language test users such as administrators need LAL to help them make responsible decisions based on language test scores. It is therefore not surprising that virtually every chapter in this volume has noted the importance of, and has called for, increased language assessment literacy as one of the chapter recommendations.

37.4 Future Directions—Suggestions, Recommendations

Highlighting challenges described and implications for readers such as test users and students of language testing or language teaching, as well as real-world insights for language testing professionals, was the focus of this edited volume. The chapters in this volume represent a variety of ways by which language teaching professionals have learned more about language assessment through the challenges they have encountered, which is seen in the huge diversity of issues raised by the authors. A great variety of research methods was employed to deal with these challenges.

What has become clear to the volume co-editors based on their experiences in putting together this book is that there is diversity—and sometimes a disconnect—between what language testing specialists think is needed and what is actually germane in real-world language classrooms and other assessment contexts. We encourage language teaching professionals to not be deterred from conducting research on their own language testing challenges because of a lack of sophisticated research skills and statistical knowledge. Local problem solving does not require generalizability, but users in many contexts can make their own connections. There is a huge diversity in contexts, but, as was evident in this volume, similar challenges occur around the world—rubrics, practices, lack of assessment knowledge on the part of administrators and test developers. And we all could benefit from working together and meeting head-on these challenges with language testing around the world.

References

Allen, J. A., Reiter-Palmon, R., Crowe, J., & Scott, C. (2018). Debriefs: Teams learning from doing in context. *American Psychologist, 73*(4), 504.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice.* Oxford: Oxford University Press.

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. 476–500). Oxford: Blackwell.

Brown, J. D. (2010). Adventures in language testing: How I learned from my mistakes over 35 years. In *English Language Testing: Issues and Prospects. Proceedings of the 2010 Annual GETA International Conference* (pp. 18–28). Gwangju, Korea: Global English Teachers' Association.

Brown, J. D. (2012). The perils of language curriculum development: Mistakes were made, problems faced, and lessons learned. In H. Pillay & M. Yeo (Eds.), *Teaching language to learners of different age groups. Anthology Series 53* (pp. 174–193). Singapore: SEAMEO Regional Language Centre.

Brown, J. D. (2014). Adventures in language research: How I learned from my mistakes over 35 years. In J. Settinieri, S. Demirkaya, A. Feldmeier, N. Gültekin-Karakoç, & C. Riemer (Eds.), *Empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache: Eine Enführung (Empirical research methods for German as a foreign and second language: An introduction)* (pp. 269–279). Paderborn, Germany: Ferdinand Schöningh UTB.

Maslow, A. H. (1966). *The psychology of science.* New York: Harper & Row.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA: Blackwell.