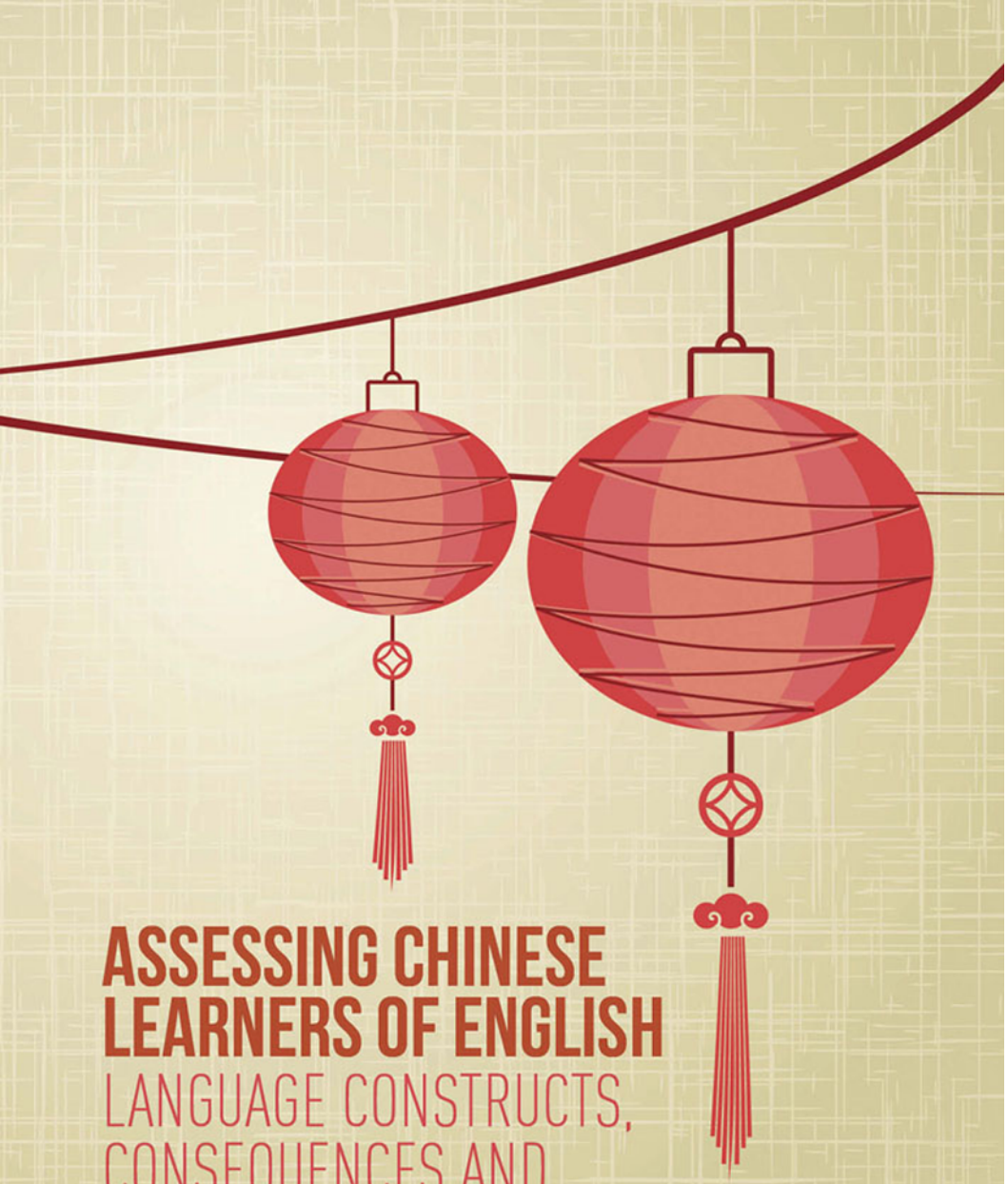


ASSESSING CHINESE LEARNERS OF ENGLISH

LANGUAGE CONSTRUCTS, CONSEQUENCES AND CONUNDRUMS

EDITED BY **GUOXING YU** AND **YAN JIN**

Assessing Chinese Learners of English

Also by Yan Jin

AN EMPIRICAL INVESTIGATION OF THE COMPONENTIALITY OF L2 READING IN ENGLISH FOR ACADEMIC PURPOSES (*co-author*)

Assessing Chinese Learners of English

Language Constructs, Consequences and Conundrums

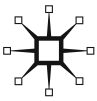
Edited by

Guoxing Yu
University of Bristol, UK

and

Yan Jin
Shanghai Jiao Tong University, China

palgrave
macmillan



First published 2016 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

ISBN 978-1-349-55397-6 ISBN 978-1-137-44978-8 (eBook)
DOI 10.1057/9781137449788

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Assessing Chinese learners of English : language constructs, consequences and conundrums / edited by Guoxing Yu, University of Bristol, UK ; Yan Jin, Shanghai Jiao Tong University, China.
pages cm

1. English language—Study and teaching—Chinese speakers. 2. English language—Ability testing. 3. English language—Conversation and phrase books—Chinese. 4. Second language acquisition—Methodology. 5. English language—Acquisition—Methodology. 6. China—Languages. I. Yu, Guoxing, 1971– editor. II. Jin, Yan, 1965– editor.

PE1068.C5A88 2015
428.0071'051—dc23 2015019855

Typeset by MPS Limited, Chennai, India.

Contents

List of Figures, Tables and Appendices

Figures

Tables

Appendices

Foreword

Guoxing Yu and Yan Jin point out in their introduction that a phenomenal number of Chinese learners of English are taking English language tests. English is one of the three key subjects (the other two being Chinese and mathematics) in Gao Kao – the national university entrance examinations. The College English Test (CET) has the most test takers of any test in the world every year, e.g. in 2012 alone it had 18 million test takers. There has been a substantial increase in the number of Chinese taking international English language tests. In 2010 there were over 300,000 Chinese who took International English Language Testing System (IELTS), and a similar number of Chinese taking TOEFL iBT (Test of English as a Foreign Language, internet-based test). Given the huge numbers of students whose lives are affected by local and international English language tests it is critical that test providers understand how policies and practices of assessing Chinese learners of English as a foreign language are intertwined with the social, political and educational systems in which the tests operate and in turn impact upon.

This volume makes a contribution to deepening the understanding of all those involved in testing Chinese students. It provides empirical evidence for test validation as well as insightful examples of research efforts which help us to better understand the characteristics of Chinese test takers, constructs of assessment (speaking in particular), assessment methods, purposes and impacts of assessment and assessment policies/innovations.

The authors look in detail at the characteristics of the Chinese learners being assessed, what makes Chinese learners of English different from learners of other first languages, the language constructs that underlie some of the tests sat by Chinese learners, various assessment methods and innovations, to what extent the social, political and educational systems in China affect the students' learning motivations and test preparation strategies, Chinese students' performance on a number of English language tests and variables affecting this performance, how different stakeholders cope with assessment policy changes and the consequences of assessment. It is a welcome addition to the increasing number of publications on the assessment of Chinese learners of

English and also contributes to the general knowledge base of English language assessment.

It is now over twenty-five years since I started working with Chinese colleagues on the CET test at Shanghai Jiao Tong University and the TEM test at Shanghai International Studies University. In relation to these Chinese tests alone a substantial contribution to test theory and practice has been made.

The College English Test Validation study I was involved with in 1991–1995 was the first of its kind in China since large-scale standardized language tests came into being in the mid-1980s (and among the first in the world on major examinations). Through collaborative research, the study contributed significantly to the growth and development of professional language testing expertise in China. A full history of the validation project can be found in Yang, H. and Weir, C.J. *Validation Study of the National College English Test* by Shanghai Foreign Language Education Press (1998).

I was also involved with the Test for English Majors (TEM) validation project in 1993–1996. The immediate purpose of the project was to review the existing TEM-4 and TEM-8 in terms of content, construct, predictive and concurrent validity and to establish their reliability through statistical analysis of the test data. By developing enhanced procedures for item writing and marker standardization, it was hoped that future tests would better reflect the English language performance of the test takers. The project's long-term aim was to improve the positive washback effects on ELT teaching and learning in Chinese universities. The study was published as Shen, Z., Green, R and Weir, C.J. *The Test for English Majors (TEM) Validation Study* by Shanghai Foreign Language Education Press (1997).

Such cases of extended international collaboration (as do those in this book) certainly helped meet the local needs in test development and validation by providing a global perspective and also helped to develop the capacity of language testing research and practice in China itself. However, the benefits are never only one-way. The socio-cognitive framework, first comprehensively elaborated in my book *Language Testing and Validation* (Palgrave, 2005) has its roots in my earlier academic work (see *Communicative Language Testing* (1990) and *Understanding and Developing Language Tests* (1993)), which arose out of this earlier collaborative work in China first as senior UK consultant on the national College English Test (Yang and Weir 1998). It developed further in work on the Test for English Majors (Shen, Green and Weir 1997) and the Advanced English Reading Test (Weir, Yang and

Jin 2000). Working with Chinese colleagues on these tests involved developing a clearer specification of the operations and performance conditions underlying language test performance. These provided the conceptual basis for the cognitive and contextual validity parameters that appear in my 2005 book for reading, listening, writing and speaking, which were further developed in the constructs volumes in the Studies in Language Testing (SiLT) series (Shaw and Weir 2007, Khalifa and Weir 2009, Taylor (Ed.) 2011, and Geranpayeh and Taylor 2013) by Cambridge English and Cambridge University Press.

As Bachman (2009) pointed out in his Foreword to an earlier volume in this area edited by Cheng and Curtis (2009) – *English Language Assessment and the Chinese Learner*, "the language testing issues discussed … are not unique to the assessment of Chinese Learners' English …." The studies in this volume similarly make an important contribution to the global knowledge base of English language assessment as well as to our knowledge of the testing of Chinese learners in particular. In addressing Chinese learners of English, the authors make an important contribution to better understanding the complexity and dynamics of assessing Chinese learners of English in different educational contexts and levels. The studies clearly illustrate the need to take into account the social, political and educational contexts in which English language tests and assessment innovations and policies take place in China.

Cyril J. Weir
Centre for Research in English Language Learning and Assessment (CRELLA)
University of Bedfordshire
April 2015

Acknowledgements

As editors, we would like to express our thanks to all the contributors to this volume. We are fortunate to have had their enthusiastic support and willingness to share with us and the readers their experience, expertise and wisdom in assessing Chinese learners of English as a foreign language. Our special thanks are also due to Elizabeth Forrest, Rebecca Brennan and Olivia Middleton at Palgrave Macmillan, for their extraordinary patience and professionalism that helped us to keep the project on track; and to the three reviewers for their constructive feedback and trust.

Notes on Contributors

Editors

Guoxing Yu is Reader in Language Education and Assessment at the Graduate School of Education, University of Bristol. He is an Executive Editor of *Assessment in Education*, and on editorial boards of *Assessing Writing*, *Language Assessment Quarterly*, *Language Testing*, and *Language Testing in Asia*. His articles have appeared in international journals including *Applied Linguistics*, *Assessing Writing*, *Assessment in Education*, *Educational Research*, *Language Assessment Quarterly*, and *Language Testing*. Email: Guoxing.Yu@bristol.ac.uk

Yan Jin is Professor of Applied Linguistics at the School of Foreign Languages, Shanghai Jiao Tong University. She is also Chair of the National College English Testing Committee in China. She is on the editorial board of international journals such as *Language Testing*, *Classroom Discourse*, and Chinese journals such as *Foreign Languages in China*, *Foreign Language World*, *Foreign Language Education in China*, *Contemporary Foreign Languages Studies*. She is also co-editor of *Language Testing in Asia*. Email: yjin@sjtu.edu.cn

Contributors

Chan Chen is a lecturer in the School of Foreign Languages at Zhejiang Gongshang University in China. Her research interests are discourse analysis, language testing and corpus linguistics. Email: 13588161993@163.com

Qiuxian Chen is an associate professor at Shanxi University, China. She has over two decades' experience in English language teaching practice and research. She received her PhD in 2011 from the Queensland University of Technology, Australia. Her expertise is in the field of English as a Foreign Language assessment and assessment policy change. Email: chenqx@sxu.edu.cn

Xiangdong Gu is Professor and Director of the Research Centre of Language, Cognition and Language Application in Chongqing University, China. She holds a PhD in Linguistics and Applied Linguistics from Shanghai Jiao Tong University, and furthered her study and research at the University of California, Los Angeles and

University of Cambridge as a visiting professor. She is an academic consultant of Cambridge English and an external reviewer of several academic journals and university presses in China. She has authored and presented widely on language assessment. Her current interests mainly focus on content validity studies and the impact of large-scale and high-stakes English tests in China. Email: xiangdonggu@263.net

Liz Hamp-Lyons is a senior consultant to the College English Test (CET). She was Head of English and Chair Professor at the Hong Kong Polytechnic University and is now a member of CRELLA (Centre for Research in English Language Learning and Assessment) at the University of Bedfordshire, UK. Her research interests include the development and validation of English language writing and speaking assessments, assessment for academic and specific purposes, learning-oriented language assessment and language teacher assessment literacy. Email: lizhamp-lyons@outlook.com

Daniel M.K. Lam is a PhD candidate in Linguistics and English Language and Teaching Fellow in TESOL at the University of Edinburgh. His primary research interests are in conversation analysis and language testing, in particular the qualitative validation of speaking tests and assessments. Email: s0964731@exseed.ed.ac.uk

Shwu-wen Lin is an assistant professor at the General Education Center, National Taipei University of Nursing and Health Sciences, Taiwan. She holds a PhD in Language Testing from the University of Bristol. She has completed a research project on GEPT Advanced writing funded by the Language Training and Testing Center in Taiwan. Email: andrelsw@ntunhs.edu.tw

Lyn May is a senior lecturer in TESOL at the Queensland University of Technology, Australia. Her research interests focus on second language assessment and pedagogy, interactional competence, and the oracy demands of tertiary study. Email: lynette.may@qut.edu.au

David D. Qian is Professor of Applied Linguistics in the Department of English at The Hong Kong Polytechnic University. He is also the founding Co-President (2014, 2015) and President (2016, 2017) of the Asian Association for Language Assessment. His publications cover a variety of topics in applied linguistics, ranging from standardised English language testing, teacher-based assessment, corpus linguistics and ESL/EFL vocabulary research. As a Principal Investigator, he has directed over 20 research projects funded respectively by the Educational Testing Service,

USA, Research Grants Council of Hong Kong, Language Training and Testing Center, Taiwan, and Hong Kong Polytechnic University. Email: David.Qian@Polyu.edu.hk

Nick Saville is a member of the Cambridge English Senior Management Team and is responsible for directing the work of the Research and Thought Leadership Division. He holds a PhD from the University of Bedfordshire in language test impact, and degrees in Linguistics and in TEFL from the University of Reading. Nick is the elected Manager of the Association of Language Testers in Europe (ALTE) and has close involvement with European initiatives, including the Council of Europe's Common European Framework of Reference (CEFR). His long-term research interests include the implementation of quality management into assessment systems, and the investigation of test impact in school contexts using mixed methods research designs. Email: saville.n@cambridgeenglish.org

Jonathan Schmidgall is an associate research scientist at the Educational Testing Service in the United States. His research has focused on the assessment of oral proficiency, and takes a broad view on how the purpose and context of interaction may impact various components of the validity argument for test use. He recently received his PhD in applied linguistics with a certificate in advanced quantitative measurement in educational research from the University of California, Los Angeles. Email: jschmidgall@ets.org

Chanchan Tang has taught English in Wenzhou No.2 Secondary Vocational School, China, for five years. As a teacher, she is interested in studying how Chinese students can learn English more effectively and has published a few pieces of research on this area. Email: tccwhale@hotmail.com

Jeanine Treffers-Daller is Professor of Second Language Education, Institute of Education, University of Reading, UK. She has published widely on the measurement of vocabulary knowledge and use among bilinguals and second language learners and has co-edited two volumes on this topic: Daller, Milton and Treffers-Daller (2007). *Modelling and assessing vocabulary knowledge* and Richards et al. (2009). *Vocabulary Studies in L1 and L2 acquisition: the interface between theory and application*. She is a member of the editorial board of the *International Journal of Bilingualism* and of *Bilingualism, Language and Cognition*. Email: j.c.treffers-daller@reading.ac.uk

Lei Wang is a professor at the School of Foreign Languages at Zhejiang Gongshang University in China, where she teaches and supervises at undergraduate and graduate levels in linguistics and applied linguistics. Her research interests include discourse analysis, pragmatics, language testing and teaching English as a foreign language in China. Email: wanglei@zjgsu.edu.cn

Yuan Wang is a senior research assistant at the Educational Testing Service in the United States. She has been involved in a variety of research projects on English language assessment and learning. She received her master's degree in TESOL from Teachers College, Columbia University. Email: ywang@ets.org

Xiaoming Xi is Senior Director of the Research Center for English Language Learning and Assessment at the Educational Testing Service in the United States. She has published widely in areas including validity and fairness issues in the broader context of test use, validity frameworks for automated scoring, automated scoring of speech, and task design, scoring and rater issues in speaking assessment. She holds a PhD in second/foreign language assessment from the University of California, Los Angeles. Email: xxi@ets.org

Shasha Xu received her PhD in English Language and Literature from Zhejiang University, with a special focus on the washback effect of a high-stakes language test in China. She has participated in several funded research projects including washback studies and development of computerized adaptive tests. She has published in *Language in Society* and *Journal of Second Language Writing*. Her primary research interests are the impact of large-scale tests on teaching and learning, and the teaching of speaking and writing. Email: xushashaecho@126.com

Lin Zhang is a lecturer at the School of Foreign Languages, Shanghai Jiao Tong University. She has also been working at the Administration Office of the National College English Testing Committee since the year 2004. Her research interests include oral assessment, language test development and validation. Email: zhang_lin@sjtu.edu.cn

Ying Zhang completed her PhD research in Monash University, Australia and currently works as Assessment Manager at the OET Centre, Cambridge Boxhill Language Assessment, Australia. She was formerly an associate professor of English, College of Foreign Languages, Tianjin Normal University, China. Her research interests are language

assessment and testing, English for specific purpose, EIL and TESOL. Email: barbara.zhang@oet.com.au

Ying Zheng is a lecturer at the Faculty of Humanities, University of Southampton. She specializes in psychometric analysis of large-scale language testing data, English as second/foreign language learner characteristics and quantitative research methods. The courses she teaches include Assessment of Language Proficiency, Research & Inquiry in Applied Linguistics and Quantitative Research Methods. Email: Ying.Zheng@soton.ac.uk

1

Assessing Chinese Learners of English: The Language Constructs, Consequences and Conundrums – An Introduction

Guoxing Yu and Yan Jin

1.1 Introduction

In this chapter, we introduce the context and the rationale for the edited volume on assessing Chinese learners of English as a foreign language. In specific, we will discuss the constant challenges and conundrums in understanding the language constructs, the various assessment methods, Chinese learners' preparation for and performance on English language tests, as well as the wide-reaching consequences of assessing Chinese learners of English. This introduction chapter also presents the logic of the sequence of the individual chapters and the overall organisation of the edited volume. The central question that we keep asking ourselves throughout this edited volume – *What have we learned from research on assessing Chinese learners of English?* – helps us to draw together, though very much tentatively, the implications of the findings of the studies reported in this volume which represents our collective endeavours as researchers to contribute to solving part of the conundrums.

1.2 The context and rationale

Understanding how Chinese students are being tested, how they are preparing or being prepared for different purposes, at different educational levels, and for different tests, will lend some insight into not only the validity of the tests per se but also the wider issues in relation to local and global impacts of the tests. English language assessment as a social practice is hugely complex in terms of assessment policies, practices and hence its impacts at different educational levels. The

uses, misuses and abuses of English language assessment transcend the traditional studies focusing exclusively on the reliability and validity of tests. The policies and practices of assessing Chinese learners of English as a foreign language are intertwined with the social, political and educational systems in which the tests operate; as a result, the impacts of English language tests are social, political and educational in nature. As Ross (2008) rightly pointed out: "Language assessments for high-stakes purposes invariably involve policy making at some level. Language assessment policy analysis requires an appreciation of the social, economic, and historical contexts in which assessment policies are introduced, modified, extended, or abandoned" (p. 5).

To understand the current status of English language assessment in China, it is imperative and inevitable that first and foremost we take into account the history of Chinese imperial examinations and the impact of the examinations on the present social, political and education systems. It is widely accepted that China is the origin of large-scale examinations of individuals' abilities for selection purposes (Bowman, 1989; Martin, 1870). Although the system of imperial examinations was abolished in 1905, its influence is still permanently embedded in the present education and assessment systems in China. *Sit for the exam and fight for the rank* – was and still is not only a manifestation of the nature of competitiveness in all aspects and levels of educational assessment in China but also one of the key mechanisms used by the Chinese government to manage resources and social mobility. Issues in educational access, equity and quality (Davey, Lian, & Higgins, 2007; Hannum, An, & Cherng, 2011; Rong & Shi, 2001; Wang, 2008), social justice and political centralism (Feng, 1995) are the main criticisms of the selection purposes of education assessment in China (see Yu & Jin, 2014).

Compared to imperial examinations, English language assessment, which probably started in the 1860s in China (Cheng, 2008; Fu, 1986), is relatively a "small baby" in terms of its history. However, in terms of its size, scope and reach of influence, English language assessment is colossal; it now permeates every aspect and moment of Chinese society. A phenomenal number of Chinese learners of English, from nursery to higher education institutions and beyond, are taking English language tests. English is the compulsory school subject from year three almost everywhere in China, rural and urban. English is one of the three key subjects (the other two being Chinese and mathematics) in Gao Kao – the national university entrance examinations. College English Test (CET) has millions of test takers every year, e.g., in 2012 alone it had 18 million test takers. There has been a substantial increase in

the number of Chinese taking international English language tests. In 2010 there were over 300,000 Chinese who took International English Language Testing System (IELTS), and a similar number of Chinese taking TOEFL iBT (Test of English as a Foreign Language, internet-based test). Educational Testing Service, the owner of TOEFL iBT, reported a 19% increase of Chinese test takers in 2011 from 2010, and a further 32% increase in 2012 from 2011. According to a recent ETS publication (Liu, 2014), Chinese test takers represent about 20% of the TOEFL iBT population. Test preparation courses, especially for TOEFL iBT and IELTS, have been the major income sources of some public listed Chinese companies such as New Oriental at NYSE and Global Education and Technology at NASDAQ which was purchased by Pearson in December 2011. To gain a sense of the scale of English language learning and assessment, this TED video by Jay Walker is particularly telling:

http://www.ted.com/talks/jay_walker_on_the_world_s_english_mania.html

English language assessment affects not only millions of people within China but also has far-reaching global effects, academically and financially, on recruitment and education of Chinese students in English-speaking universities. According to the UK Council for International Student Affairs (UKCISA), there were 428,225 international students in UK higher education institutions in the 2010–2011 academic year; they made up of 48% of full-time research degree students, 70% of full-time taught postgraduates, and 14% of full-time first degree students. Several UK universities recruited a substantial percentage of their students from overseas (e.g., LSE 66%, Imperial College 40%, UCL 38%, Cambridge 30%, Warwick 30%, and Edinburgh 28%). In the USA, there were 723,277 international students in colleges and universities in 2010–2011 academic year. In Australia, there were 184,830 international university students enrolled as of July 2012. In New Zealand, there were 22,811 international university students enrolled as of April 2012 (around 13% of university enrolments). China is the leading place of origin for international students enrolled in the aforementioned countries; and the number of Chinese students has been increasing substantially year on year. For example, UK higher education institutions enrolled 17% more Chinese students from mainland China in 2011/12 than 2010/11 (Source: UKCISA). As a well established but highly debatable, global practice, universities use students' English language test results as one of the most important admission criteria (Rea-Dickins, Kiely, & Yu,

2007). As a result, we witness an increasing number of Chinese taking TOEFL iBT and IELTS year on year as we described above. These English language tests shape and are shaped by the globalising higher education sector. The English language abilities of Chinese students have an impact on the extent to which the students can access and benefit from their higher education experiences, and affect their lives as students and the overall quality of higher education.

In addition to Chinese from the mainland, there are similarly a large number of Chinese learners and test takers of English in Hong Kong and Taiwan who share in many aspects the cultural, linguistic and educational traditions and values as their mainland Chinese counterpart. In this edited volume, we use Chinese or China as terms associated with the Chinese language and people, rather than as a political entity, unless otherwise stated explicitly.

Among policy makers, curriculum designers, material writers, English language instructors, and assessment professionals, at all educational levels, there are substantial and sustainable interests in understanding the issues surrounding the assessment of Chinese learners of English. A number of academic publications have recently appeared or are under preparation to address these issues. For example, *Researching Chinese Learners: Skills, Perceptions and Intercultural Adaptations*, (Editors, Jin & Cortazzi, 2011, Palgrave), *English Language Assessment and the Chinese Learner*, (Editors: Cheng & Curtis, 2009, Routledge), *English Language Education and Assessment: Recent Developments in Hong Kong and the Chinese Mainland*, (Editor: Coniam, 2014, Springer Singapore). *Assessment in Education* (Taylor and Francis) published a special issue on the assessment of Chinese learners of English, edited by Yu and Jin (2014). *Language Assessment Quarterly* (Taylor and Francis) published a special issue on English language assessment in Taiwan (Guest Editor, Vongpumivitch, 2012). Another special issue on high-stakes English language testing in China is under preparation by Professors David Qian (a contributor to this edited volume) and Alister Cumming (OISE, University of Toronto), to be published by *Language Assessment Quarterly*. Together, these publications make incremental contributions to understanding the constructs and consequences of assessing Chinese learners of English.

1.3 The chapters

Given the nature and scope of the complexity of the issues in assessing Chinese learners of English, no single volume would be able to capture all. This edited volume is intended to provide some insights into

language constructs of assessment, various assessment methods and innovations, Chinese students' preparation for and performance on a number of English language tests, and consequences of assessment. These chapters are arranged broadly in line with the fundamental questions that have been continuously challenging the field of language assessment: who, what, how and why to assess.

- What are the characteristics of Chinese learners we are assessing?
- What makes Chinese learners of English different from learners of other first languages?
- To what extent do the social, political and educational systems in China affect the students' learning motivation and test preparation strategies?
- How are Chinese learners being assessed?
- What are the underlying language constructs of assessment?
- What is Chinese learners' performance in English tests, and what affects their performance?
- What are the consequences of assessment?
- What are the policy and pedagogical implications of requiring students to reach a certain English language proficiency level before they are allowed to graduate?
- How do different stakeholders cope with assessment policy changes? For example, how do teachers implement formative assessment in response to government assessment mandate?

These are the main questions that the research studies reported in this edited volume endeavour to address, from different perspectives. The authors of the chapters come from Australia, mainland China, Hong Kong and Taiwan, UK and USA. Some are seasoned researchers who have published widely on language assessment, and some are recent PhD graduates; however, it is our shared experience in assessing and working with Chinese learners of English that brings us together to address collectively a number of perennial issues in assessing Chinese learners of English.

Below we briefly introduce the focus of each chapter.

In Chapter 2, Hamp-Lyons, as one of the main architects of School-Based Assessment (SBA) in English in Hong Kong, reflected on the aims and structure of SBA, and the challenges and issues in developing and implementing SBA in this fervently examination-oriented society. SBA is a typical example of Hong Kong government's initiative to address the dominant culture of summative assessment in schools. As a kind of teacher-based assessment, SBA is intended to serve both summative and

formative purposes, in the high-stakes English Language examinations for secondary school students at age 15–16/16+. SBA was introduced by the government, seeking a balance between summative and formative assessment to make a major educational shift in assessment; however, it met with strong resistance from teachers initially. Hamp-Lyons explained that some of the cultural and political influences helped and hindered the effective implementation of SBA in Hong Kong. She argued that a rigorous teacher's professional development programme and a carefully developed and validated set of assessment criteria and standards are two essential components for successful implementation of assessment innovations. Although SBA is increasingly being accepted by teachers and other stakeholders (including researchers), there are a number of issues that have remained problematic in the nearly ten years of this innovation. In this Chapter, Hamp-Lyons highlighted two of these issues. The first issue is related to the planning or preparation time for Group Interaction tasks in SBA English. Thanks to the "test prep" culture that is "ubiquitous" in Hong Kong, variation in planning time for the Group Interaction tasks could potentially pose threats to the validity of the tasks. The second issue has something to do with the different interpretations of "fairness" – fairness often viewed as equivalent to reliability in the examination-oriented societies, and fairness in terms of opportunity for learning, an opportunity for every student to develop and demonstrate their knowledge and ability to the best of their capabilities. This chapter clearly demonstrates what Ross (2008) argued, which is that language assessment policy analysis requires an appreciation of the broader social, cultural and political contexts in which educational assessment policies or innovations operate, but more importantly, Hamp-Lyons presented a very interesting and thought-provoking first-hand, first-person narrative of the challenges and issues that SBA English faced and still faces in the nearly ten years of implementation.

Following on the same topic, but from a more technical perspective of the implementation of SBA English, Lam reported in Chapter 3 a validation study on Group Interaction tasks. Lam observed that there was a considerable variation in the amount of planning or preparation time given to students for Group Interaction tasks – one of the two continuing challenges that SBA English faces as Hamp-Lyons pointed out in Chapter 2. He looked at how the task was implemented in schools and the authenticity of engagement in student interactions. Based on conversation analysis of student interactions and the stimulated recall interviews with students and teachers, Lam reported that the spoken

discourse of the Group Interaction tasks exhibited some superficial features of authentic interactions and that the students' pre-task planning activities revealed the "contrived and pre-scripted nature" of such interactions. The interactions observed were essentially a "staged performance of pre-scripted dialogues", in other words, "the product of students acting out a composed dialogue based on their knowledge and perceptions of what interactional competence is, rather than students' spontaneous performance of the competence that involves moment-by-moment monitoring of and contingent reaction to each other's talk in real time". The findings of this study can have important implications for designing SBA Group Interaction tasks and the assessment criteria. More generally, as group and paired speaking tasks aiming to assess students' interactional competence often have "planning time" as a key task condition, the findings of this study offer further evidences on the effects of planning time on the features of interactions in such tasks. The next three chapters (4–6) continue the same topic on speaking assessment. Chapter 4 reports on the communication strategies used by test takers in computer-based and face-to-face discussion tasks, Chapter 5 on test takers' use of single words and multi-word clusters in a paired speaking test, and Chapter 6 on test takers' use of formulaic sequences (similar to multi-word clusters in Chapter 5) in a monologue story-retelling task.

In Chapter 4, Jin and Zhang reported a small-scale exploratory study investigating the comparability in test takers' use of communication strategies in two different modes of speaking tasks. Data were collected from six pairs of test takers who sat both the computer-based and the face-to-face College English Test – Spoken English Test (CET-SET). Like Lam in Chapter 3, Jin and Zhang conducted conversation analysis of test takers performance in the two discussion tasks, and found a high level of similarity in both the quantity and variety of communication strategies used by the test takers. They also reported that test takers were generally capable of making effective turn-taking decisions in the computer-based discussion task. Furthermore, in both computer-based and face-to-face discussion tasks, test takers who were awarded a high score on communicative effectiveness made more frequent use of interaction strategies while low performers made more frequent use of production strategies. This small-scale study provided some supporting evidences for the implementation of computer-based CET-SET discussion tasks. Given the number of students taking CET annually, these are particularly welcoming evidences to support the on-going reform and improvement of the delivery of the test.

In Chapter 5, Xu compared the basic spoken vocabulary used in face-to-face interactions by Chinese learners of English and English native speakers. Xu analysed the high-frequency single words and multi-word clusters in the College Learners' Spoken English Corpus and the broadcast conversation and discussion component of British National Corpus. Xu reported that Chinese university students tended to underuse lexical items of interactive functions (e.g., interactive words, interjection in discourse markers) and clusters of vagueness and approximation function; but they tended to overuse conjunction and hesitation in discourse markers. The analysis of the learner corpus also revealed that Chinese students used only a limited number of multi-word clusters in interactions and that they often used them repeatedly, in a sharp contrast to the diverse use of multi-word clusters by English native speakers in similar contexts or genres. Xu argued that the considerable differences between Chinese learners and English native speakers in their use of single as well as multi-word clusters might be attributable to the lack of emphasis or opportunity to learn these aspects of language in the English curricula in Chinese schools. She suggested that "interactive words, discourse markers and clusters of politeness and vagueness functions that enhance communicative competence should be introduced at an early stage of language learning" as the key implications of the findings of her study.

In Chapter 6, Wang and Chen examined the features of formulaic sequences used by test takers in a story-retelling task of Spoken Test for English Majors – Band 4 (STEM4). Test takers listened to a story (about 300 words) twice, taking notes while listening, and then retold the story within three minutes, without any preparation time after listening. To some extent, the story-retelling was a listening/speaking-integrated task, as test takers had to understand the source before being able to retell the story. The extent to which test takers used formulaic sequences directly from the source text or modified them could provide some glimpses of (a) the role that short-term memory might have played for successful completion of the task and (b) the validity of story-retelling task as a measure of speaking ability. The use of formulaic sequences from the source was found to be helpful for test takers to construct fluent texts with less effort. However, the formulaic sequences in the source text were not readily useable unless test takers made a full use of language knowledge and their ability to memorize (though short-term) formulaic sequences to reproduce meaningful and grammatically correct sentences in English. In other words, memorization of formulaic sequences alone did not guarantee successful completion of the

story-retelling task. Story-retelling may be an old-fashioned method for assessing speaking ability, but useful as a pedagogical task for developing speaking ability.

Chapters 5 and 6 investigated Chinese university students' lexical knowledge as demonstrated in their speaking test performance. In Chapter 7, Tang and Treffers-Daller reported the effects on incidental vocabulary acquisition of six reading tasks in a secondary vocational school. The six reading tasks had different levels of involvement ("need", "search" and "evaluation") according to the Involvement Load Hypothesis (ILH) proposed by Laufer and Hulstijn (2001). As Tang and Treffers-Daller pointed out, learning English vocabulary is particularly challenging for Chinese students because the typological distance between the two languages means that there are hardly any cognates between Chinese and English. The results of the experiments showed that students learned more words in reading tasks with a higher involvement load and they also retained more words as shown in a delayed and unexpected post-test. In terms of the contribution of the three different components of involvement – "need", "search" and "evaluation", Tang and Treffers-Daller found that "evaluation" was the most important and "search" the least important of the three. When involvement load was the same, students who carried out output-oriented reading tasks did not outperform those who did input-oriented reading tasks.

Unlike chapters 2–7 that report Chinese learners' performances in English speaking or vocabulary tests, the studies reported in the next five chapters (8–12) focused on the perceptions of the two key stakeholder groups, learners and teachers. Stakeholders' perceptions are essential for test validation purposes. In specific, chapters 8–11 report Chinese learners' attitudes and reactions to assessment innovations and assessment policy changes; and Chapter 12 reports teachers' attitudes towards the use of Standard English and Chinese English in assessing Chinese learners.

SpeechRater℠ is an automated scoring system which is used to provide quick score feedback on the speaking section of the TOEFL® Practice On-line (TPO) test. Xi, Schmidgall and Wang (Chapter 8) investigated the perceptions of 227 prospective TOEFL iBT test takers from China about automated speech scoring and the impact of the use of SpeechRater℠ on their test taking strategies. They also looked at the participants' perceptions, interpretations and uses of SpeechRater scores. The research team administered an online survey to Chinese TPO users of various background characteristics and interviewed 35 of them after the survey. The data suggested that the majority of the

participants considered human scoring more accurate than computer scoring and would prefer human scoring. The combination of human scoring with computer scoring was considered more favourably than computer scoring alone for high-stakes decisions. If only computer scoring were used for high-stakes decisions, the participants indicated that they would try to trick the system. However, there was a good level of acceptance among the participants of using SpeechRater for low-stakes purposes, i.e., for test preparation or practice online. The use of SpeechRater did not change the way they responded to TPO speaking section when preparing for TOEFL iBT. As the authors rightly pointed out, users' perception of automated scoring was much under-researched, but critical for understanding the impacts of automated scoring on test takers' strategies for language learning, test preparation and test taking. In terms of test taking strategies, how would test takers, knowing that they are being assessed by an automated scoring system, respond or interact with various assessment tasks, and to what extent would the outcome and the quality of their speaking and writing performances be affected? The authors suggested a number of other interesting research topics, from users or stakeholders' perspectives, to complement the current studies on automated scoring, which often focus on the technical quality of the systems.

In Chapter 9, Qian reported a study on students' attitudes towards the implementation of project-based group assessment (PBGA) in a Hong Kong university. From the perspectives of assessment for learning, PBGA is widely used for evaluating student works. However, for high-stakes purposes, PBGA as a formal assessment of student academic achievement is less accepted. In this study, Qian administered a semi-structured questionnaire with a sample of 62 English major students (42 first-year and 20 senior-year students) in the English Department of the Hong Kong university. The purpose of the survey was to understand the students' views of using PBGA for assessing their performances in the English language classroom. The data suggested that the senior-year students were more positive than the first-year students. The first-year students tended to focus on negative aspects of PBGA and considered "free-riding" as the main drawback of the assessment method. The senior-year students were generally more positive, although they felt that there were some persistent issues with PBGA, especially the issues of fairness in assessment. As indicated in the data, the majority of the students preferred to be assessed through individual rather than group projects. Other issues such as how to put students into groups, the effects of students' personality, availability, learning motivation and

commitment on group dynamics, the development and use of transparent and fair assessment criteria, were among those discussed by Qian in this chapter. Qian offered some suggestions on how to improve the implementation of PBGA as a formal assessment method, while stressing the importance of taking into consideration the nature of competition at individual level in Hong Kong's education and assessment (see also Chapter 2).

In Chapter 10, Chen and May reported on Chinese university students' reactions to the government's initiative to include formative assessment in College English aiming to promote students' learning and engagement. It was a case study involving interviewing one College English teacher and four of her students, observations of her six lessons consecutively and a survey with 100 students of the College English teacher. In this chapter, Chen and May presented the profiles of the four students – two considered as active in learning and assessment and the other two inactive. The analysis of the profiles of the four students indicated that their responses to the change of assessment policy, especially the inclusion of their performance and participation in classroom towards the final grade that they would receive for College English, were influenced by a number of sociocultural factors. The imbalanced economic development in different regions in China did not seem to have a direct impact on the students' behaviour in classroom or their attitudes towards the change of assessment policy; however, it was evident that students from the more developed areas had higher English language proficiency, especially in speaking, than those from disadvantaged areas. Their previous experience in English language learning and assessment, especially their learning styles, and the degree of their willingness and motivation to play the assessment game, in other words whether they were test or learning-oriented, seemed to be the key factors that influenced these students' responses to the assessment policy change implemented by the university authority. However, it is important to bear in mind that only one student was really responsive to the assessment policy change. The other students would probably be more active in classroom participation if there had been a greater weighting of classroom participation in the calculation of the final grade that they would receive for College English.

Student motivation to learn English is also one of the focuses of Chapter 11. In this chapter, Zheng reported Chinese university students' views on what affected their English language learning and their performance in College English Test Band 4 (CET-4). As part of the larger study, Zheng selected 12 from over 800 students, who responded

to a survey, to conduct in-depth interviews with them. The interview data demonstrated that the students had a variety of motivations to learn English, from contributing to the country's globalization and economic development to more personal reasons. Overall, due to the fact that English is used as a main *lingua franca* in the world, the students attached great importance to learning English; however, they were also mindful that they should keep a balance between learning English and maintaining their Chinese culture and identity. In addition, some students also reported that their motivation for learning English changed over time, particularly from the time when they took the national university entrance examination to the time when they took CET-4. Differences between male and female students with regard to their commitment to learning English were also observed; so were differences between high-proficiency and low-proficiency students in their resources to learn English. The influences from the society, teachers, parents and peers were all instrumental for the students to put efforts into learning English. Passing CET-4 was considered one of the major external forces that motivated the students to learn English since the test provided them with some kind of direction for them to learn the language. The students also made a number of suggestions for improving English language assessment in China with reference to CET-4. For example, they suggested that it would be wise for the test provider to learn from international English language tests to include integrated assessment tasks in CET-4, and that it should make available the speaking test for every test taker of CET-4 rather than just for those who can manage to achieve a certain score in the written test.

In Chapter 12, Zhang investigated teachers' attitudes towards Standard English and Chinese English. The sample of her survey included 20 native English-speaking teachers (ten teaching English as a second language or a language-related subject in Australia; and the other ten working in Chinese universities but originally coming from USA, New Zealand or Canada) and 20 Chinese teachers of English in universities. All the teachers in the study had exposure to Chinese English. It was found that the two groups of teachers were generally in good agreement about the ownership of English and the definitions of the native speakers of English. However, the native English speakers were more open-minded about the use of different varieties of English than the Chinese teachers of English who preferred Standard English (i.e., American/British English) as the norms for learning, teaching and assessment. The teachers working in Chinese universities, regardless of their first language (English or Chinese), were reluctant to accord

Chinese English as a variety of English in its own right, in contrast to the Australia-based native English speaking teachers who considered Chinese English as a language. Teachers' different attitudes towards varieties of English can have a number of implications for teaching and assessing Chinese learners of English. In terms of assessment, teachers who have different attitudes towards varieties of English may operate different constructs of language from teachers who think Standard English should be the norm for assessment; and as a result, their assessment criteria may differ.

The final two chapters (13–14) of this edited volume report the impact of the requirement of English language proficiency for graduation in Taiwanese universities, and the impact of Cambridge English tests in schools in mainland China respectively. A number of Taiwanese universities set an exit requirement for English language proficiency before a student can graduate, aiming to improve students' English language proficiency. The universities have the autonomy to decide whether or not to implement this policy and which tests are acceptable for this purpose. In Chapter 13, Lin reported part of her PhD study that investigated the impacts of the requirement of English language proficiency for graduation. In specific, she examined the impacts of the requirement on the English for Academic Purposes curriculum for non-English majors, in relation to teaching and learning within the English language classroom and students' learning outside classroom. Lin collected data from two universities – one with the requirement for English language proficiency and the other without. Her main data included relevant policy documents from the universities, lesson observations and interviews with teachers and students. One of the findings of her study, reported in this chapter, showed that the locally developed General English Proficiency Test (GEPT) had only limited influence on teaching, compared to international tests such as IELTS and TOEFL iBT. However, the influence of GEPT in local universities was much reinforced by the implementation of the exit requirement and the importance that the general public attached to GEPT because it is arguably the best-known test in the society.

In Chapter 14, Gu and Saville briefly reviewed Cambridge English examinations in China in the last two decades, and looked at the notion of impact in the Chinese context and the effects and consequences that Cambridge English examinations exerted on English language teaching and learning in China. They reported in this chapter two studies on the impact of "Cambridge English: Key for Schools", "Cambridge English: Preliminary for Schools" and "Cambridge English

Business Certificates". In both studies, they used a structured questionnaire and interview as the main instruments to collect data with on a number of important aspects of the impact of the three tests. The main data reported in this chapter included test takers' characteristics, their perceptions of the tests, motivations for and anxiety about taking the tests, and preparation for the tests. In Study One, which examined the impact of the two tests on schools, they found that Chinese students taking the two tests were younger than the targeted groups in the rest of the world. They attributed this to the prevailing culture of test taking in China. Even at a young age, the students were very positive towards the tests and highly motivated, for a variety of reasons, to take the tests. The data also evidenced some negative impacts of the tests, especially in relation to anxiety and extra workload that the students experienced. In Study Two, Gu and Saville looked into the impact of Cambridge English Business Certificates – Vantage and Higher – on university students. Similar findings to those in Study One were noted, with regard to test takers' highly positive views about the tests and high motivations for taking the tests. However, it was also found that the Chinese test takers were less familiar with certain aspects of the tests, e.g., the rating scales for speaking and writing, than similar cohorts of test takers in the rest of the world. They attributed the test takers' lower familiarity or awareness of the rating scales to the fact that Chinese university students are typically not assessed in their day-to-day learning in the same way as in the Cambridge examinations.

In summary, the studies reported in this edited volume covered a diverse, but still focused range of issues that we encounter when assessing Chinese learners of English. Chapter 2 reflected on the issues and challenges in the development and implementation of School-Based Assessment in English in Hong Kong. Chapters 3–6 reported on Chinese students' performances in speaking tests, notably in relation to their communication strategies, interaction and lexical knowledge. Chapter 7 reported on the assessment of secondary school students' vocabulary acquisition through reading. Chapters 8–12 looked at Chinese learners' and their teachers' attitudes and reactions to assessment innovations and assessment policy changes. These chapters covered a range of assessment issues, from automated speech scoring, project-based group assessment, formative assessment, motivation to learn English, and the use of Standard English and Chinese English in teaching and assessing Chinese learners of English. Chapters 13–14 presented studies on the impact of local and international English language tests on teaching, learning and specific test preparation efforts.

1.4 Conclusion

Across the studies, it is evident that we must take into account the social, political, educational contexts in which English language tests and assessment innovations and policies operate in China in order to better understand the impacts of the tests and policy initiatives, and how well Chinese students perform in various tests and why. At the individual level, it is equally important for test validation purposes to investigate Chinese students' attitudes towards assessment innovations, their language learning motivations as well as other personal characteristics which are arguably shaped by the wider social, political and educational contexts.

This volume, as our collective efforts to address a number of perennial issues in assessing Chinese learners of English, makes an important contribution to better understanding the complexity and dynamics of assessing Chinese learners of English in different educational contexts and levels. This volume provides some insights into a number of selected key challenges and issues in English language assessment, e.g., language constructs of assessment, assessment innovations and methods, test preparation and performance, and consequences of assessment. However, they only represent the tip of an iceberg, the enormous challenges that we face in assessing Chinese learners of English; not only because of the sheer number of test takers but also because of the wide-reaching influences that the use of test results exerts within China and globally. To some extent, we agree with what Bachman (2009) wrote in the foreword to the volume edited by Cheng and Curtis (2009) – *English Language Assessment and the Chinese Learner*, "the language testing issues discussed … are not unique to the assessment of Chinese Learners' English at all. Rather, the enormity of the enterprise in this context magnifies kinds of problems that are faced by language testers everywhere, and makes it more difficult to find justifiable solutions" (p. x). In this sense, the studies in the present volume also make essential contributions to the global knowledgebase of English language assessment.

References

Bachman, L. F. (2009). Foreword. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. x–xii). New York: Routledge.

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist, 44*(3), 576–578. doi: 10.1037/0003-066X.44.3.576.b

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15–37. doi: 10.1177/0265532207083743

Cheng, L., & Curtis, A. (2009). *English language assessment and the Chinese learner*. New York: Routledge.

Coniam, D. (2014). *English language education and assessment: Recent developments in Hong Kong and the Chinese mainland*. Singapore: Springer.

Davey, G., Lian, C. D., & Higgins, L. (2007). The university entrance examination system in China. *Journal of Further and Higher Education, 31*(4), 385–396. doi: 10.1080/03098770701625761

Feng, Y. (1995). From the imperial examination to the national college entrance examination: The dynamics of political centralism in china's educational enterprise. *Journal of Contemporary China, 4*(8), 28–56. doi: 10.1080/10670569508724213

Fu, K. (1986). *A history of foreign language education in China (中国外语教育史)*. Shanghai: Shanghai Foreign Language Education Press.

Hannum, E., An, X., & Cherng, H. Y. S. (2011). Examinations and educational opportunity in China: Mobility and bottlenecks for the rural poor. *Oxford Review of Education, 37*(2), 267–305. doi: 10.1080/03054985.2011.559387

Jin, L., & Cortazzi, M. (Eds.). (2011). *Researching Chinese learners: Skills, perceptions and intercultural adaptations*. Basingstoke: Palgrave Macmillan.

Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: the construct of task induced involvement. *Applied Linguistics, 22(1),* 1–26.

Liu, O. L. (2014). Investigating the relationship between test preparation and toefl iBT performance. *ETS Research Report Series*, n/a-n/a. doi: 10.1002/ets2.12016

Martin, W. A. P. (1870). Competitive examinations in China. *The North American Review, 111*(228), 62–77. doi: 10.2307/25109555

Rea-Dickins, P. M., Kiely, R., & Yu, G. (2007). Student identity, learning and progression: The affective and academic impact of IELTS on 'successful' candidates. In P. McGovern & S. Walsh (Eds.), *IELTS research reports volume 7* (pp. 59–136). Canberra: IELTS Australia and British Council.

Rong, X. L., & Shi, T. (2001). Inequality in Chinese education. *Journal of Contemporary China, 10*(26), 107–124. doi: 10.1080/10670560124330

Ross, S. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing, 25*(1), 5–13. doi: 10.1177/0265532207083741

Vongpumivitch, V. (2012). English-as-a-foreign-language assessment in Taiwan. *Language Assessment Quarterly, 9*(1), 1–10.

Wang, L. (2008). Education inequality in China: Problems of policies on access to higher education. *Journal of Asian Public Policy, 1*(1), 115–123. doi: 10.1080/17516230801900444

Yu, G., & Jin, Y. (2014). English language assessment in China: Policies, practices and impacts. *Assessment in Education: Principles, Policy & Practice, 21*(3), 245–250. doi: 10.1080/0969594x.2014.937936

2

Implementing a Learning-Oriented Approach within English Language Assessment in Hong Kong Schools: Practices, Issues and Complexities

Liz Hamp-Lyons

2.1 Background: educational and assessment innovation in Hong Kong

It is becoming increasingly well understood that every educational innovation thrives or flounders within a social-political ideological context (Henrichsen 1989; Kellaghan & Greaney 1992; Wall 2005). Hong Kong has for many years had a traditional norm-referenced examination system for school placement, promotion and exit (for a historical overview of the public examination system in Hong Kong, see Choi & Lee 2009). While this system is congruent with a traditional Chinese cultural heritage context, educators have long felt there is something fundamentally flawed about a system in which students may fail *every* school subject in which they take a formal exam. In the English subject, for example (by no means one with the worst results), between 1997 and 2007 (the last year of norm-referenced results reporting) 41–60% of students failed the Syllabus A English and 60–78% failed the more difficult Syllabus B English. Steps have been taken at several points in the past 30 years to reform the educational system to better fit the needs of school students, and to ensure that the right individuals enter tertiary education and that sound educational opportunities are available to those not entering tertiary education (King 1994; Qian 2008). The need for a more liberal and broad approach to examinations was one of the points made in the 1997 Education Commission Report, *Quality School Education*: this Report may well have had foreknowledge of one of the recommendations of a review of the public examination system which was at that time reaching a conclusion. The review, which has become referred to as the *ROPES Report (Review of public examination system*),

was not made public, but the general outlines of its recommendations became known. Among the members of the review consultancy team was Patricia Broadfoot from the UK, well-known for her progressive views on inclusive and humanistic assessment, and so it is not surprising that one comment made in the review was that worldwide there is "a pronounced shift in responsibility for assessment of student achievement to a blending of the information available from both (traditional and classroom) sources" (as reported by Berry 2008). However, as Fok, Kennedy, Chan & Yu (2006) comment: "Hong Kong is famous for its examination-dominated culture, which heavily relies on public examinations. So ingrained has it become that the whole society is sensitive to any change in such an assessment mechanism" (p. 1).

One recommendation of the ROPES Report was to expand school-based assessment in the Hong Kong Certificate of Education (HKCE) and Hong Kong Advanced Level (HKAL) Examinations. Christina Lee, who joined the Hong Kong Examination Authority (HKEA) in 1990 as English Language Subject Officer and was the General Manager of the Assessment Development Division, describes this history:

"HKEAA has had something what we called 'teachers assessment scheme' (TAS) as early as 1978, so we didn't call it … SBA.., but in effect, they are like the same thing, I would say, you know, the teachers assessment scheme and SBA, but, basically, when I started working with the exam authority in the early 1990s, TAS was already going on in Chemistry, Biology, and then later on Physics also joined." (C. Lee personal interview, 16 May 2007)

Yung (2002) describes the form of 'school-based assessment' implemented in the early years of reform as comprised mainly of fairly traditional formative assessment, in which essay marking is thought of as feedback as well as 'traditional' assessment; and in which 'school-based' referred primarily to a de-centralization of some components of assessment to the schools. In 1999 the Hong Kong Education and Manpower Bureau, the governmental body with oversight of the HKEA, commissioned its own 'Strategic Review of the HEAA" which was published in 2003 (IBM Corporation 2003). While continuing the direction of the 1998 review of the public examination system, this was a very ambitious and progressive review with far-ranging recommendations, aiming simultaneously at financial and managerial reform of the Authority, and at reforming the testing culture in Hong Kong to be less rigid and more supportive of modern thinking in curriculum, teaching and learning. The re-naming of the HKEA as the HKEAA (Hong Kong Examinations AND ASSESSMENT Authority) was a direct result of this review. One of

the most immediate recommendations of the Strategic Review was to "progressively move to the further development of school-based assessment … and to assessment based on defined standards" (p. 100). But the Review also acknowledged that "… Hong Kong must urgently engage in fundamental debate right across its education community – and with the public – to raise 'assessment literacy'. (p. 3). In fact, the consultants' consciousness of the complexity of educational and assessment reform and innovation permeates their review. This recommendation was accepted by the Hong Kong Legislative Council in December 2003 [(LC Paper No.CB(2)634/03-04(01)] and funding was earmarked. As the innovation was gradually introduced in various school subjects, responses from teachers and parents were mixed (Cheung 2001).

2.2 Aims and structure of the Hong Kong SBA in English language

2.2.1 Aims and first steps

In 2004 the HKEAA put out a call for tenders for the development of a new component to the Hong Kong Certificate of Education Examination (HKCEE) English Language syllabus which would be a school-based assessment (SBA) component of speaking, and would first be introduced to secondary forms 4 and 5 in the 2005–2006 school year, and included in exam results reported at the end of the 2007 school year. The purpose of this innovation from the government viewpoint was to align assessment more closely with developments in the English Language teaching syllabus and the Senior Secondary curriculum (Curriculum Development Council 1999). The new syllabus, set by the central Curriculum Development Institute of the Hong Kong Education Department, was to include a speaking component taught and assessed in the classroom by the teacher, which would contribute 15 percent of the student's total English grade. A proposal in response to an invitation to tender for the project was submitted to the HKEAA by a team in the Faculty of Education at Hong Kong University (HKU), with Dr. Chris Davison as Project Leader and Prof. Liz Hamp-Lyons as Principal Researcher in September 2004. The proposal was for a pilot study with ten to 12 schools in the first year; however, the Hong Kong Education Bureau (EDB) were concerned that not to include all schools would be seen as unfair, and determined that this innovation should be introduced full-scale. The proposal was therefore revised in December 2004 and awarded on 24 December 2004.

2.2.2 Early challenges

Work on the design and development of this new approach, to be called School-based Assessment (SBA) – English, began in January 2005 with a deadline of August 2005 for full functionality. The development team was asked to present an overview of and rationale for this new approach in March 2005, to a live audience of about 1,000 teachers. The plan from the Educational and Manpower Bureau (EMB) was for it to be rolled out in September of that year with all secondary schools in Hong Kong (a little over 500) without prior trailing or a familiarization period. This sudden and rapid change in assessment coupled with the inevitable (and essential) changes in teachers' teaching strategies caused anxiety among school principals, parents, and especially teachers, and attracted considerable media attention. This negative attention led to intense scrutiny and discussion at senior education levels and in January 2007 a statement was issued by Dr K Chan, the Principal Assistant Secretary (Curriculum Development) in the Hong Kong Government's EMB and Dr Francis Cheung, then Deputy Secretary General of the HKEAA, strongly supporting the introduction of the SBA and offering consultation on implementation. As a consequence, the plan was changed and every school was given the choice of introducing the new English SBA on the original timetable (i.e., formal reporting at the end of the 2006–2007 school year), or of waiting one year before introduction, or waiting two years before introduction. This staged introduction of the assessment innovation enabled some interesting and informative research, which the HKU consultancy team also conducted for the HKEAA, and which was reported as two 'longitudinal studies', the earlier of which can be found on the HKEAA website:

[http://www.hkeaa.edu.hk/DocLibrary/Resources/Longitudinal_Study-SBA_HKCEE_English_dec2010.pdf].

2.2.3 Structure of the Hong Kong SBA in English Language

The Hong Kong School-based Assessment is tied to the same standards–referenced assessment approach as the rest of the public exam system. However, as the Deputy Secretary General of the HKEAA, Dr. Pook, says on the HKEAA website: http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/sba_definition.html, the SBA component of the English Language exam aims to go further by giving a more comprehensive appraisal of learners' skills and abilities, including aspects that cannot be easily assessed in traditional large-scale examinations. The key audience for the DVDs and web-based materials for the English language

SBA are teachers, and at http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/atl_interrelationship.html Liz Hamp-Lyons explains the interrelationships between assessment, teaching and learning, and Dylan Wiliam discusses the differences between formative and summative assessment.

The major components and requirements of SBA in HKCE English Language are described in Davison and Hamp-Lyons (2009), Lee (2008) and Davison (2007), and are summarized below:

- Students are assessed on four domains of speaking custom-designed to fit the two speaking task-types: individual presentation and group interaction.
- Students are assessed during class time by their own English teacher.
- The assessment is embedded into the curriculum.
- Teachers video or audio record a range of the student assessments they observe and assess, in order to assist with later standardization.
- Teachers submit the best two sets of scores out of three assessment tasks.
- SBA speaking scores constitute 15% of each student's total HKCE English Language result.
- The content for students' speaking is drawn from an extensive reading/viewing programme using four different text types: print fiction, non-print fiction, print non-fiction and non-print non-fiction.

This reading/viewing-into-speaking process is closely similar to a curriculum sequence familiar in both English and other disciplines, where students read or work with other input materials, meet in groups to discuss them and expand their individual understanding by the sharing of viewpoints with others, and the ***group interaction task*** aims to model such sharing. The ***individual speaking task*** is seemingly less 'authentic', but within classroom instruction the individual report-back is quite a common task, as is the more formal individual presentation in the senior secondary years as well as in the universities, in many disciplines. These task types and parameters were established by the consultancy team during extensive consultations with a core group of volunteer teachers from a group of over 50 schools who participated in seminars and workshops, action research in their own classrooms, and the trialing and validation processes of the SBA English Language as it developed. Some of these requirements were predetermined by the HKEAA in order to comply with the curriculum set by the Hong Kong Curriculum Development Council (CDC).

These tasks, the individual presentation (IP) and the group interaction (GI), will not sound very unusual or innovative to most readers, and indeed they are not, and they probably suggest speaking activities that are quite inauthentic, as Lam (this volume) suggests. However, it should be remembered that in Hong Kong there had in 2005 been little overt *teaching* of speaking skills, and that the *assessment* of speaking is itself quite new in Hong Kong (Andrews & Fullilove 1994). The formal speaking test that was introduced in 1994 was a close imitation of speaking tests used at that time in the UK, and was (and remains) extremely formal and narrow, and was not very reliable. Nevertheless, Andrews and Fullilove (1994) argue that the sheer fact that speaking was now officially assessed played a part in bringing teachers' and schools' attention to the teaching of speaking.

However, what was innovative in Hong Kong in 2005–2006 was the expectation that classroom teachers would design and carry out these task types in their own classrooms, and that they would be allowed to score their own students' performances. We quickly found that most teachers needed strong support in developing appropriate reading-into-speaking tasks for their students so that they would have content to work with in fulfilling a task. An IP task might be, for example, describing an interesting character in a book they have read or film they have viewed. Teachers asked for guidance on how to plan the tasks and how to tie the assessable task to the curriculum. Davison (2007) shows an exemplar IP task taken from the early stages of the materials development. The GI task is more complex, as described by Gan, Davison and Hamp-Lyons (2009), being a dialogue or exchange of short turns between two or more speakers. Turns are expected to be comparatively short and quite informal. A student taking part in a GI needs to attend to turn-taking skills, and to be able to initiate, maintain and control the interaction through suggestions, questions and expansion of ideas. These interactive skills are deliberately mentioned and rewarded in the assessment domains, criteria and levels. The expectations of effective student performance on the GI task also include the capacity to speak intelligibly and reasonably fluently with suitable intonation, volume and stress, using pauses and body language such as eye contact appropriately and effectively; to use a range of vocabulary and language patterns that are accurate and varied, and language that is natural and interactive, not memorized or read aloud. The Guidelines also note that some use of formulaic language may appear when appropriate for structuring, but that overuse of set phrases is discouraged. The assessment criteria for the GI can be found at http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/doc/Assessment_Criteria.pdf

2.2.4 Expansion into the new 3–3–4 senior education structure: SBA 2009-present

By the time the SBA for English Language in the HKCEE reported its first results in 2007, the plans to change Hong Kong's senior secondary and university structure from 3–4–4 to 3–3–4 (i.e., from four years of senior secondary school and three years of university to three years of senior secondary and four years of university – or, from the British to the American system) were already well advanced. As a consequence of this major restructuring, the HKCEE was replaced and the Hong Kong Diploma of Secondary Education (HKDSE) was introduced from September 2009. The HKEAA mandated that in the HKDSE the SBA English Language should be revised and extended to cover the full three years of senior secondary education, and in the third year there should be the opportunity for students to use SBA within their chosen elective subjects. The original consultancy team were asked to revise and expand the existing materials for teachers to fit the new requirements. The HKEAA provided support for further materials development and for extension and renewal of the professional development materials and courses. The revised, three-year, SBA English materials and additional professional development (PD) materials were introduced in 2009, and PD courses continued to be run until the funding ended in December 2011.

2.2.5 Seeking balance between summative and formative paradigms

Successfully transitioning schools, teachers, learners and parents into this 'school-based' approach to assessment is much more than a matter of developing and distributing sets of task types and parameters, sample materials, criteria for assessing spoken language in these contexts, and samples of performance. We did all those things. The Guidelines for SBA, which explained the nature of the oral text-types to be assessed, the mandatory assessment conditions, bureaucratic issues such as record-keeping and particularly the importance of standardization, were extensively shared among schools, in seminars run at the Hong Kong University, in schools, and for the teachers' support network HKEdCity [http://www.hkedcity.net/article/project_sba_eng/050902-002/]. The Guidelines were (and are) available online. The team worked with a clear focus on the need to work with teachers to create successful innovation reform. A key aspect in this was 'accountability': not only the education system but perhaps even more the teachers and school principals wanted to be assured of the trustworthiness and fairness of this new kind of assessment.

Given the tensions of the Hong Kong traditional exam context (Hamp-Lyons 2007), we had to find a balance between the formative thrust of AfL and the summative purposes teachers were accustomed to. This balance, though explicitly argued for in the stated goals of the ROPES review in 1998, and again in the IBM report 2003, and despite the movement towards a modest version of school-based assessment in other subject areas (Cheung 2001; Yung 2002), had by no means been achieved before our project began, and as Qian (2014) points out, it has still been only partially achieved. As my colleague Chris Davison has recently commented, "… teachers do not work in isolation; their attitudes and beliefs about change are inextricably linked to those of other members of the educational community, not just colleagues, but supervisors, parents and students" (Davison 2013, p. 271). In addition to the key concepts and principles of AfL, the professional development programme focused on helping teachers to unpack and understand the structure of the assessment instrument so that they can not only use it for assessing students in their own classes at the appropriate times, but use or adapt it within the teaching and learning they do during instructional units (Davison & Hamp-Lyons 2009).

The movement towards AfL begins from the challenge to summative assessment as being inflexible, behaviourist and teacher-directed, and proposes modes of assessment that are more reactive to the needs of contexts and learners, drawing on constructivist views of learning, and emphasizing the role of assessments in helping students learn how to learn. An early draft of the Guidelines, which laid out these principles as well as the processes involved in putting them into practice, was introduced and explained in face-to-face seminars to over 600 local teachers of English during April and May 2005 (immediately after the plans for English SBA had been made public, but before the system itself had been fully developed). The processes schools and teachers use have remained the same since the implementation of the English Language SBA:

- All teachers teaching with SBA meet within their school, view a range of student performances, consider tasks and criteria, and discuss scores.
- Teachers are not required to alter their own scores after discussion, though they may do so.
- Groups of schools' "SBA coordinators" meet once a term to view a range of their own schools' performances, discuss issues within SBA [not only scores] and refresh their shared understanding of what the levels look like (anchor sets are available).

- Scores are not changed, the discussion is advisory.
- Scores from each school are reported to HKEAA, which conducts a statistical "moderation" that may or may not lead to score adjustments.

This very simple description of a process overlays a complex set of procedures that were carefully developed with the aim of achieving two opposing goals: to satisfy the HKEAA and the Hong Kong Education Bureau's desire for rigorous safeguards for reliability and fairness; and to encourage in teachers the kinds of flexible handling of student learning in live classrooms implicit in the *assessment for learning* movement. These Guidelines were revised twice more before the first teachers carried out the first implementation of a formal (i.e., scored and recorded) SBA teaching and assessment sequence with their classes, and they have been revised several times between 2007 and 2012. It was, and is, regrettable that, in their formulation of a simple format intended to make the SBA process seem non-intimidating to teachers, they acquired a bureaucratic flavour and a less than humanistic tone that sits uneasily beside the very humanistic aims of the innovation.

During the initial development stage, as we worked closely with our core group of teachers, they provided feedback on the emerging Guidelines; and as we visited schools and classes and saw these teachers working with SBA, we became more and more convinced of the need for professional development support for teachers working with this very different kind of assessment for the first time. Our collaboration with some of these teachers continued all the way through to the end of the project in 2011. The PD support we developed was of two broad kinds: general language assessment literacy, and specific familiarization/training and standardization materials.

2.3 Supporting teachers

2.3.1 Language assessment literacy in Hong Kong

As Kennedy (2013) has pointed out, increased teacher and student workload, lack of community confidence in school-based processes and even lack of confidence by teachers themselves emerge as key issues during the implementation of SBA in Asian cultures; but he also considers that teachers are reluctant to accept responsibility for high stakes school-based assessment. Carless (2011) has discussed at length the characteristics of examination-driven education in what he calls the 'Confucian-heritage cultures' (CHC) of China, Japan, South Korea, Taiwan, Singapore and Hong Kong. All these cultures have been directly

or indirectly influenced by the Chinese Imperial civil service examination system, success in which was essential for a respectable post in the civil service and thus, a comfortable life style. Elman (2000) emphasizes the competitive nature of the Imperial system and also the draconian regulations surrounding exam-taking. The Imperial exams focused largely on understanding and interpreting Confucian thought as represented by the key Confucian texts, memorization and exegesis of these texts at a high literary level was most prized. Scholars have argued that this long tradition has led both to the examination-oriented approach to education inherent in these cultures, and also to a view of the reproduction of knowledge as a significant purpose of education. Exams which privilege knowledge-telling rather than knowledge-making lend themselves more easily to cheating.

Hong Kong English teachers have somewhat limited teacher education before they begin teaching, and a substantial proportion of those entering schools as teachers are not qualified, but begin work on the basis that they will qualify during their initial years as teachers. Until very recently, it was quite common for teachers of other subjects to be assigned to teach English because their own English was judged (by the school principal) to be good, or "good enough". With these varied backgrounds in pedagogy and subject knowledge, teachers bring not just lack of knowledge about assessment but also their own experiences of teaching, learning and assessment in the rigid environment of the typical Hong Kong secondary school. They bring an ingrained belief in external exams as 'fair' mechanisms of selection, as well as the belief that it is possible to succeed in any exam with enough hard work. Their students bring these same values, and they are likely to have a little 'help' in their own studies and exams from cram schools.

The lack of assessment literacy by teachers is a major stumbling block for any assessment innovation, just as lack of the appropriate knowledge and understanding is a stumbling block for any educational reform. Similar issues have been identified and discussed in mainland China (Chen & Klenowski 2009). Despite a more established structure of English teacher education in Hong Kong, and in spite of the long history of teaching English in elite secondary schools, the situation has been very similar (Boyle & Falvey 1994). This seemingly anomalous situation is in part explained by the social/political history of education and the medium of education in Hong Kong, and is well reported by Evans (2000). 'Shadow education', or private tutoring and 'cram schools', is very prevalent in Hong Kong (Bray 2007, 2012). Chan and Bray (2014), focusing on Hong Kong in particular, comment that, "Much of the

shadow education focuses on techniques for performance in external examinations, and is not consistent with the emphases stressed by teachers and the government." (p. 361). It is often argued that reforms in mainstream teaching are undermined by the use of much more traditional methods such as complete dependence on a single custom-written textbook to any new exam syllabus, and the use of read-repeat-recall (rote memorization) techniques in cram school classes. Although there is very little empirical research into the methods and culture of cram schools, one thing we do know is that cram or tutorial schools in Hong Kong may not be officially registered; tutors in cram schools are often not qualified as teachers, and may have BAs in the subject but rarely more advanced subject qualifications; and the charges in these schools as well as the service they provide are uncontrolled. Tutors rarely attend professional development seminars because they are not on the radar, and might not attend anyway as their legal status is questionable. This "shadow education" system therefore influences students' and parents' attitudes toward every potential educational innovation.

The English Language School-based Assessment was introduced to Hong Kong within the complex context described above, and this reality limited and to some extent shaped what we could do. We were not surprised when teachers expressed concern about this attempt to bring assessment for learning into Hong Kong's English language classrooms, but the opposition of some Hong Kong teacher educators, despite the initiative having been debated and promoted by the government for more than five years, and despite it being officially supported by the universities, did surprise us. Several major articles and letters were published in the Hong Kong English language and Chinese language press, and these were mainly negative/cautionary. For example, the article by Icy Lee (2005), then an assistant professor in education at Hong Kong Baptist University, was titled "Latest initiative adds to burden". Lee wrote:

> "While most educators applaud the move away from norm-referenced to standards-referenced assessment in the public exam component, the implementation of SBA is causing teachers to shudder. The simple reason is that SBA will put the onus on English teachers to take charge of the whole business of assessing students, and to ensure that the assessment is fair and reliable. The task is daunting, and teachers are ill-equipped for it."

This comment sums up well the sentiments of most teachers at that point. Chris Davison and I met group after group of English teachers, in

all of which there were some teachers who spoke up forcefully against the implementation of SBA on the grounds of workload and concern about its fairness. As the project proceeded we learned a great deal about what we could not assume about teachers' assessment knowledge (and also about their pedagogical knowledge), and about what we could not do with the SBA as a classroom-based assessment. We quickly realized that the key to the successful introduction of this planned innovation would be the planned provision of the essential aspects of language assessment literacy to teachers who would be teaching with SBA – some 7,000 teachers from over more than 500 schools during the first five years, as well as provision of extensive familiarization/training and standardization materials.

2.3.2 SBA familiarization/training and standardization materials

During 2005–2007, substantial work and time went into the development of several kinds of training and familiarization materials. Our core group of teachers helped us by allowing project staff into their schools and classrooms to video record both group interactions (GIs) and individual presentations (IPs): we collected more than 500 such videos. These formed the basis of both the familiarization materials, with professional development activities built around them, and the materials to be used by Group Coordinators (key teachers, mainly taken from the core group, who provided locally-based training and support in 39 sub-districts of Hong Kong). The team also developed a face-to-face professional development programme which supported and augmented the training materials, and which was delivered a number of times each year between 2009 and 2012, serving a total of almost 8,000 teachers. During these formal courses teachers saw a wide range of video samples and became familiar with the criteria and scales. Every term, teachers of classes using SBA meet with the SBA coordinator in their own school to discuss video samples from their own students, and to refresh their understanding of the criteria and standards at each scale level by relating the local samples to a selection from the various DVD collections of samples with commentaries we created during the early years of development. The school coordinators then meet with the district coordinators to share both local and 'official' samples, discuss any issues, and 're-calibrate' as a group, feeding back to the teachers in their own schools. At the time of development, this was the most extensive programme of professional development support for English Language school-based assessment that Chris Davison and I were aware of.

2.3.2.1 Statistical Moderation

Added to this school-based approach to accountability, or what we might call *social moderation*, there is also a moderation of results across schools by the HKEAA. The process of statistical moderation was added by the HKEAA during the introductory year of the SBA and continues to the present. The consultancy team argued against this added process, but it was felt by HKEAA and by the advisory committee that, given the negative publicity the introduction of the English Language SBA had attracted, the public wanted the "reassurance" of a form of reliability they could recognize.

The official HKEAA document explaining statistical moderations says: "There are essentially two ways in which differences in marking standards may affect SBA scores. First, teachers in a given school may be either harsher or more lenient than teachers in other schools. Second, they may tend to either overly bunch students' scores together or spread them apart too much." In fact, experience with SBA English has shown that careful professional development, the processes of in-school teacher moderation, and the support of district coordinators have led to very few concerns of those kinds (Lee 2008). In statistical moderation, the mean score and the spread of SBA scores of students in a given school are compared to a 'moderator variable', which is the same students' results in the formal examination (HKEAA 2006). In the moderation process, the internal rank order of the SBA scores remain the same and no student's marks are changed; but the means and the group overall SBA profile of each school are compared with its own profile on the whole English exam, and with the school's position in the ranking of all schools' English performance. A school's ranking will not be changed but marks may be added or deducted during moderation to bring that school closer to its moderator. In practice this results in a kind of 'norming out' of all schools. The HKEAA statistical moderation booklet is available to all teachers, and a complete explanation of this rather complex process, with diagrammes, can be found at: http://www.hkeaa.edu.hk/DocLibrary/SBA/booklet_sba.pdf.

Christina Lee, then the HKEAA Subject Officer for English, reported that the moderation revealed that not only was the SBA speaking reliable, it had proved to be more reliable than the formal speaking test, and correlated better with all the other aspects of the HKCEE English Language exam (Lee 2008). However, discussions with teachers suggest that even now, few teachers really understand how the process works, and many of them see this external intervention as a sign that their school-based assessments are not trusted.

2.4 Continuing challenges

The literature on educational innovation and assessment reform consistently tells us that reform is difficult, and innovation often does not 'stick'. Two issues in particular have remained problematic throughout the nearly ten years of this initiative.

2.4.1 Planning time

A key issue that arose with the GI was the question of planning: should students be allowed to prepare? Should they have specific planning time in their groups, and if so, how much? This is proving to be a complex question not only for SBA but in language learning as a whole. Until fairly recently, most experts would have said that planning time and overt planning were beneficial (e.g. Skehan & Foster 1997; Yuan & Ellis 2003), but more recently evidence is suggesting that planning is beneficial only in quite specific conditions, and that too much planning time can be counter-productive to success. Ellis (2009) found that planning has a beneficial effect on fluency, but that the effects on complexity and accuracy are less clear. Lam (Chapter 3) found that "what can be observed in the SBA assessed interactions is often not students' *in situ* execution of interactional competence in L2, but a 'canned' product of students' execution of the competence *prior to* the assessed interaction *in L1* during pre-task planning" (p. 56). There is much that is true in this, although in all my school observations, and on all videos, the use of Chinese occurred only occasionally in groups where students were at extremely low levels of English proficiency. However, as I watched the 500+ video clips of GIs while preparing standardising samples and training exemplars for English SBA, I saw many that were over-prepared and formulaic, showing evidence of substantial 'planning time'. I also saw some excellent, close to authentic-seeming, GIs. I saw too, videos that included students doing their own planning before beginning a formal GI: these always seemed more lively, more interactive and on some measures 'better' than the formal GIs after planning. Puntel Xhafaj, Muck, and de Souza Ferraz D'Ely (2011) found, in a close review of this literature, that the factors influencing the effects of planning time on task performance are manifold and very difficult to separate out, but there are good arguments for up to ten minutes planning/preparation time in some contexts. Nitta and Nakatsuhara's (2014) literature review found that while "findings have varied, depending on the nature of planning, task types, and proficiency levels of learners … a general consensus by these researchers is

that relatively long planning times (e.g. ten minutes) in classroom and laboratory settings provide clear benefits to task performance in terms of fluency, but to a lesser extent to complexity and accuracy" (p. 148). The SBA Guidelines for Teachers suggest up to ten minutes planning time before a GI.

What has happened, however, is that the culture of 'test prep' that is ubiquitous in Hong Kong has been brought to bear on this innovation, like every other in Hong Kong. Many Hong Kong teachers have fallen prey to the influence of textbooks 'preparing' students for SBA and to the influence of the cram school culture, which infiltrates their classrooms with their pupils. There is no longer much that is 'unplanned' about the GIs due to the prevalence of training materials and sets of "advice" to students. We were, perhaps, over-optimistic in the early days about the possibility of achieving 'authentic' discourse with the GI, given the reality of the summative role of SBA as part of the Hong Kong formal exam structure.

2.4.2 Fairness

Many issues deserve much more attention than I can give them here; however, the concern about the fairness of school-based assessment when used in a high stakes context such as HKCEE or HKDSE has been pervasive, and still continues, and it must be addressed. As Qian (2008) has pointed out, fairness is as important an issue for assessment for learning and school-based assessment as for traditional tests. This issue was taken very seriously in developing the Hong Kong English Language SBA, and was the focus of one of the inaugural seminars held by the consultancy team (Hamp-Lyons 2006). Qian (*op. cit.*) was very doubtful about SBA at its introduction, commenting that "fairness cannot be reasonably predicted from the SBA, at least during the initial years of its implementation"; and that "it will be technically impractical to fairly and accurately equate all assessment results from different teachers and different schools, who may apply the same set of criteria according to their own understanding, own value systems and their individualized contexts" (p. 105). These concerns were firmly in the minds of the consultancy group while planning the introduction of SBA: indeed, it was not unreasonable that traditional language testers such as Qian should have had these concerns at the early stages. Fortunately, this concern had been addressed by the authors of the ROPES review, which was approved by the Hong Kong Legislative Council. This gave the HKEAA the credibility at senior levels to support us as we went through the multiple stages of the processes of collaboration, sharing and mentoring

described above. This has given the majority of teachers – those who take the SBA task as seriously as the rest of their teaching – a common understanding of the standards expected at each level of the assessment domains. With the confidence of clear and shared standards, many teachers have been convinced that fairness *can* be achieved by alternative processes that are trustworthy.

In its 2013 Handbook for students, the HKEAA describes its approach to fairness:

> How will fairness be ensured in SBA?
> The HKEAA will:
>
> - provide detailed guidelines, assessment criteria and exemplars to ensure consistency in teachers' assessments;
> - provide professional development training to help teachers become familiar with how to conduct the SBA of their subject(s);
> - appoint district coordinators to support schools in the conduct of SBA for individual subjects;
> - moderate SBA marks submitted by different schools to iron out possible differences among schools in marking standards. [http://www.hkeaa.edu.hk/DocLibrary/Media/Leaflets/SBA_pamphlet_E_web.pdf]

The last of these has been discussed above. The first three were developed and built into the support structure by our consultancy team for the SBA English Language, and over time have gradually been implemented in SBA for other subjects, with varying degrees of enthusiasm. The introduction of district coordinators grew directly out of our experience with our core group of teachers in the development year (2005); many of these enthusiastic early adopters (Hall & Hord 2006) proved instrumental in initiating change in their own schools, and went on to become district coordinators, and we did not want to lose their expertise and commitment.

Behind all these structures is the concern for fairness: but the understanding of fairness that we have tried to convey within SBA English Language is a little different from that usually understood in the Hong Kong exam culture. In Hong Kong, fairness is synonymous with reliability. We promote a view of 'fairness' as ensuring that every student has the opportunity to develop their knowledge and ability to the best of their capabilities. This means that both learning and assessment situations

should be structured to make that possible. The SBA professional development materials emphasize that, like teaching tasks, assessment tasks should be adjusted to fit the level of the individual learner: in the same class some students may read an easier text, or be asked a linguistically-simpler question, than others; but each learner should have the opportunity to show what they can do, and be stretched to do a little more. In that way, the class teacher can use her knowledge of her own students to make sure that the scores each student receives on assessable tasks are valid representations of their language at that stage of learning. Since every student is assessed on the same scales and criteria, they can be scored according to their actual performance. This approach has led to far fewer 'no shows' for the speaking exam as students gain more confidence in speaking; and there are far fewer students who receive zero scores because they simply cannot say anything in English.

Writing in 2014, Qian had moderated his view of SBA somewhat; reporting a small-scale study of 'front-line' teachers who were users of SBA, Qian found that 73% of his 33 respondents strongly supported or supported the implementation of SBA while only 12% expressed negative views. This is a much better result than in the early years of the longitudinal study of the implementation of English Language SBA. Qian concludes: "as the English SBA is an assessment-for-learning component within a traditional assessment-of-learning examination of English Language, this new component should be viewed with a different set of expectations from what is expected of a traditional examination, in terms of validity, reliability and fairness" (Qian 2014: 18). This gradual shift in attitude toward SBA and other forms of assessment for learning from an influential early opponent suggests that SBA is succeeding, if slowly. Indeed, the greatest indicator of the success of the innovation is Qian's comment that: "the English SBA is here to stay as an oral assessment component within the new HKDSE Examination of the English Language. Therefore, it is highly advisable to raise the awareness of teachers, students as well as school administrators with respect to the important role the English SBA can play in students' learning of English Language" (p. 19).

2.5 Conclusion

From the distance of some ten years since the beginning of this assessment reform project, it seems to me that two elements have been key to the degree of success this assessment innovation has achieved: a rigorous professional development programme that continued over six years,

and was accessible to every teacher across Hong Kong teaching students at this level; and a carefully developed and validated set of assessment criteria and standards, which were available for use in all schools. These two elements were possible because of the coming together of Chris Davison, a very experienced language teacher educator with a real interest in language assessment, and myself, very experienced in developing and implementing performance assessments and with a real interest in teacher education. I would like to think that we have moved some way towards assessment reform: but if we are to change language assessment practice we must impact teachers', and teacher educators', core beliefs and understandings about testing/assessment.

Acknowledgements

The SBA Consultancy Team were based in the Faculty of Education, the University of Hong Kong, from 2005–2012. The projects were led by Dr. (later Prof.) Chris Davison and Prof Liz Hamp-Lyons; the Project Manager was Ms Wendy Leung, the Project Coordinator Ms Karri Lam, and the IT Officer Miss Jo Wong. Many research assistants and doctoral students helped with the project at various times, and I thank in particular Gao Manman and Xie Qin for stimulating discussions. Thanks must go to the core group of almost 50 English teachers who worked with us through several years of development, piloting, feedback and implementation, and many of whom continued on to become Area Coordinators. They deserve the most appreciation of all, because without their enthusiasm, dedication and skills this innovation could not have been successful. Earlier versions of this chapter have been presented at the IAEA Conference Brisbane, 2009; the ITC Conference, Hong Kong, 2010; and the Language Testing Forum, Lancaster UK, 2011. In reflecting on and drawing together my experiences working in this important project over so many years, I acknowledge the support of the HKEAA and the Hong Kong Research Grants Council (RGC HKU 7268/04H) for funding the several rounds of this research, and the extensive assistance with the study of the SBA Consultancy Team. I am particularly appreciative of the stimulating and challenging working relationship I shared with my co-researcher and co-investigator, Chris Davison. Thanks and appreciation also go to my colleague at the University of Hong Kong, Prof. Stephen Andrews, with whom I worked on RGC Grant HKU7483/06H on assessment innovation, and the doctoral student participating in that project, Ms Yu Ying.

References

Andrews, S. & Fullilove, J. (1994). Assessing spoken English in public examinations – why and how? In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong* (pp. 57–86). Hong Kong: Chinese University Press.

Berry, R. (2008). *Assessment for learning*. Hong Kong: Hong Kong University Press.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5 (1), 7–68.

Boyle, J. & Falvey, P. (Eds.). (1994). *English language testing in Hong Kong*. Hong Kong: Chinese University Press.

Bray, M. (2007). *The shadow education system: Private tutoring and its implications for planners*. (Fundamentals of Educational Planning 61, 2nd Ed.). Paris: UNESCO International Institute for Educational Planning.

Bray, M. (2012). Wolves lurking in the shadows of education. *South China Morning Post*, 19 July 2012.

Carless, D. (2011). *From testing to productive student learning*. New York: Routledge.

Chan, C. & Bray, M. (2014). Marketized private tutoring as a supplement to regular schooling: Liberal Studies and the shadow sector in Hong Kong secondary education. *Journal of Curriculum Studies*, 46 (3), 361–388.

Chen, Q. & Klenowski, V. (2009). Assessment and curriculum reform in China: the College English test and tertiary English as a foreign language education. In: *Proceedings of the 2008 AARE International Education Conference*, 30 November–4 December 2008, Queensland University of Technology, Brisbane.

Cheung, D. (2001). School-based assessment in public examination: Identifying the concerns of teachers. *Educational Journal*, 29 (2), 105–123.

Choi, C-c. & C. Lee. (2009). Developments of English Language assessment in public examinations in Hong Kong. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 60–76). New York: Routledge.

Curriculum Development Council (1999). *English language teaching syllabus and the Senior Secondary curriculum*. Hong Kong: Hong Kong Education Bureau.

Davison, C. (2007). Views from the chalk face: English language school-based assessment in Hong Kong. *Language Assessment Quarterly*, 4 (1), 37–68.

Davison, C. (2013). Innovation in assessment: Common misconceptions and problems. In K. Hyland & L. L. C. Wong (Eds.*), Innovation and change in English language education* (pp. 263–275). Abingdon: Routledge.

Davison, C. & Hamp-Lyons, L. (2009). The Hong Kong certificate of education: school-based assessment reform in Hong Kong English language education. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 248–262). New York: Routledge.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, 30 (4), 474–509.

Elman, B. (2000). *A cultural history of civil examinations in late imperial China*. Los Angeles: University of California Press.

Evans, S. (2000). Hong Kong's new English language policy in education. *World Englishes*, 19 (2), 184–204.

Fok, P. K., Kennedy, K., Chan, K. S. J., & Yu, W. M. (2006). Integrating assessment of learning and assessment for learning in Hong Kong public examinations:

Rationales and realities of introducing school-based assessment. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore, 21–26 May 2006. Accessed on 14 January 2013. http://www.iaea.info/documents/paper_1162a1b7ea.pdf

Gan, Z-d. Davison, C. & Hamp-Lyons, L. (2009). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics*, 30 (3), 315–334.

Hall, G. E. & Hord, S. M. (2006). *Implementing change: Patterns, principles, and potholes*. Boston, ME: Allyn and Bacon.

Hamp-Lyons, L. (2006). Fairness as an issue in school-based assessment. Inaugural seminar series on English language school-based assessment: Integrating theory and practice, 9 January 2006. Hong Kong University.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching: Ideologies and alternatives. In J. Cummins & C. Davison (Eds), *The international handbook of English language teaching, Vol. 1*. (pp. 487–504). Norwell, MA: Springer.

Henrichsen, L. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956–1968*. Westport: Greenwood Press.

Hong Kong Examinations and Assessment Authority. (2006). *2007 HKCEE English Language, Introduction to the School-based Assessment Component.*

IBM. (2003). *Strategic review of Hong Kong examinations and assessment authority: Final report.* Examinations Authority: Hong Kong. Accessed 10 September 2007 from http://www.hkeaa.edu.hk/doc/isd/Strategic_Review.pdf

Kellaghan, T. & Greaney, V. (1992). *Using examinations to improve education: A study in fourteen African countries*. Washington, DC: World Bank.

Kennedy, K. (2013). High stakes School Based Assessment and cultural values: Beyond issues of validity. Key Note Address, Seminar on 'School based assessment: Prospects and realities in Asian contexts', 3 June 2013, Kuala Lumpur, Malaysia. Accessed 11 November 2014 from http://www.cambridgeassessment.org.uk/Images/139719-sba-seminar-papers.pdf

King, R. (1994). Historical survey of English language testing in Hong Kong. In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong* (pp. 3–30), Hong Kong: Hong Kong Chinese University Press.

Lee, C. (2008). The beneficial washback of the introduction of a School-based Assessment component on the Speaking performance of students. Paper presented at the 34th IAEA Conference, Cambridge, September 2008. Accessed 13 August 2014 from http://iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/152128_Lee.pdf

Lee, I. (2005). Latest initiative adds to burden. *South China Morning Press*, 7 May 2005. Accessed 1 April 2008 from http://www.scmp.com/article/499634/latest-initiative-adds-burden

Nitta, R. & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-taskplanning effects on paired oral test performance. *Language Testing*, 31 (2), 147–175.

Puntel Xhafaj, D. C., Muck, K. E. & de Souza Ferraz D'Ely, R. C. (2011). The impact of individual and peer planning on the oral performance of advanced learners of English as a foreign language. *Linguagem & Ensino, Pelotas*, 14 (1), 39–65 (January/June 2011).

Qian, D. D. (2008). English language testing in Hong Kong: A survey of practices, developments and issues. *Language Testing*, 25 (1), 85–110.

Qian, D. D. (2014). School-based English language assessment as a high-stakes examination component in Hong Kong: Insights of frontline assessors. *Assessment in Education: Principles, Policy & Practice*, 21 (3), 251–270.

Skehan, P. & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1 (3), 185–211.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory* (Studies in Language Testing Series, Volume 22). Cambridge: Cambridge ESOL and Cambridge University Press.

Yuan, F. & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologicoral production. *Applied Linguistics*, 24 (1), 1–27.

Yung, B. H. W. (2002). Same assessment, different practice: Professional consciousness as a determinant of teachers' practice in a school-based assessment scheme. *Assessment in Education: Principles, Policy & Practice*, 9 (1), 97–117.

3

Contriving Authentic Interaction: Task Implementation and Engagement in School-Based Speaking Assessment in Hong Kong

Daniel M.K. Lam

3.1 Introduction

In 2007, a School-based Assessment (SBA) component combining the assessment of speaking with an extensive reading/viewing program was introduced into the Hong Kong Certificate of Education Examination (HKCEE). Having operated on a trial basis for several years, SBA is now fully integrated in the new secondary school exit examination, the Hong Kong Diploma of Education Examination (HKDSE), since 2012.

The SBA component accounts for 15% of the total subject mark for HKDSE English Language, consisting of two parts. Part A is made up of two assessments, one *individual presentation* and one *group interaction* (otherwise commonly known as the 'group discussion' task), with one to be carried out in Secondary 5 (S5) and the other in Secondary 6 (S6). The speaking tasks are based on an extensive reading/viewing program. Therefore, students engage in either an individual presentation or a group discussion on the books they have read or movies they have viewed. Part B consists of one assessment in either the group interaction or individual presentation format, based on the Elective Modules (e.g. social issues, workplace communication) taught in the upper secondary curriculum. This is to be carried out either in the second term of S5 or anytime during S6. Thus, a total of three marks (each weighing 5%) are to be submitted by the teacher. Further details of the SBA assessment tasks can be found in the Teachers' Handbook (HKEAA, 2009) available online.

This study focuses on the *Group Interaction* task, whereby students in groups of three to five (mostly four) carry out a discussion of around

eight minutes. While the peer group interaction format has been used in the public exam for many years, the SBA task differs from its public exam counterpart in that students would be interacting with their classmates rather than unacquainted candidates, and are assessed by their own English teacher instead of unfamiliar external examiners. Moreover, one of the discussion tasks would be based on a book or movie that students have experienced as part of the extensive reading/viewing program.

The objectives of the SBA initiative are to elicit and assess 'natural and authentic spoken language' (HKEAA, 2009, p. 7), providing an assessment context 'more closely approximating real-life and low-stress conditions' (p. 3), and for students to 'interact in English on real material' (Gan, Davison, & Hamp-Lyons, 2008). Thus, the assumption is that *authentic oral language use* constitutes the basis of the validity of the assessment task, as has been reiterated in the published guidelines (HKEAA, 2009) and in validation studies (Gan *et al.*, 2008; Gan, 2010).

As an assessment-*for*-learning initiative, the assessment policy for SBA places considerable emphasis on flexibility and sensitivity to students' needs in the design and implementation of the assessment tasks, a marked departure from the public exam where standardized tasks, conditions, and practices are strictly adhered to for reliability and fairness. As stated in the Teachers' Handbook,

> the SBA process, to be effective, has to be highly contextualised, dialogic and sensitive to student needs (i.e. the SBA component is *not* and cannot be treated as identical to an external exam in which texts, tasks and task conditions are totally standardised and all contextual variables controlled; to attempt to do so would be to negate the very rationale for SBA, hence schools and teachers must be granted a certain degree of trust and autonomy in the design, implementation and specific timing of the assessment tasks). (HKEAA, 2009, p. 4)

The recommended practice is for teachers to give students the 'general assessment task' to prepare a few days in advance, and to release the 'exact assessment task' shortly before the assessment to avoid students memorizing and rehearsing the interaction (*ibid.*, p. 37).

Although some recommendations for task implementation are included in the Teachers' Handbook and in teacher training seminars, the emphasis on flexibility in the assessment policy has translated into diverse assessment practices (see Fok, 2012). There is considerable variation in when the discussion task with question prompts is released to

students, in other words, in the length of preparation or pre-task planning time during which students have the opportunity to talk to group members about the upcoming assessed interaction (Note: the term *preparation time* is used in official documents published by HKEAA, whereas *pre-task planning time* is used extensively in the SLA and language testing literature. The two terms are used synonymously in this chapter). Varied practices in task implementation are evident, both in previous studies and my own. Gan *et al*. (2008) and Gan (2012) reported that the specific assessment task was made known to students about ten minutes beforehand. In the school that Luk (2010) investigated, students received the discussion prompt one day before the assessment, which was also when they were told who their group members are. Of the eight schools whose teachers Fok (2012) interviewed, four gave students the actual discussion questions one day or more before the assessment, three gave students similar sample questions a few days before but the actual questions only minutes before the assessment, and one allowed no preparation at home but gave students the actual questions shortly prior to the assessed interaction. As for the two schools in my own study, one (School L) released the discussion prompt to students ten minutes before the assessment, and group members were not allowed to talk to each other during preparation time. The other school (School P) released the discussion prompt to students a few hours before the assessment, and students who formed their own group could plan their interaction together.

Such variation in the pre-task planning time allowed generates group interactions that are considerably different in nature. As will be seen, students having a few hours or more to prepare display an overwhelming tendency to pre-script an interactive dialogue followed by reciting and acting out the scripted dialogue, rather than participating in a spontaneous interaction as students having only 10–15 minutes of planning time do. This chapter explores what students do during the preparation time and how it affects their subsequent group interaction; and examines whether the task, as it is implemented, elicits authentic oral language use. Before outlining the details of data and methodology, I shall review some previous research relevant to this study.

3.2 Literature review

Since its implementation, there has been a growing body of research that examines different facets of SBA. One strand of research looked at perceptions towards the SBA initiative by various stake-holders, for example, teachers' and students' initial responses at the first stage of

implementation (Davison, 2007); students' and parents' views (Cheng, Andrews, & Yu, 2011); and teachers' perceptions and readiness of administering SBA at the frontline (Fok, 2012). Another strand of research focused on the assessed speaking performance. Some studies engaged in micro-analysis of the test discourse and students' interaction (Gan *et al.*, 2008; Gan, 2010; Luk, 2010), to be reviewed in more detail below. Others compared the discourse output elicited by the two task types (Gan, 2012), and examined the extent to which students' personality (extroversion/introversion) influences their discourse and test scores (Gan, 2011). At a more theoretical level, Hamp-Lyons (2009) outlined a framework of principles guiding the design and implementation of large-scale classroom-based language assessment, drawing on the case of SBA in Hong Kong.

3.2.1 Validity of SBA group interaction

Validation studies of the SBA Group Interaction task to date have yielded mixed results regarding whether the task has achieved its aim of eliciting students' authentic oral language use. Gan *et al.* (2008) presented a detailed conversation analysis of one group interaction from a databank of 500, focusing on topic organization and development. They identified two types of topic shifts: 'marked topic shifts', where the speaker used particular turn design features to signal the introduction of a new topic, and 'stepwise topic shifts', where the speaker referred to the content in the previous turn and introduced new elements as something relevant. The authors concluded that the similarities in topic negotiation and development to everyday conversation serve as evidence for authenticity, hence validity, of the task.

In another study, Gan (2010) compared the students' discourse in a higher-scoring group and a lower-scoring group from the same databank of 500. He found that, in the higher-scoring group, participants responded contingently to each other's contributions. By fitting their comments closely to the previous speakers' talk, these participants displayed understanding of the preceding discourse. Participants in the lower-scoring group, by contrast, often reacted minimally. Their discourse was more 'structured' and reliant on the question prompts, but there was also some negotiation of form and meaning, where students helped one another search for the right forms to express meaning. In alignment with Gan *et al.* (2008), he concluded that the discourse exhibited characteristics of an authentic task that 'emphasize[s] genuine communication and real-world connection' and 'authentically reflects candidates' interactional skills' (Gan, 2010, p. 599).

The study by Luk (2010) painted a considerably different picture. She found the group interactions characterized by features of ritualized and institutionalized talk rather than those of everyday conversation. In her discourse analysis of 11 group interactions involving 43 female students in a secondary school, participants were seen to engage in orderly turn-taking practices with turns passed on in an (anti-)clockwise direction, and to front those speaking turns in which each member delivered extended, pre-planned speech before the whole group started giving responses. There was little evidence of on-line interaction and contingent responses to previous speaker contribution, manifested in the frequently deployed surface agreement (e.g. 'I agree with you') that came without further elaboration, therefore appearing superficial and possibly perfunctory. Students also avoided seeking clarifications from each other, but concealed problems instead. These findings mirrored those of He & Dai (2006) on the group discussion task in the College English Test in China, where candidates were observed to exploit the time when others were speaking to organize and formulate their own ideas in upcoming turns, and accordingly, to focus on expressing their own ideas rather than responding actively and relevantly to previous speakers' talk. With students' interview responses as supplementary evidence, Luk (2010) concluded that students were engaging in the endeavor of managing an 'impression of being effective interlocutors for scoring purposes' rather than in 'authentic communication' (p. 25).

As shown above, the findings and conclusions about the validity of the SBA Group Interaction task in terms of the authenticity of students' discourse elicited are mixed. It is not difficult to note a marked difference in the amount of preparation time between the first two studies and Luk's (2010) study, although none of them investigated in detail what students do during the planning time, or attributed the observable interactional patterns to students' pre-task planning activities. However, as will become evident in Spence-Brown's (2001) study (reviewed below) and my own, there are cases where the candidates' discourse ostensibly suggests authentic language use, but close inspection of their task engagement during the planning stage yields contrasting evidence.

3.2.2 Effect of pre-task planning time on task performance

On the question of whether pre-task planning time benefits subsequent task performance, studies in testing and non-testing contexts to date have also produced different results. As reviewed in Nitta & Nakatsuhara (2014), previous research on TBLT (task-based language teaching) has found planning time beneficial from a cognitive perspective, having a

positive effect on subsequent task performance most notably in fluency, and to a lesser extent in terms of accuracy and complexity (see Ellis, 2009, for an overview of these studies). However, as pointed out by Nitta & Nakatsuhara, these studies focused primarily on the cognitive complexity and linguistic demands of the task, and did not investigate the interactional aspects of the task performance.

According to Wigglesworth & Elder (2010), evidence that pre-task planning time benefits subsequent task performance in language testing contexts is less clear. While a few studies attested to a positive impact on accuracy (Wigglesworth, 1997), complexity (Xi, 2005), or both, along with 'breakdown' fluency (Tavakolian & Skehan, 2005), others found little or no benefits on test scores or the discourse output (Wigglesworth, 2000; Iwashita, McNamara, & Elder, 2001; Wigglesworth & Elder, 2010). Again, the overwhelming majority of the studies have focused on proficiency measures – accuracy, fluency, and complexity – of the discourse output. This can be readily accounted for by the fact that testing studies on the effect of pre-task planning time to date have been exclusively on monologic rather than interactive tasks (Nitta & Nakatsuhara, 2014).

Nitta & Nakatsuhara's (2014) pioneering study of the impact of planning time on performance in a paired speaking test revealed a potentially detrimental effect on the quality of interaction. Analysis of the candidates' discourse showed that the interactions without the three-minute planning time were characterized by collaborative dialogues, where candidates engaged with each other's ideas and incorporated their partner's ideas into their own speeches. In contrast, the planned interactions consisted of more extended monologic turns where candidates only superficially responded to their partner's talk and concentrated on delivering what they prepared. The significance of the study is that, while the planning time was found to be slightly beneficial to candidates' test scores, the qualitative analysis of interactional patterns indicated that planning time might inhibit the task from tapping into the construct that the task is meant to measure: the ability to interact collaboratively.

Evident from the above review is that, in both SLA and testing research, the focus of pre-task planning effects has mostly been on proficiency measures in the discourse output; and in testing studies, there is a gap in looking at pre-task planning effects on candidates' performance in interactive (paired or group) task formats. Further, there seems to be a general lack of studies that investigate what candidates actually do during the pre-task planning time (Wigglesworth & Elder, 2010), let alone drawing links between the planning activities and the

extent of candidates' authentic engagement in the subsequent dialogic task. This is perhaps because in most high-stakes assessment contexts, candidates are not given extended preparation time or the opportunity to talk to fellow candidates in the same pair/group before the assessment. Therefore, the classroom-based assessment situated within a high-stakes examination in the present study, with the assessment task implemented in such conditions that follow from a flexible assessment policy and engender particular kinds of pre-task planning activities and strategies, creates a unique, interesting context for the study.

3.2.3 Call for research on task implementation

Given the mixed results on the authenticity of the SBA Group Interaction task in previous studies, and the possible detrimental effect of pre-task planning time identified by Nitta & Nakatsuhara (2014), the importance of investigating how the assessment task is implemented and engaged in by student-candidates is becoming apparent. In the language testing literature, several authors have called for studies on task implementation. In concluding her study on the effect of planning time on subsequent speaking performance, Wigglesworth (1997) recommended looking into what candidates actually do during pre-task planning time in future studies. Building on earlier arguments by Messick (1994), McNamara (1997) asserts that validity cannot be achieved through test design alone, but needs to be established with empirical evidence from actual test performance 'under operational conditions' (p. 456). Applying this to the case of SBA Group Interaction, validation studies need to include an examination of students' activities during the preparation time, which is a non-assessed yet integral part of the assessment task. How important it is for test validation studies to look at task implementation and authenticity of engagement is most elaborated and empirically attested to in Spence-Brown (2001).

The assessment task that Spence-Brown (2001) examined involved students in a Japanese course at an Australian university conducting tape-recorded interviews with a Japanese native speaker whom they had not previously met. Data comprised students' discourse in the interview, scores and raters' comments, and retrospective interviews with students incorporating stimulated recall. The analysis identified several aspects of students' task engagement that posed threats to the authenticity and validity of the task. Besides selecting a known informant and pretending otherwise, as well as rehearsing and re-taping the interview, students approached the task by preparing questions, predicting answers and appropriate responses to them. This enabled students

to appear to be engaging in authentic interaction without actually taking the risk of doing so. In a particularly noteworthy case, a student predicted the informant's answer to a question and pre-planned his response to the answer. The surface discourse in the interview suggested successful interaction, with the student giving an appropriate response. However, the stimulated recall revealed that the student did not actually understand the informant's answer, but drew on a rehearsed response that suggested he did. Based on such findings, Spence-Brown (2001) challenged the validity of the task: while the task is designed to engage students' use of 'on-line' linguistic competence, it in fact does not. She cautioned that because the nature of task engagement is not always transparent in the task performance (the taped interview in this case), it is more meaningful to examine authenticity from the view of implementation rather than task design alone.

3.2.4 The present study

Informed by the findings and recommendations from the previous research outlined above, the present study sets out to examine the validity of the SBA Group Interaction task by looking at aspects of task implementation and student-candidates' engagement. Specifically, it seeks to answer the following research questions:

1. Does the SBA Group Interaction task elicit authentic oral language use from students in accordance with the task's stated aim?
2. What do students do during the pre-task planning time, and how does this affect their discourse in the group interaction?

3.3 Data and methodology

The data reported in this chapter comes from a larger study, in which three types of data were collected: (1) video-recordings of test discourse, (2) stimulated recall with student-candidates and teacher-raters, and (3) mock assessments. This section provides details of the data collected for the entire research project and the data selected for in-depth case study in this chapter.

First, video-recordings of the group interaction task completed by 42 groups in two secondary schools (School P and School L) were obtained. Among them, 23 were from Part A of the SBA, and 19 were from Part B, with some of the Part B group interactions conducted by the same students as Part A in either the same or different grouping. To explore how extended preparation time as a task implementation condition

might impact on the subsequent assessed interaction, this chapter focuses on the case of School P, where students were given a few hours of preparation time (cf. ten minutes in School L). In the following section, two extracts from two different group interactions in School P will be presented. They were selected on the basis that, at first glance, the students appeared to be engaging in authentic interaction, while close analysis and additional data (explained below) revealed the contrived nature of their interactional exchange. The first extract (Extract 3.1) was part of a group interaction for Part A in which students were asked to talk about the misunderstanding between the two main characters in the movie *Freaky Friday*. In the second extract (Extract 3.2), students in a group interaction for Part B assumed the roles of marketing team members, and the task was to choose a product to promote and discuss the promotional strategies. The interactions were transcribed in detail following Jefferson's (2004) conventions (see Appendix 3.1 for additional transcription symbols used), and analyzed following a conversation analytic approach.

To supplement the test discourse data, retrospective interviews incorporating stimulated recall were conducted for 15 assessed interactions (eight from Part A, seven from Part B) with the relevant student-candidates and teacher-raters in the two schools who were available at the time of data collection. Depending on the mutual availability of the participants and the researcher, the time gap between the assessment and the interview varied between a few days and two months. During the interviews, the video-recordings of the assessed interactions were played and paused at intervals for the students/teachers to comment on. Additional questions about particular parts of the interactions (e.g. episodes which appear to be authentic interactional exchange) and the participants' views about the assessment in general were also asked. The stimulated recall procedure enabled me, as the researcher, to gain insights on the kinds of pre-task planning activities student-candidates engaged in, and how the interactional exchanges were perceived by the teacher-raters. All interviews were conducted in Cantonese, and the interview transcripts were translated into English. The only exceptions were two interviews (for Part A and Part B respectively) with one teacher-rater, conducted in English in accordance with her preference. The following section presents the relevant stimulated recall data for the group interaction extracts analyzed.

The third type of data was from a mock assessment, where the whole assessment process from preparation time to the assessed interaction, as well as the post-interview immediately after the assessment, was video-recorded. This was to capture the fine-grained details of students' pre-task planning activities and allow closer inspection of such activities in subsequent analysis. The limitations were that, due to constraints

on the participants' availability, it was possible to carry out the mock assessment with only two groups, and with reduced preparation time. These two groups of students (four in each group) were selected from the 19 group interactions for Part B, where ostensibly authentic episodes of talk exchange were found in the initial analysis of their test discourse. The two groups were each given a discussion task adapted from their Part B assessment. One group was given approximately one hour of preparation time, and the other group approximately ten minutes as part of an investigation of whether and how the amount of preparation time impacts on the subsequent group interaction. In the post-interview, students were asked to compare their experience in the mock and the actual assessments, in particular what preparation work they did for the actual assessment and what they were unable to do before the mock assessment, and these responses were taken as complementary evidence to the video-recording of the preparation time. Extracts 3.4–3.6 in the section below illustrate some of the planning activities engaged in by the student group with approximately one hour of preparation time.

3.4 Data analysis

3.4.1 Discourse in assessed interactions

I begin by presenting a conversation analysis of two extracts from two group interactions, where the discourse ostensibly suggests authentic interaction among the student participants.

Extract 3.1 (PA11: 48–60)

```
 1 W:  Do you remember there is a scene showing that the
 2     door of Anna's- (..) bedroom had been removed by
 3     Mrs Coleman¿ ((R nods and turns her head to N just
 4     before N begins her turn))
 5 N:  Yeah. I can even \\remember the phrase on her room's
 6                      \\((R looks briefly at W))
 7     door. Parental advisory, uh keep out of my room.
 8     So::, what you're trying to say i::s
 9 W:  >What I'm trying to< say is privacy. ((R turns to D))
10 D:  I see what you mean. I think: (.) privacy is::- should
11     be: (.) important to anyone. Uhm just like me, if my
12     right (.) if my right to play computer game is being
13     >exploited by my mom<, I think I will get mad on
14     her.=So, I think: lack of (.) privacy is the main cause.
```

The group has been talking about the various aspects of misunderstanding between the mother, Mrs. Coleman, and the daughter, Anna, in the movie *Freaky Friday*. Extract 3.1 shows a sequence where the group discusses another cause of misunderstanding between the two characters.

In lines 1–2, W asks the co-participants if they recall a particular scene from the movie. This takes the shape of a pre-telling, whereby W checks the requisite condition for a forthcoming telling. The next speaker, N, offers an affirmative 'yes', and provides further recalled details showing the condition has been met (lines 4–6). The sequence does not immediately proceed to W's telling, however. In lines 6–7, N issues a clarification request in the 'fill-in-the-blank' format ('what you're trying to say is'). This displays her orientation to W's prior turn as projecting more talk – the thrust of the telling sequence for which W's recall question has been laying the groundwork. Interestingly, on the one hand, N's clarification request displays her alignment with the trajectory of a telling W has been setting up, amounting to a 'go-ahead' for W to make her point. On the other hand, N modifies this trajectory by opening up another sequence, of which the clarification request is the first-pair-part (FPP).

Note how W's following response (line 8) displays sensitivity to the contingency of the unfolding sequence. Instead of staying on her own course and designing her turn like the FPP of the main telling sequence following the pre-telling, W aligns with the new trajectory of talk set up by N through formatting her turn as the answer second-pair-part (SPP) to N's question, with the preface 'what I'm trying to say is' mirroring the shape of the question FPP. Throughout these three turns (lines 1–8), then, both participants construct their responses in ways that are sensitive to and contingent on the previous speaker's talk. In other words, they seem to engage in each other's talk and develop on each other's contribution, showing evidence of authentic interaction.

Rather strikingly, however, the main telling towards which all the previous interactional work seems to have been building ends up with one word, 'privacy' (line 8). Here, D acknowledges receipt and claims understanding of W's telling, provides an affiliative assessment of the point about privacy, offers an example from his personal experience, and finally formulates the upshot of the whole sequence ('lack of privacy is the main cause'). Interestingly, then, W is seen to leave it for D to spell out the thrust of the sequence.

Thus, we see a rather odd sequential development in which W seems to (willingly) relinquish the rights to making her point, after all the

preliminary interactional work that has built towards it and would have sequentially ratified an extended telling turn on W's part for such purpose. The task of bringing home the point about privacy as a main cause of misunderstanding is conveniently re-allocated to another participant, D. This raises questions as to whether this has truly been how the interaction has unfolded, or something pre-planned prior to the assessment.

Indeed, close examination of co-participants' non-verbal behavior yields preliminary evidence that this interactive sequence has been pre-scripted. In lines 2–3, towards the end of W's question, R nods and turns her head to N just before N commences her turn. Meanwhile, despite generally being the most active participant, R does not even offer a minimal verbal response such as 'mm' or 'yes' here, let alone elect herself to answer W's question. As N begins answering W's question, R glances at W again (line 5) instead of focusing her gaze on N to display listenership. Finally, in line 8, R turns to D right at the end of W's turn and just before D's, as if she has already known that D would be the next speaker.

Students confirmed in the stimulated recall that this sequence (and the whole interaction) was pre-scripted, and R explained that this was to create an opportunity for a group member who wouldn't have spoken for a while to take a turn.

Extract 3.2 below shows another group interaction, one that simulates a marketing team meeting for the promotion of a new product. The discourse in this episode, with reference to turn design and sequential development, gives some indication of students' authentic engagement in the simulated interactional context, and in challenging each other's ideas.

Extract 3.2 (PB14: 10–25)

```
 1 L:  Mm. Yes, our company has just released (.) our
 2     beauty products in- eh- uhm the teenagers. Mm::
 3     (.) mm:: (1.9) uhm: so: are you guys clear about
 4     the special features of the product?
 5 K:  °Mm.° I've heard that the new products .h are composed
 6     of a traditional Chinese medicine. That is quite
 7     special.
 8     (..)
 9 T:  Uhm:: but, do you think that the traditional Chinese
10     medicine .h have strong and strange smell? Many people
11     may refuse to use our ↑pro↓duct.
```

```
12 S:  Hey. You've missed out a ↑po↓int. That is our product
13     also includes (.) natural ingredients (.) li:ke lav
14     ender (.) which is successfully cover (.) the::
15     ↑smell brought by the traditional Chinese medicine.
16 L:  Mm::. (.) It's one of the fo- ma- m- main focus, that
17     uh to promote our product. .h Uhm, it is not smelly
18     even if we have added the traditional Chinese medicine
19     into it.
```

The sequence begins with L, who assumes the role of team leader, initiating the topic about special features of their skincare product (lines 1–3). She discursively constructs her authoritative role through announcing the release of their product, and asking other team members if they are 'clear about the special features', thereby claiming epistemic superiority over other group members. K responds by introducing the feature of traditional Chinese medicine as product ingredient, and adds a positive assessment (lines 4–5). In providing this answer to L's question, she ratifies and co-constructs L's role as team leader. The turn design of prefacing her response introducing the Chinese medicine with 'I've heard that …' also displays K's commitment to their contextual roles as marketing team members (as people who should know about the product's features but did not create the product themselves).

K's positive evaluation of Chinese medicine as product ingredient is then met with a disagreeing response from T (lines 7–9). This begins with prolonged hesitation 'uhm', followed by a negative assessment of the Chinese medicine framed as a question. Neither K nor T orients to the question as projecting an answer, as T continues to offer a further account for disagreement predicting negative consumer reactions. The turn shape of T's disagreeing response in itself is noteworthy, indeed striking. It differs markedly from formulaic disagreeing responses such as 'I'm sorry I can't agree with you' that feature an explicit disagreeing component, and which frequently occur in other group interactions in the data.

Equally striking, perhaps, is the following response by S, which counters T's disagreement by commenting that T has 'missed out a point' – another feature of their product (line 10). This type of sequential development, where a disagreeing response is followed by another disagreeing response countering the first, is rarely observed in the data. However, S is then able to conveniently introduce this neglected feature both as a counter argument and as a new idea that she contributes on the topic, as she elaborates on how other natural ingredients such as lavender can solve the problem of the smell brought by Chinese medicine. Such

a design enables S to both topicalize previous speaker's idea of Chinese medicine and make her own contribution about other ingredients.

During the stimulated recall, the teacher-rater paused the video and gave her positive evaluation on this episode of talk exchange:

Extract 3.3 (PB-TR-B stimulated recall, English original)
((TR pauses the video after line 9 in Extract 3.2))

TR: Uh I like it how she responded to something that K said. So rather than say something else … she asked about it.

The teacher-rater positively remarked that T raised a question about K's idea in her response, topicalizing the previous speaker's contribution rather than focusing on delivering her own idea. Subsequently, the teacher-rater also gave a favorable evaluation of S's response, in which she further topicalized the feature of Chinese medicine and elaborated on how the problem with its smell could be solved. Throughout the stimulated recall, the teacher-rater commented several times that this group's interaction was 'authentic'.

Nevertheless, the stimulated recall with students again revealed that the entire interaction was pre-scripted and rehearsed. Within the test discourse, students' intonation and the strangely 'neat' speaker transition without many gaps and overlaps might have been a giveaway. More importantly, the students' unique ways of doing disagreement (cf. using formulaic expressions), which ostensibly suggested authentic interaction, was precisely one of the clues to a pre-planned, contrived interaction. Though performed in a playful tone here, the kind of unmitigated negative comment directed at a co-participant (line 10) rarely occurs in spontaneous assessed interactions, as it would probably constitute a direct face threat to a co-participant.

3.4.2 Pre-task planning activities

Further insights about the kinds of pre-task planning activities students engage in, including pre-scripting, were gained through close examination of the video-recorded one-hour preparation time for one of the mock assessments. Figure 3.1 below is a schematic representation of the planning activities carried out during the preparation time.

As shown in Figure 3.1, students' pre-task planning for the mock assessment can be roughly divided into three stages (represented in solid lines). The first stage involves students brainstorming for ideas about the discussion topic, researching information and relevant

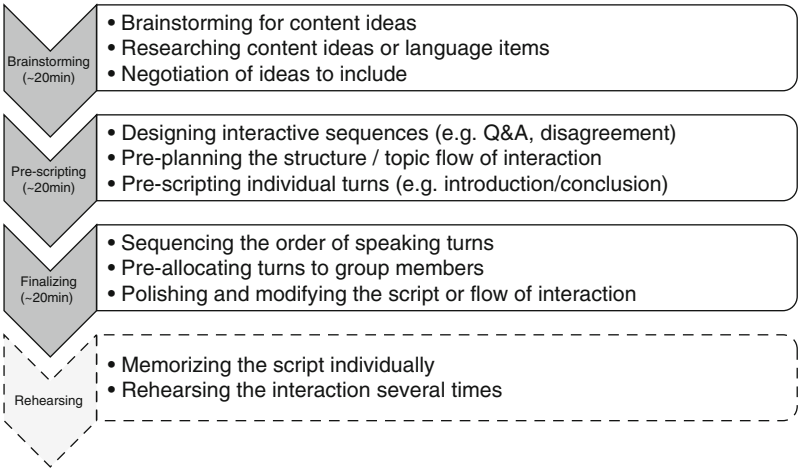


Figure 3.1 Students' pre-task planning activities

vocabulary items with their smartphones, and negotiating what ideas to include and exclude in the assessed interaction. In the second stage, students decide together on the structure or topic flow of the interaction. They also design interactive sequences such as question-and-answer or disagreement, and pre-script particular speaking turns such as the opening and concluding turns. In the final stage, students fix the sequence of speaking turns and assign each turn to a group member. Any final touch-ups to the script or flow of interaction are also done at this time.

It should be noted that these activities are not actually carried out in a strictly linear sequence, and are only presented in approximate order. For instance, form-focused planning activities such as looking up vocabulary items and English translation of brand names, and checking them with others in the group, are recurrent and interspersed throughout the preparation time. In the post-interviews with the two groups participating in the mock assessment, supplementary information about students' pre-task planning activities was gained regarding what they did before the actual assessment and, correspondingly, what they did not manage to do during the preparation time for the mock assessment. Students reported not having enough time for pre-scripting the interaction *verbatim* before the mock assessment. They also reported an additional stage before the actual assessment

(in dotted lines) that involved memorizing the script individually and rehearsing the interaction (referred to as 「試演」'trial acting') several times.

In the following, I discuss three types of pre-task planning activities that pose threats to the authenticity of the assessed interaction.

First, students were observed to pre-negotiate the pros and cons of certain ideas in the brainstorming stage, with differences of opinion dealt with and consensus reached. Consider the following extract of students' pre-task planning discussion (Extract 3.4):

Extract 3.4 (PB11MockPrep 24:00)
((Previously, someone suggested hiring three spokespersons for their three target age groups of customers))

Y: But have you guys considered the cost? It's very expensive, if we get three spokespersons.
K: Well, so maybe we can *ban* the idea of three spokespersons. *Ban* three spokespersons.
R: No. We should first have someone say let's get one spokesperson, then someone else *ban* the idea, and say we actually have three *target* groups, so why don't we have one spokesperson for each *target* group.
S: But it's mainly adults who would buy [vitamin pills] after all. Isn't one spokesperson enough?
Y: Wait. Let's get a 'mum'. Getting a 'mum' [as the spokesperson] will work!
K: We can say it's usually housewives who buy [vitamins for the whole family]. It's not the children who would buy them.

Instead of having it as a point for debate in the assessed interaction, the group pre-determined their final decision of having only one spokesperson, and pre-planned how they would work their way through the different proposals to reach such consensus in the assessed interaction. This pre-task discussion therefore eliminates the information and opinion gaps that could create a genuine need for communication and negotiation in the group interaction task proper.

Related activities which threaten the authenticity of the assessed interaction include students pre-scripting interactive episodes, pre-sequencing their turns and assigning them to individual group members. Extract 3.5 below shows the final stage of pre-scripting the discussion on the 'spokesperson' topic.

Extract 3.5 (PB11MockPrep 55:45)

S: ((points to Y)) She will introduce [the topic of] *spokesperson*
K: OK. So I'll then suggest three. ((writing on note card simultaneously)) I'll say since we have three *target groups*, why don't we get three *spokespersons*.
R: ((points to K)) You say that, you'll suggest that, right? So you suggest having three spokespersons. And then who's gonna *ban* the idea? You *ban* it, S.
S: Sure, I'll *ban* it. I'll *ban* it.
R: And after *ban*ning it I'll lead to [the topic of] '*place*'. Alright, let's do it like this.
S: ((writing simultaneously)) I'll do the banning. The cost is too high.
R: ((writing simultaneously)) 'Three *spokespersons'* is by K, and then S *bans* the idea, because the cost is too high. And then I'll agree with her, and afterwards I'll introduce [the topic of] '*place*'.

As seen in the transcript, the students are assigning roles and finalizing the interactive sequence where they would propose having three spokespersons, challenge the idea, then agree on the alternative of having one only, and shift to another topic. The sequence of assigned speaking turns, and the order of proposing, disagreeing, and finally reaching consensus on an idea, were all written down on their note cards as what the students themselves called the 'route map' (「路圖」) of the assessed interaction.

Finally, there was an instance of a student helping a less capable group member (Y) pre-script her turns:

Extract 3.6 (PB11MockPrep 41:40)

K: Oh so you can also mention this. You say 'let's start with '*product*', but I can't think of promotional ideas because it's difficult when there're so many *competitors*, so what ideas do you guys have?' And then we'll respond to her.

Thus, what Y eventually said in that turn during the assessed interaction was not even entirely her 'original work', let alone a spontaneously produced contribution.

On scrutinizing students' pre-task planning activities, we now have good evidence that what might appear as authentic exchange in the assessed interaction can in fact have been contrived. Overall, the data

in School P indicates an overwhelming tendency of students engaging in contrived rather than spontaneous interaction, supported by the fact that all students in School P interviewed admitted having pre-scripted the assessed interaction. As a result of the aforementioned pre-negotiation of ideas and the subsequent pre-scripting of the relevant discussion, what the students perform and are evaluated on during the assessed interaction is, at best, a re-presentation of their pre-task interaction conducted in L1. It is not an authentic and spontaneous interaction conducted in L2 spoken English, the target of the assessment. Instances of authentic, spontaneously produced exchanges were found in interactions with only ten minutes of preparation time (in School L and in one of the groups in the mock assessment), but are beyond the scope of this chapter. These cases and their comparison with contrived exchanges warrant equally detailed analysis and discussion, and will be taken up in future published work.

3.5 Discussion and conclusion

3.5.1 Findings and implications

This chapter has sought to contribute to the body of validation work for the SBA Group Interaction task, and to reveal some of the complexities in ensuring the task's validity implicated by the 'flexibility' element in the assessment policy and the corresponding practices. A main objective of this study was to examine whether the Group Interaction task, in the way it is implemented, elicits authentic oral language use. Previous studies have gauged the task's (lack of) construct validity mainly in terms of *authenticity* and its real-world connection with everyday conversation. Indeed, the relationship between authenticity and validity of a task has long been an issue in theoretical debates. Bachman (1990) attributed the preoccupation with authenticity to 'a sincere concern to somehow capture or recreate in language tests the essence of language use' in the target domain (p. 300). However, Spolsky (1985) contended that test behavior can never be an entirely authentic reflection of non-testing behavior, as interactions in testing and non-testing situations follow different rules. Some authors (e.g. Widdowson, 1979; van Lier, 1996) distinguish between *genuine* – employing texts used by native speakers for everyday communication in pedagogic tasks; and *authentic* – related to processes of engagement. Building on this distinction, Spence-Brown (2001) introduced the notion of *authenticity of engagement* in evaluating the validity of assessment tasks.

In answer to the research questions of this study – whether the SBA Group Interaction task elicits authentic oral language use, and how it is affected by students' pre-task planning activities – we can conclude that, while the task has authenticity in terms of task content, it has questionable authenticity of engagement by students. The discussion tasks do have some real-world connection, with students interacting on 'real material' (movies), or simulating real-life situations (work meetings). Students' discourse yielded ostensible evidence of authentic engagement in interacting with each other, for instance, modifying one's response to align with previous speaker's talk (Extract 3.1), and natural, non-formulaic ways of doing disagreement (Extract 3.2). Some of these were recognized and favorably evaluated by the teacher-rater. Nonetheless, stimulated recall with the students and video recording of preparation time before the mock assessment revealed that these interactive episodes were part of a staged performance of pre-scripted dialogues.

Therefore, what the assessed interactions showed was essentially the product of students acting out a composed dialogue based on their knowledge and perceptions of what interactional competence is, rather than students' spontaneous performance of the competence that involves moment-by-moment monitoring of and contingent reaction to each other's talk in real time. Several authors have included this element of 'spontaneity' in defining competence in interaction. Bachman (1990) describes 'communicative language ability' as 'consisting of both knowledge, or competence, and the capacity for implementing it, or executing that competence in appropriate, contextualized communicative language use' (p. 84). Barraja-Rohen (2011) asserts that interactional competence involves, among other skills, 'precision timing and a quick analysis of speakers' turns' (p. 482). Spence-Brown (2001) questions the validity of the tape interview task based on its failure in eliciting learners' 'on-line linguistic competence' (p. 471). Similarly, what can be observed in the SBA assessed interactions is often not students' *in situ* execution of interactional competence in L2, but a 'canned' product of students' execution of the competence *prior to* the assessed interaction *in L1* during pre-task planning. Furthermore, Kramsch (1986), in her seminal work on interactional competence, describes interaction as relative and unpredictable in nature, and it is on this premise that talk exchange takes place, with the objective of reducing uncertainty of 'intentions, perceptions, and expectations' (p. 367). However, we have seen evidence of pre-task planning activities closing the information or opinion gap for interaction, with aspects of uncertainty and

unpredictability (otherwise matters to deal with in the assessed interaction) being reduced or eliminated.

Some of the key emphases of the School-based Assessment policy, as outlined in the Introduction, were on flexibility, sensitivity to students' needs, and low-stress conditions, all constitutive of an explicit departure from standardized language assessments. In a way, the face of the assessment practices matched the policy. First, as seen in previous studies reviewed and my own, diverse practices in task implementation, rather than standardized tasks and task conditions, were found across different schools. Moreover, extended preparation time given in some schools catered for weaker students' needs, as it could reduce anxiety in the otherwise highly stressful assessment situation (Wigglesworth & Elder, 2010), as well as enable prepared speech for those who lack confidence in spontaneous L2 interaction. The greatest tension, then, is perhaps not just about aligning policy and practice, but lies between some of the above principles behind this set of policy and practice, and the target L2 interactional competence by which the validity of the assessment task is determined. This competence, as argued above, entails spontaneous production of talk exchange in L2 predicated on genuine needs for communication (information/opinion gaps to bridge).

The findings of this study also bring to light the immense difficulty to reconcile the formative and summative elements of an assessment-for-learning initiative such as the SBA in Hong Kong. This is best summarized in Hamp-Lyons's (2009) remark that it needs to be 'meaningful at the level of the individual school and classroom', and at the same time, 'be accountable territory-wide' and 'meet the traditional expectations of rigour for summative reporting' (p. 525). The current practices in task implementation by teachers and task engagement by students, as reflected in this study and some of the previous research (Luk, 2010; Fok, 2012), seem to primarily serve the aim of creating optimal impressions of performance for scoring purposes (Luk, 2010). As it stands, the English SBA has yet to accomplish being a valid assessment that fully reflects the L2 interactional competence the task is designed to assess, and to serve the pedagogical goal of developing students' competence in conducting spontaneous L2 interaction with peers. More research is needed to refine the implementation of assessment for learning, both in the Hong Kong context and in general, in order for it to truly fulfill its purpose.

Based on the findings from this study, and subject to further empirical validation, the following recommendations for the assessment policy on task implementation can be made. Students can be given an amount

of preparation time just enough to brainstorm ideas and research on language items, but not for pre-scripting the interaction. Alternatively, aligning with the assessment-for-learning initiative, teachers can allow pre-planning and pre-scripting the interaction in practice assessments at early stages of the upper-secondary curriculum to accommodate weaker students, with a goal of gradually moving students towards spontaneous interaction in the graded assessments.

3.5.2 Limitations and future directions

This investigation of task implementation and engagement is necessarily exploratory. Given a small sample and the known diversity of assessment practices, I do not claim extensive generalizability of the study results. However, there is reason to believe that aspects of task implementation and engagement shown in this study are representative of a common practice in Hong Kong schools, as Fok (2012) and Luk (2010) have also provided evidence of pre-scripting. Furthermore, the mock assessment data can be considered a faithful reflection of the pre-task planning activities students engage in before the assessed interaction. Students were cooperative and did not exhibit any behavior that oriented to the mock assessment as anything less serious than the actual assessment. As acknowledged before, preparation time was reduced, and some differences in the planning activities were thus inevitable, but these were addressed in the post-interview. Future studies can, where practical conditions allow, gather larger samples of mock assessments for more generalizable results about pre-task planning activities. Controlled experimental studies would also be useful to determine the optimal pre-task planning time and conditions for the assessed interaction.

Acknowledgements

I would like to thank Prof John Joseph, the two Editors of this volume, and an anonymous reviewer for their very helpful comments on earlier drafts of this chapter. All remaining errors are my own.

Appendix 3.1 Additional transcription symbols

\\words	beginning of non-verbal action simultaneous with speech
\\((actions))	
first letter underlined	sequence of words each uttered with hearable effort or emphasis
…	rest of the turn omitted

References

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

Barraja-Rohen, A-M. (2011). Using conversation analysis in the second language classroom to teach interactional competence. *Language Teaching Research*, *15*(4), 479–507.

Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing*, *28*(2), 221–249.

Davison, C. (2007). Views from the chalkface: School-based assessment in Hong Kong. *Language Assessment Quarterly*, *4*(1), 37–68.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *19*(4), 474–509.

Fok, W. K. (2012). *HKCEE English Language school-based assessment: Its implementation at the frontline*. Unpublished doctoral thesis, Durham University.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*(4), 585–602.

Gan, Z. (2011). An Investigation of Personality and L2 Oral Performance. *Journal of Language Teaching and Research*, *2*(6), 1259–1267.

Gan, Z. (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Language Assessment Quarterly*, *9*(2), 133–151.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2008). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics*, *30*(3), 315–334.

Hamp-Lyons, L. (2009). Principles for large-scale classroom-based teacher assessment of English learners' language: An initial framework from school-based assessment in Hong Kong. *TESOL Quarterly*, *43*(3), 524–530.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET–SET group discussion. *Language Testing*, *23*(3), 370–401.

HKEAA (2009). 2012 Hong Kong diploma of secondary education examination English language: School-based assessment teachers' handbook. Retrieved February 21, 2014, from http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/doc/SBA_handbook_2012_ENG_240709.pdf

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*(3), 401–436.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 13–31). Amsterdam: John Benjamins.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372.

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly*, *7*(1), 25–53.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446–466.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.
Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, *31*(2), 147–175.
Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing*, *18*(4), 463–481.
Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing*, *2*(1), 31–40.
Tavakolian, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam: John Benjamins.
van Lier, L. (1996). *Interaction in the language curriculum. Awareness, autonomy and authenticity.* London: Longman.
Widdowson, H. G. (1979). *Explorations in applied linguistics*. Oxford: Oxford University Press.
Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 101–122.
Wigglesworth, G. (2000). Issues in the development of oral tasks for competency-based assessments of second language performance. In G. Brindley (Ed.), *Studies in immigrant English language assessment* (Vol. 1. Research Series 11, pp. 81–124). Sydney: National Centre for English Language Teaching and Research, Macquarie University.
Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*(1), 1–24.
Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing*, *22*(4), 463–508.

4

The Impact of Test Mode on the Use of Communication Strategies in Paired Discussion

Yan Jin and Lin Zhang

4.1 Introduction

In order to meet the growing demands for oral English proficiency certification among college students in China, the National College English Testing Committee, in collaboration with an IT company, has recently developed the computer-based College English Test—Spoken English Test (CET-SET) to replace the traditional face-to-face interview format. The previous face-to-face version used the group format of testing, where two examiners interviewed three candidates at a time. The candidates undertook several tasks as a group, one of which was a discussion task requiring the three of them to participate in a five-minute discussion among themselves (see National College English Testing Committee, 1999 for a detailed description). The computer-based CET-SET adopts similar testing procedures and tasks. However, it takes an examiner-absent paired format, in which two candidates are randomly paired and work on the test tasks on their own. One of the tasks is paired discussion, where the two candidates discuss a given topic through earphones. The reason for the switch from a group format to a paired format is the difficulties that raters may encounter in discerning more than two voices in the recordings when scoring the discussion task. This computer-mediated non-face-to-face paired discussion task, however, deserves further investigation because computer-based speaking tests typically require test-takers to talk to the computer and produce monologues only. A major fairness concern with this innovative approach to assessing speaking is that the mode of test delivery might constitute a construct-irrelevant factor affecting test performance on the paired discussion task. In other words, some test-takers could be disadvantaged by the computer-mediated interaction.

Test fairness has always been a central concern among language test developers and test users, especially in the context of high-stakes testing. Some researchers argue that fairness is an aspect of validity, that is, a test has to be fair to be valid (Willingham & Cole, 1997; Xi, 2010). Whatever weakens fairness also compromises the validity of a test. Based on this conceptualisation of fairness, gathering multiple types of fairness evidence should be an important part of test validation. To ensure test fairness, as Xi (2010) points out, it is important that construct-irrelevant factors, among others, produce no systematic and appreciable effects on test results. There is thus an urgent need to build fairness/validity arguments for the paired discussion task used in a computer-delivered speaking test. A preliminary small-scale validation study was therefore conducted at the initial stage of the implementation of the computer-based CET-SET in the hope of gaining useful insights into the potential effect of the test mode on test performance in the paired discussion task with a special focus on test takers' use of communication strategies.

4.2 Literature review

4.2.1 Oral communication strategies

Selinker (1972) first proposed the notion of communication strategies as one of the five central processes involved in second language learning. Canale and Swain (1980) referred to communication strategies as strategic competence within their framework of communicative competence. The incorporation of strategic competence into the communicative competence framework and the increasing importance attached to communicative language skills have aroused considerable interest in researching communication strategies in the field of second language acquisition.

4.2.1.1 Conceptualisation of communication strategies

There have been essentially two approaches to defining communication strategies: the psycholinguistic definition and the interactional definition. From the psycholinguistic perspective, communication strategies are defined as "potentially conscious plans for solving what to an individual presents itself as a problem in reaching a particular communicative goal" (Færch & Kasper, 1980: 81). Interactionally oriented researchers view communication strategies as "a mutual attempt of two interlocutors to agree on a meaning in situations where requisite meaning structures do not seem to be shared" (Tarone, 1980: 420).

As manifested in the definitions, the psycholinguistic approach focuses overwhelmingly on individual production, whereas the interactional approach places emphasis on the joint construction of discourse between two interlocutors. Because of their differing opinions on conceptualising communication strategies, researchers also have divergent views as to the taxonomies of strategies. What the two approaches have in common is that there was a tendency in earlier research on strategy use to restrict the concept to problem-solving activity (Kasper & Kellerman, 1997). When faced with a communication problem, speakers may either change or abandon their original communicative goal by using avoidance/reduction strategies, or attempt to maintain their original aim by resorting to achievement/compensatory strategies.

Canale (1983) extended the notion of communication strategies to include "any attempt to enhance the effectiveness of communication" (Dörnyei & Scott, 1997: 179). Thus, communication strategies are seen as more than devices for handling communication problems. Bachman (1990) gives strategic competence a central position in his model of communicative language ability, viewing it as "an important part of all communicative language use, not just that in which language abilities are deficient and must be compensated for by other means" (p. 100). Similar to Bachman's broader view of strategic competence, the Common European Framework of Reference for Languages (Council of Europe, 2001, henceforth CEFR) defines communication strategies as "the adoption of a particular line of action in order to maximise effectiveness" (p. 57).

The CEFR differentiates a series of categories of communication strategies in its description of communicative language activities. Of relevance to oral activities are three categories: production, interaction and nonverbal. Production strategies are defined as attempts made by language users to mobilise and balance their internal resources in order to complete the task successfully. Following the reduction-achievement distinction in the literature, the CEFR describes production strategies as consisting of achievement and avoidance strategies (North, 2000). In explicating the category of interaction strategies, the CEFR gives an account of strategies exclusive to spoken interaction which entails the collective creation of meaning among participants, including turn-taking (taking the floor), cooperating (e.g. eliciting and referring to others' contributions), and meaning-negotiation (e.g. asking for clarification). Nonverbal strategies are not the focus of the study and will not be discussed here.

4.2.1.2 Methodologies for communication strategy research

Proponents of the psycholinguistic approach generally resort to elicitation tasks (e.g. picture description) for eliciting and analysing strategy tokens, focusing heavily on individual production. The interactional perspective, as Ellis (1985) argues, is best tackled by discourse analysis, which considers the joint contribution of the two interlocutors, rather than singling out the learner's activity for separate analysis.

In order to minimise subjectivity in strategy analysis, one solution is to look for clear strategy markers in the performance data. What constitutes a strategy marker depends largely on what one considers a strategy to be. Some researchers who adopt an interactional approach take the view that the data that constitute evidence of strategic behaviour are those utterances marked by a speaker in some way as requiring specific attention on the part of the listener. The psycholinguistic approach, however, locates and identifies strategies in relation to both explicit and implicit signals. Færch and Kasper (1983) suggest looking for three kinds of problem indicators: 1) implicit signals of uncertainty, such as filled pauses and self-repairs; 2) explicit signals of uncertainty: expressions such as "I don't know how to say this"; 3) direct appeals for assistance, such as "how do you say it in English?" The second solution is to get more than one person involved in strategy analysis to enhance reliability.

In recent decades, the conversation analysis (CA) methodology has been increasingly applied to examine oral discourses for evidence of strategy use that constitutes an important source of validity evidence. There is now a general consensus in the language testing community that CA is a viable approach to understanding candidate language within the context of an oral assessment, particularly when the focus of the research is on interaction (Chalhoub-Deville, 2003; Galaczi, 2004, 2008; Lazaraton, 2002; May, 2007).

4.2.1.3 Conceptualisation of communication strategies in the present study

Similar to the broader notion proposed by the CEFR, communication strategies are defined in the study as any attempts that language users make to facilitate communication and maximise communication effectiveness. The criteria for selecting strategies include: 1) they are clearly defined in the literature; 2) they appeared in a number of major taxonomies; 3) they are relevant to the context of the present study.

Following the production-interaction distinction made by the CEFR, we distinguished two categories of strategies that are non-interactional

or interactional in nature. The reasons are twofold: 1) This study aims to investigate strategy use in spoken interaction which involves not only the speech production of an individual but also the mutual interaction between the interlocutors; 2) Both production strategies and interaction strategies, in our view, contribute to the accomplishment of communication goals and the enhancement of communication effectiveness.

Production strategies comprise achievement, avoidance and stalling strategies. The first two subcategories are conceived of as problem-solving attempts that a speaker makes by resorting to his/her own linguistic resources. Stalling strategies, which help speakers gain time to think, do not engage the interlocutor's support and are thus non-interactional in nature. Interaction strategies consist of turn-taking, cooperative and problem-related strategies. Turn-taking and cooperative strategies included in the CEFR are incorporated into our taxonomy, which, in our view, contribute much to the naturalness and interactiveness of the communication. Problem-related interaction strategies are joint efforts made by the interlocutors to solve their mutual problems arising in communication.

A coding scheme was developed based on the taxonomy of strategies we proposed for this study. After a pilot study conducted among two pairs of candidates prior to the present study, we decided to focus on thirty types of oral communication strategies which were grouped into five sub-categories (see Appendix 4.1). Turn-taking was investigated as a separate strategy of oral communication.

4.2.2 Paired discussion in a computerised testing context

In recent years, the use of computer technology has been an important feature of language assessment, including the testing of oral proficiency. However, there is still no consensus as to which method is the most effective for computer-based oral assessment and whether computer-delivered speaking tests can provide a valid alternative to face-to-face oral assessments. Among the important issues that need to be addressed concerning the use of computer technology is the potential impact of the mode of delivery on candidate discourse, in other words, the test method effect (Bachman & Palmer, 1996), which has implications for the nature of the input as well as the expected response (Kiddle & Kormos, 2011).

In the past decades, a number of studies have been undertaken to examine the equivalence of semi-direct speaking tests with their direct counterparts (e.g. Brown, 1993; O'Loughlin, 1997; Shohamy, 1994). The majority of these studies correlated scores across the two test modes,

and research findings revealed a satisfactorily high correlation between semi-direct and direct tests. Some researchers also looked at the discourse features of candidate output through qualitative analyses (e.g. O'Loughlin, 1995; Shohamy, 1994). Although correlations between the two tests have generally been high, comparative analyses of candidate discourse have unveiled distinctions between the language samples elicited by the two different modes of testing.

However, these studies almost exclusively examined candidate discourse that was non-interactive in nature, as the semi-direct tests used in these studies were basically a replication of the prompt-response format of a one-to-one oral interview. Since a paired discussion task has rarely been incorporated into a computer-mediated speaking test so far, little research is available that investigates interactional candidate discourse produced through such a mode of delivery. The present study is thus an attempt to provide insights into the impact of test mode on candidate discourse elicited in the paired discussion task by comparing a computer-delivered oral assessment with its face-to-face equivalent.

As discussed in the previous section, communication strategies are any means the speaker exploits in order to fulfil the demands of communication in context and successfully complete the task in question (Council of Europe, 2001). They certainly play a significant role in candidate interaction in a test situation. Given candidates' awareness that their performance in the discussion task may be judged based on their and their partner's joint contribution, the assumption can be made that they may employ some communication strategies more consciously and frequently than in non-test situations. Out of these considerations, the present study seeks to explore the impact of test mode on test performance in the paired discussion task with a special focus on communication strategy use. The following two research questions are to be answered in the study:

RQ1: How do test-takers' performances differ in the paired discussion task of the computer-based CET-SET and the face-to-face interview test in terms of the quantity and the variety of communication strategies?

RQ2: What is the possible relationship between the use of communication strategies and the effectiveness of communication in the paired discussion task of the computer-based CET-SET and the face-to-face interview test?

4.3 Research methodology

4.3.1 Participants

The participants were twelve (six pairs) students who sat the live computer-based CET-SET in May 2012. To have a better view of strategy use across different oral proficiency levels, the participants were selected according to their CET scores on the assumption that there is some degree of correlation between one's listening, reading and writing skills and one's oral proficiency. Three research assistants, two with a doctoral degree and one a master's degree in applied linguistics, were involved in data transcription and coding. A training session was conducted prior to the coding of the data. The inter-coder reliabilities for the major types of strategies proved satisfactory (r = .83~.92).

4.3.2 Data collection

Immediately after they completed the computer-based CET-SET, the twelve students took an experimental test, which was a replication of the face-to-face CET-SET. But instead of using the group format where three students formed a group for the discussion task, the experimental test employed a paired discussion format. The examiners of the experimental test did not participate in the discussion. The testing conditions of the discussion task in the computer-based CET-SET and the experimental test were therefore equivalent except for the way the two candidates talk to each other: via earphones or face to face. To minimise the topic effect, we chose a retired CET-SET topic for the discussion task of the experimental test to match the one used in the computer-based CET-SET in terms of the content and the degree of familiarity to test-takers (see Appendix 4.2).

The participants' responses were recorded by the computer system in the computerised test. Their performances in the experimental test were videotaped. To examine the relationship between strategy use and effectiveness of communication in paired discussion, the criteria for scoring communication effectiveness were developed on the basis of literature review (see Appendix 4.3). Two raters listened to the responses of the computer-based discussion as well as the face-to-face discussion in the experimental test and scored independently using a holistic scale of 1 to 5, with 5 indicating the highest level of effectiveness and 1 the lowest. If the scores assigned by the two raters differed by more than one band on the holistic scale, they were scored again by a third rater in order to resolve the discrepancies.

4.3.3 Data transcription and coding

The speech data elicited in the paired discussion task of the two tests were first transcribed following the CA transcription conventions (Atkinson & Heritage, 1984) with slight modifications (see Appendix 4.4). The transcripts were double-checked carefully by the researchers to ensure accuracy of transcription.

4.3.3.1 Coding the major types of strategies

Strategies were identified and coded based on the coding scheme (see Appendix 4.1). To avoid subjectivity, we identified achievement and avoidance strategies based on the problem indicators recommended by Færch and Kasper (1983). Stalling strategies can be easily singled out as they are marked by clear verbal signals. Interaction strategies, as visible behaviour involving joint contribution or attention of the two interlocutors, are also relatively easy to identify.

4.3.3.2 Coding the turn-taking strategy

Turn-taking refers to how turns at talk are structured and what order is followed. An ideal conversation, according to Sacks, Schegloff and Jefferson (1974), features one party speaking at a time. However, in much naturally occurring talk-in-interaction, the orderliness of this turn-exchange principle is often violated by incidences of simultaneous talk, such as interruptions and overlaps. To facilitate the analysis of the turn-taking strategy, the speech data were segmented into turn units, words and turn length, and coded for such conversational devices as interruptions and overlaps. Backchannels and inter-turn pauses were also examined.

- *Data Segmentation: turn units*. A turn, as defined by Stenstrom (1994), is everything the current speaker says before the next speaker takes over. But not all utterances are proper turns. Non-turns were also identified in this study and limited to backchannel responses such as "mm", "uh huh" and "yeah", which are interactional features signalling comprehension, agreement and encouragement on the part of the listener and do not involve a speaker shift (ibid.).
- *Data Segmentation: words*. The reasons for choosing the word as the analytic unit to provide an estimate of quantity of talk are twofold: 1) it has proved to be a useful measure of quantity of talk in various studies of interaction (Itakura, 2001); 2) it is a more accurate measure of amount of talk as it is directly proportionate to floor presence

(Galaczi, 2004). Following Galaczi, filled pauses and incomprehensible speech were excluded from the word count, but the unfinished words were included in the count if the meaning was recognisable.

- *Data segmentation: turn length.* The length of turns is indicated by the total number of words contained in a turn. A comparison of the distribution of turn lengths between the two discussion tasks was made in this study.
- *Data coding: interruptions and overlaps.* Interruptions and overlaps, as means to obtain the floor, are fundamental aspects of the turn-taking system. Overlap, according to Sacks et al. (1974), is simultaneous talk that begins at the transition relevance place (TRP). TRP refers to places where a speaker's talk is possibly complete and speaker change could happen (Liddicoat, 2007). Interruption is defined as simultaneous talk which starts at a non-TRP (Levinson, 1983).

4.3.4 Data analysis

To address the first research question, we adopted mainly frequency analyses for detailed comparisons of the quantity and variety of strategies used in the computer-based paired discussion (CB-PD henceforth) and the face-to-face paired discussion (FF-PD henceforth). To answer the second research question, we compared the strategy use of the two candidates who achieved the highest scores on communication effectiveness with that of the other two candidates who obtained the lowest scores on communication effectiveness.

4.4 Results

4.4.1 The quantity and variety of communication strategies

4.4.1.1 Overall frequency of strategy use

The data were first summarised by comparing the total number of communication strategies used in the CB-PD and FF-PD. Stalling strategies were presented separately for the following reasons: 1) they are somewhat distinct from the other strategy types in terms of level of consciousness involved in strategy use, or more specifically, they are often subconsciously or unconsciously employed by the speaker, whereas all the other strategies are more of conscious efforts made by the speaker; 2) The data revealed the test-takers' heavy reliance on the stalling strategies in both tests, whose number of occurrences was out of proportion to that of any other type of strategy.

Due to the variation in the length of discussions between the two tests caused by less strict time control in the experimental test, and, more importantly, differences in speech rate among speakers, we compared the number of strategies per hundred words in each test instead of the total number of strategies. The result (see Table 4.1) revealed a striking similarity in the overall frequency of strategy use between the FF-PD (4.6) and the CB-PD (5.1). The number of occurrences of each strategy was then calculated, along with the number of test-takers (TT) using each strategy. Percentages (%) of the occurrences of each strategy in proportion to the total number of strategy occurrences were also presented.

As shown in Table 4.1, a large number of communication strategies were employed by the test-takers in either FF-PD or CB-PD. Moreover, a variety of strategies occurred in both types of discussions (see highlighted strategies). No marked difference was found in the percentages of most of them between the two types of discussion tasks. In addition, the frequencies of these strategies are basically in the same order in the two tasks, from the highest, self-repair, to the lowest, paraphrase. The top six strategies used most frequently in both tasks include giving feedback, self-repair, asking a question, restructuring, referring to partner's contributions and message abandonment. And these strategies were employed by at least five test-takers in both tasks, as indicated by the number of test-takers using each strategy. By contrast, the strategies that were present in one task alone were employed by very few test-takers (almost exclusively by one test-taker), and they occurred only occasionally (mostly once or twice).

Of all the 30 types of communication strategies (turn-taking excluded), 16 strategies occurred in the FF-PD, and a similar number, 18, were used in the CB-PD. Over half of them (11 types) appeared in both testing situations, demonstrating a fairly high degree of similarities in the variety of communication strategies used in the two tasks. Despite the marked similarities in overall strategy use, some variation was observed in the strategy of asking a question, which was employed more frequently and by more test-takers in the CB-PD than in the FF-PD.

In addition to the specific types of strategies, the various categories and subcategories of strategies were also examined. As seen in Table 4.2, the use of cooperative strategies was notably more frequent in the CB-PD than in the FF-PD (see DISCUSSION). Avoidance strategies were used least frequently in both tasks, and problem-related interaction strategies were also infrequently employed.

Table 4.1 Frequency of strategies used in each discussion task and number of test-takers using each strategy

		FF-PD			CB-PD		
		frequency	**%**	**TT**	**frequency**	**%**	**TT**
1	paraphrase	1	0.9	1	1	0.7	1
2	approximation	–	–	–	1	0.7	1
3	use of all-purpose words	1	0.9	1	–	–	–
4	restructuring	10	9.1	8	6	4.1	5
5	word-coinage	–	–	–	–	–	–
6	self-repair	41	37.3	11	46	31.5	11
7	literal translation	–	–	–	–	–	–
8	foreignising	–	–	–	–	–	–
9	code-switching	1	0.9	1	–	–	–
10	nonverbal	2	1.8	2	–	–	–
11	topic avoidance	–	–	–	–	–	–
12	message abandonment	5	4.5	5	7	4.8	5
15	asking a question	12	10.9	5	30	20.5	12
16	eliciting opinions	–	–	–	1	0.7	1
17	giving feedback	21	19.1	11	37	25.3	12
18	referring to partner's contributions	5	4.5	5	8	5.5	6
19	appealing for help	3	2.7	2	–	–	–
20	responding to help	–	–	–	–	–	–
21	asking for clarification	–	–	–	–	–	–
22	responding to clarification requests	–	–	–	–	–	–
23	requesting repetition	–	–	–	1	0.7	1
24	responding to repetition requests	–	–	–	1	0.7	1
25	seeking confirmation	3	2.7	1	2	1.4	2
26	responding to confirmation requests	3	2.7	1	2	1.4	2
27	comprehension checks	–	–	–	1	0.7	1
28	responding to comprehension checks	–	–	–	1	0.7	1
29	expressing non-understanding	2	1.8	1	–	–	–
30	other repair	–	–	–	1	0.7	1
Total		110	100		146	100	
Number of words		2367			2879		
Strategies per hundred words		4.6			5.1		

Note: FF-PD: Face-to-face paired discussion. CB-PD: Computer-based paired discussion. %=Percentage of occurrences of each strategy in proportion to the total number of strategy occurrences. TT=Number of test-takers using each strategy.

Table 4.2 Frequency of subcategories of strategies used in each discussion task

Strategy	FF-PD		CB-PD	
	Frequency	%	Frequency	%
SC1: Achievement strategies (1–10)	56	50.9	54	37.0
SC2: Avoidance strategies (11–12)	5	4.5	7	4.8
SC4: Cooperative strategies (15–18)	38	34.5	76	52.1
SC5: Problem-related strategies (19–30)	11	10.0	9	6.2
Subtotal: Production strategies (1–12)	61	55.5	61	41.8
Subtotal: Interaction strategies (15–30)	49	44.5	85	58.2
Total	110	100	146	100

Note: FF-PD: Face-to-face paired discussion. CB-PD: Computer-based paired discussion. SC=Subcategory. SC3, stalling strategies, was dealt with separately.

4.4.1.2 Use of stalling strategies

Stalling strategies consist of use of fillers and self-repetition. The results (see Table 4.3) show that stalling strategies occurred with a high frequency in both FF-PD and CB-PD, far more frequently than other individual or subcategories of strategies (see Tables 4.1 and 4.2), manifesting the test-takers' heavy reliance on stalling strategies in performing the paired discussion task.

4.4.1.3 Use of turn-taking strategies

The data in relation to turn-taking strategies were first summarised by counting the number and length of turns in each dialogue of the two types of discussion tasks, the results of which are presented in Table 4.4. In both FF-PD and CB-PD, the two test-takers in each pair produced a similar number of turns, indicating that they were generally able to turn-take effectively in the discussion and manage the interaction, irrespective of the presence or absence of examiners. The speech samples produced in the CB-PD, however, contained more turns and words than those elicited in the FF-PD.

When investigating the length of turns generated in the two types of tasks, we found a marked difference in the proportion of brief turns with less than five words in length, as indicated in Figure 4.1. As the speech data showed, many of these brief turns were conversation openings like "Hello", and conversation management techniques like "You first?", which are characteristic of non-face-to-face interactions. Another attributing factor was the presence of indiscernible turns (calculated as zero in turn length) in the CB-PD arising from simultaneous talk (see

Table 4.3 Use of stalling strategies (SC3: 13–14) in each discussion task

	Frequency	Number of words	Strategies per hundred words
FF-PD	412	2367	17.4
CB-PD	593	2879	20.6

Note: FF-PD: Face-to-face paired discussion. CB-PD: Computer-based paired discussion.

Table 4.4 Number of turns and range of turn lengths in each discussion task

		FF-PD			CB-PD		
Pair	**Test-taker**	**No. of turns per test-taker**	**Words per test-taker**	**Range of turn length**	**No. of turns per test-taker**	**Words per test-taker**	**Range of turn length**
1	1	7	268	11–70	9	286	1–76
	2	6	245	7–73	8	257	5–66
2	3	10	198	1–42	15	364	1–78
	4	10	181	1–48	15	238	1–68
3	5	4	126	1–52	7	151	1–75
	6	5	137	3–51	6	254	1–109
4	7	12	209	1–53	12	244	0–79
	8	12	169	0–96	10	163	1–84
5	9	3	248	63–106	6	197	1–93
	10	3	355	63–173	5	332	1–110
6	11	2	80	31–49	5	129	1–51
	12	2	151	56–95	6	264	1–84
Total		76	2367	0–173	104	2879	0–110

Note: FF-PD: Face-to-face paired discussion. CB-PD: Computer-based paired discussion. Turn length: Total number of words contained in a turn.

DISCUSSION). In addition, compared with the face-to-face discussion, the CB-PD generated fewer turns between 10 and 50 words in length, but more turns ranging from 70 to 90 words in length.

To gain a fuller picture of the turn-taking strategies, we also calculated the occurrences of overlaps, interruptions, backchannels and inter-turn pauses in the talk. As seen in Table 4.5, two different turn-taking styles were evident. Two dyads (Pairs 2 and 4) applied a very active turn-taking strategy and produced a highly interactive conversation featuring brief turns, swift speaker change, and fairly frequent use of backchannel responses, overlaps and interruptions. Most of the dyads, in contrast, produced longer turns and followed a somewhat mechanical turn-taking sequence.

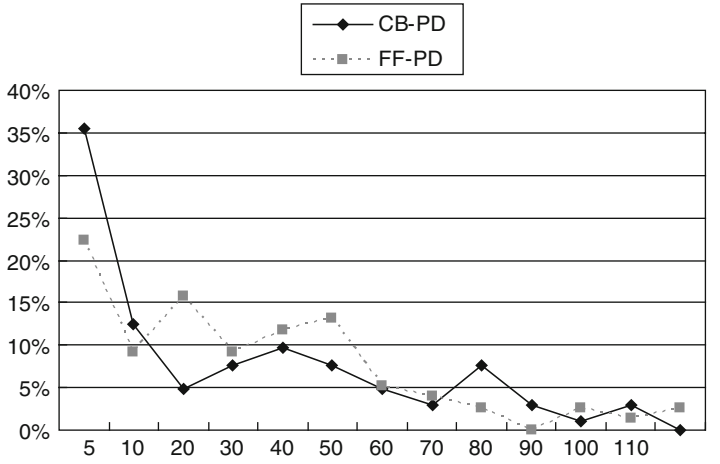


Figure 4.1 Proportion of turn lengths in the two tasks

Table 4.5 Frequency of turn-taking strategies

	Pair 1		Pair 2		Pair 3		Pair 4		Pair 5		Pair 6		Total
	1	2	3	4	5	6	7	8	9	10	11	12	
FF-overlap			1	1									2
FF-interruption	1		1					2					4
FF-backchannel response			2	1				2		2			7
FF-inter-turn pause length of pause							4 2–3s				1 7s		5
FF-simultaneous talk													0
CB-overlap		1	4					1					6
CB-interruption								1				1	2
CB-backchannel response	1		3	2				2					8
CB-inter-turn pause length of pause	3 2s		3 2–3s		4 2–3s		4 2s		5 1–3s		1 1s		20
CB-simultaneous talk			7				7		2		3		19

Note: FF: Face-to-face. CB: Computer-based. s = Second.

It should tbe noted that some unexpected findings emerged in the process of analysing the turn-taking strategies. As shown in Table 4.5, the inter-turn pauses occurred only occasionally in the FF-PD. But they appeared in all the dialogues in the CB-PD and very frequently in some

of them. Meanwhile, frequent occurrences of simultaneous talk were also apparent in the CB-PD, particularly in the two highly interactive dialogues produced by dyad 2 and dyad 4. However, much of the overlapping speech was hard to discern.

4.4.2 The relationship between strategy use and communication effectiveness

The relationship between strategy use and communication effectiveness in the discussion task was also explored to gain some insight into the role of communication strategies in the completion of the paired discussion task. The raters assigned a holistic score on the communication effectiveness in the discussion task of each test based on the scoring criteria (Appendix 4.3). Worthy of note is the finding that frequent occurrences of simultaneous talk in the two highly interactive dialogues in the CB-PD, which made part of the speech difficult to comprehend, affected test-takers' scores on communication effectiveness.

Therefore, two test-takers who scored the highest on communication effectiveness in both FF-PD and CB-PD and the other two who scored the lowest in both tasks were compared in terms of their strategy use. Table 4.6 illustrates the frequency of the two categories of strategies (stalling strategies excluded) and the types of strategies used by each test-taker.

As can be seen from Table 4.6, the high-scoring test-takers were characterised by their frequent use of interaction strategies (more than 70% in both FF-PD and CB-PD). The low-scoring test-takers, in contrast, resorted much less frequently to interaction strategies in the two tasks. Moreover, the high-scoring test-takers employed a wider range of interaction strategies than their low-scoring counterparts in both tasks, as indicated by the types of strategies used. The score on communication effectiveness in the paired discussion task, therefore, seems to correlate with both the quantity and the variety of interaction strategies used by the test-takers in both tasks. However, this conclusion is only tentative, due to the small sample size.

4.5 Discussion

The results of the study indicate, on the whole, a high degree of similarities in the quantity and variety of communication strategies used in the computer-based paired discussion and the face-to-face paired discussion, though some differences were found in the frequencies of cooperative strategies. The test-takers resorted to a range of achievement

Table 4.6 Strategies used by test-takers with the highest or lowest scores on communication effectiveness

TT	FF-PD					CB-PD				
	Score	Production strategies		Interaction strategies		Score	Production strategies		Interaction strategies	
		%	Type	%	Type		%	Type	%	Type
1	4	28.6	1	71.4	3	4.5	13.3	1	86.7	6
10	4	25.0	2	75.0	3	4	11.1	1	88.9	3
11	2	75.0	3	25.0	2	3	44.4	2	55.6	2
12	3	85.7	3	14.3	1	3	76.9	2	23.1	2

Note: FF-PD: Face-to-face paired discussion. CB-PD: Computer-based paired discussion. TT=Test-taker. Score=Score of communication effectiveness.

strategies, rather than avoidance strategies, in both types of discussion tasks, suggesting that in most cases, they attempted to keep to, instead of reducing or abandoning, their original communication goals at times of language difficulty. They also employed a variety of interaction strategies, following the cooperative principle, to make joint contributions to the development of the conversation. It is worth noting that the strategy of asking a question, one type of cooperative interaction strategy, was used more frequently in the computer-based discussion than in the face-to-face discussion. The analysis of the speech samples showed that the strategy was more often used as a means of giving floor or initiating a topic in the computer-mediated discussion to help maintain the non-face-to-face interaction where nonverbal clues (eye contact, facial expressions, etc.) were lacking.

The data also unveiled an excessive use of stalling strategies in both types of discussion tasks, which may be attributable to the fact that as learners of a foreign language, the test-takers, with limited target language resources, needed more time to formulate messages. Unless speech is pre-planned, as Fulcher (1993) points out, hesitations and reformulations will abound for native and non-native speakers, since it is in practice linked to forward planning and lexical choice. The overuse of these strategies may also be a result of transference of L1 speaking habits (He & Liu, 2004).

The investigation of turn-taking strategies indicated that the test-takers were generally able to turn-take effectively in the discussion and exchange ideas actively with their partner. While only a few test-takers displayed a highly interactive conversation style featuring brief turns and frequent speaker shift, most speakers produced longer turns and preferred a neat and orderly turn-taking style. Well worthy of mention is the finding that simultaneous talk and inter-turn pauses took place more frequently in the computer-based discussion than in the face-to-face discussion. An in-depth discourse analysis suggested the test-takers had difficulty predicting the transition relevance place (TRP) accurately without any nonverbal clues in the computer-mediated interaction, so they waited until they made sure they partner's turn had completely ended, which resulted in an increased amount of silence. Some test-takers may have misinterpreted some features in their partner's speech, such as a brief pause or a falling intonation, as turn-closing signals. Upon receiving these signals, they took over the floor, only to find that their partner was still holding the turn. Simultaneous talk of this kind occurred more often in the two highly interactive dialogues, generating a number of incomprehensible utterances. These findings seem to

suggest that test-takers favouring a highly interactive conversation style may be unfairly disadvantaged in the computer-based paired discussion task.

The results of the investigation into the relationship between communication effectiveness and strategy use in the paired discussion task seem to suggest that interaction strategies contribute to improving the effectiveness of communication and accomplishing the communication goals in the discussion. Therefore, effective use of these strategies may help enhance test performance in a speaking task involving peer-to-peer interaction.

The results of the study have important practical implications for further improvements of the test design. The research findings, for example, unveiled some problems of the computer-mediated interaction, the most serious being that frequent occurrences of incomprehensible simultaneous talk in the discussion arising from misinterpretation of turn-closing signals affected adversely the test-takers' test performance. A possible solution to the problem is to allow the two test-takers in the pair to sit next to each other so that they can talk face to face through earphones in the computerised test. Or they can see each other on the computer screen with web cameras and high-speed connection, which is similar to video chatting.

The limitation of the sample size in the study, however, is apparent due mainly to practical difficulties with data collection and the subsequent analysis of the large amounts of data. First, lack of motivation on the part of test-takers was one of the major obstacles. The students participated in the experiment on a voluntary basis. They needed to sit an additional oral interview test after they took the live speaking test. Moreover, the results of the interview test would not affect in any way their final score on the live computer-based speaking test. The cost involved in data collection and the degree of collaboration on the part of the test centre were also determining factors. Last but not least, the amount of work in the data analysis phase had to be considered as well because data transcription and data coding proved extremely labour-intensive and time-consuming.

4.6 Conclusion

This chapter has reported the findings of a small-scale validation study which, using the CA methodology, compared the use of communication strategies in the paired discussion task of the face-to-face oral interview test and the computer-mediated speaking test. The delivery

mode of testing, as the data suggested, had little impact on the use of communication strategies in the paired discussion task, thus providing some empirical validity and fairness evidence that lends support to the use of paired format in a computer-based speaking test. In terms of research methodology, this study has shown that CA, which has been successfully used in the past to examine interactional patterns and discourse features in speaking tasks, can also be successfully applied in the investigation of communication strategy use in peer-to-peer interaction.

The results of the study, however, should be interpreted with caution due mainly to the small sample size. Future studies need to involve a larger number of participants in order to generate more reliable results. It is also felt necessary in future studies to exercise a tighter control of the time for the discussion task so that the quantity of communication strategies used in the two tests can be compared directly. In addition, post-test interviews with test-takers can be conducted for more supporting evidence of their strategy use as well as their perceptions of the computerised test. It is also worth investigating variables such as gender, personality, computer anxiety, and computer familiarity, all of which may cause variation in the use of communication strategies between the two testing situations.

Difficult as it is to collect and analyse data of strategy use in speaking tests, we believe that studies on this topic will provide substantial groundwork for continued research into the use of a paired discussion task in a computer-based speaking test.

Appendix 4.1 Coding scheme used in the study

Strategy	Explanation
Production strategies: *Achievement strategies*	
1 paraphrase	rewording of the desired item in L2 by means of description, definition, exemplification, etc.
2 approximation	using a word, such as a superordinate or a related term, which shares semantic features with the target item
3 use of all-purpose words	extending a general lexical item to contexts where specific words are lacking
4 restructuring	executing an alternative plan to communicate the intended message after running into difficulty in implementing the intended speech plan

(*continued*)

Appendix 4.1 Continued

Strategy	Explanation
5 word-coinage	creating new L2 words or compound words related to the original intended meaning, or creating a non-existent L2 word by applying L2 morphology
6 self-repair	making self-initiated corrections in one's own speech
7 literal translation	translating literally an expression from L1
8 foreignising	using an L1 word by adjusting it to L2 phonology and/or morphology
9 code-switching	using L1 items in place of L2 words or phrases
10 nonverbal	conveying the intended meaning by using facial expressions, gestures, eye contact, mime, etc.
Production strategies: *Avoidance strategies*	
11 topic avoidance	avoiding topics for a lack of linguistic resources
12 message abandonment	leaving a message unfinished because of some language difficulty
Production strategies: *Stalling strategies*	
13 use of fillers	using filled pauses and verbal fillers to fill a gap in the exchange, to stall, and to gain time in order to maintain the floor
14 self-repetition	repeating immediately what one has said to gain time in order to hold the floor
Interaction strategies: *Cooperative strategies*	
15 asking a question	inviting the interlocutor into the discussion by asking questions (asking for opinions, seeking information, giving floor, etc.)
16 eliciting opinions	soliciting the interlocutor's opinion
17 giving feedback	providing feedback on the interlocutor's speech to help the development of the discussion
18 referring to partner's contributions	referring to what the interlocutor has already said
Interaction strategies: *Problem-related strategies*	
19 appealing for help	eliciting help from the interlocutor by asking an explicit question concerning a gap in one's L2 knowledge; or requesting assistance indirectly by expressing one's lack of knowledge of an L2 item either verbally or nonverbally.
20 responding to help	providing assistance to the interlocutor
21 asking for clarification	asking the interlocutor to give further explanations

(*continued*)

Appendix 4.1 Continued

Strategy	Explanation
22 responding to clarification requests	providing clarification upon request
23 requesting repetition	requesting repetition when not hearing or understanding the interlocutor's speech properly
24 responding to repetition requests	repeating what one has said upon request
25 seeking confirmation	requesting confirmation that one heard or understood the interlocutor's speech correctly
26 responding to confirmation requests	confirming what the interlocutor has said
27 comprehension checks	asking questions to check whether the interlocutor can follow you, whether the interlocutor is listening, or whether the interlocutor can hear you
28 responding to comprehension checks	responding to the interlocutor's comprehension checks
29 expressing non-understanding	expressing verbally or nonverbally that one did not understand what the interlocutor said properly
30 other repair	correcting something in the interlocutor's speech

Appendix 4.2 Topics for the discussion tasks

Directions for the paired discussion task read by the examiner:

Now that we've talked briefly about ... , I'd like you to develop this topic further and have a discussion for about four and a half minutes. During the discussion you may argue and ask each other questions. Our discussion is about ... Remember, this is a pair activity and you need to interact with each other. So don't keep talking without giving the other a chance. Now let's begin.

Computer-based paired discussion: The last drop of water on earth will be your tear.

Face-to-face paired discussion: Is it possible for humans to conquer nature?

Appendix 4.3 Criteria for communication effectiveness

- Can cope with problems that occur in the course of communication and help keep the discussion going (e.g. using paraphrase or other

means in times of linguistic difficulty, asking for clarification and providing clarification, etc.);
- Can adhere to the cooperative principles in the discussion and exchange ideas with partner actively and effectively;
- Can accomplish the communication task effectively.

Appendix 4.4 Transcription conventions

(adapted from Atkinson and Heritage 1984)

(.)	Pause of less than 1 second
(3)	Approximate length of pause in seconds
:	Lengthened sound or syllable; more colons indicate greater lengthening
.	Falling intonation, indicating the end of an utterance
?	Rising intonation (not necessarily a question)
,	Level or low-rising intonation, indicating the continuation of an utterance
[ ]	Overlapping speech
-	Abrupt cutting off of sound
/ /	Phonetic symbols (e.g. /ən/)
(( ))	Nonverbal action (e.g. laughter)
( )	Doubtful transcription
#	No overlaps occur, but the speaker has not ended his/her talk when the interlocutor initiates his/her talk
underline	Instance of strategy use

References

Atkinson, J. M. & Heritage, J. (1984). *Structures of social action: Studies in conversation analysis*. Cambridge: Cambridge University Press.

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

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

Brown, A. (1993). The role of test-taker feedback in the test development process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, *10*(3), 277–303.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). New York: Longman.

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

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369–383.
Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
Dörnyei, Z. & Scott, M. L. (1997). Communication strategies in a second language: Definitions and taxonomies. *Language Learning*, *47*(1), 173–210.
Ellis, R. (1985). *Understanding second language acquisition*. Oxford: Oxford University Press.
Færch, C. & Kasper, G. (1980). Processes and strategies in foreign language learning and communication. *Interlanguage Studies Bulletin*, *5*(1), 47–118.
Færch, C. & Kasper, G. (1983). On identifying communication strategies in interlanguage production. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 210–238). London: Longman.
Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language*. Unpublished PhD dissertation, University of Lancaster.
Galaczi, E. (2004). *Peer-peer interaction in a paired speaking test: The case of the First Certificate in English*. Unpublished PhD dissertation, Teachers College, Columbia University.
Galaczi, E. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89–119.
He, L. Z. & Liu, R. J. (2004). 基于语料库的大学生交际策略研究 [Corpus-based Investigation into Communication Strategies in CET-SET]. *外语研究[Foreign Languages Research]*, 1, 60–65.
Itakura, H. (2001). Describing conversational dominance. *Journal of Pragmatics*, *33*(12), 1859–1880.
Kasper, G. & Kellerman, E. (1997). *Communication strategies: Psycholinguistic and sociolinguistic perspectives*. New York: Longman.
Kiddle, T. & Kormos, J. (2011). The effect of mode of response on a semi-direct test of oral proficiency. *Language Assessment Quarterly*, *8*(4), 342–360.
Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: UCLES/Cambridge University Press.
Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
Liddicoat, J. A. (2007). *An introduction to conversation analysis*. London: Continuum.
May, L. (2007). *Interaction in a paired speaking test: The rater's perspective*. Unpublished PhD dissertation, University of Melbourne.
National College English Testing Committee. (1999). *大学英语四、六级考试口语考试大纲及样题 [College English Test-Spoken English Test Syllabus and Sample Test Papers]*. Shanghai: Shangai Foreign Language Education Press.
North, B. (2000). *The development of a common reference scale of language proficiency*. New York: Peter Lang.
O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*(2), 217–237.
O'Loughlin, K. (1997). *The comparability of direct and semi-direct speaking tests: A case study*. Unpublished PhD dissertation, University of Melbourne.
Sacks, H., Schegloff, E. A. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735.
Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(3), 209–231.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, *11*(2), 99–123.
Stenstrom, A. B. (1994). *An introduction to spoken interaction*. London: Longman.
Tarone, E. (1980). Communication strategies, foreigner talk and repair in interlanguage. *Language Learning*, *30*(2), 417–431.
Willingham, W. W. & Cole, N. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170.

5

Face-to-face Interaction in a Speaking Test: A Corpus-Based Study of Chinese Learners' Basic Spoken Vocabulary

Shasha Xu

5.1 Introduction

Natural language use constitutes the best source of linguistic evidence (Sinclair & Carter, 2004). The availability of large corpora changes considerably the possibility of research on authentic language data (Adolphs & Carter, 2013; Hunston, 2002). Recently, analyses of learner corpora against native speaker corpora provide insights for vocabulary teaching and learning. Corpus linguistics studies have included spoken language data for the purpose of developing pedagogical materials (Biber, 2006; Campoy-Cubillo, Bellés-Fortuño, & Gea-Valor, 2010; Granger, 2003). By comparing a learner corpus with a native speaker corpus, it is 'possible to identify instances of learners' underuse or overuse of spoken vocabulary, as well as to investigate how far, and in what ways, learners deviate from NS norms' (Shirato & Stapleton, 2007, p. 394).

The College Learners' Spoken English Corpus in China (COLSEC) is the first spoken English corpus of non-English major university students in China (Yang & Wei, 2005). It is composed of the transcriptions of the College English Test-Spoken English Test (CET-SET) from 2000 to 2004, with a total of 723,299 tokens. The CET-SET measures Chinese university students' oral English ability and aims to promote test takers' communicative competence. Test takers' performance is evaluated based on three criteria, that is, accuracy and range of vocabulary, percentage of contribution and discourse management, flexibility and appropriateness (Jin, 2009; Zheng & Cheng, 2008). During the CET-SET, three or four test takers form a mini-group. Each test taker is required to conduct a conversation, make an individual presentation, and interact with group members on a controversial topic. The COLSEC is regarded

as a representative source that can provide 'basic data for studies of Chinese EFL learners' spoken English features' (Wei, 2004, p. 140).

With a view to comparing basic spoken vocabulary used in face-to-face interaction by Chinese EFL learners and English native speakers, the present study compared the COLSEC with the British National Corpus (BNC), especially the broadcast discussion and conversation parts of the spoken BNC. The present study analyzed and discussed the categories, usages and functions of high-frequency single words and multi-word clusters retrieved from the two spoken corpora. The results identified Chinese EFL learners' distinguishing patterns of basic spoken vocabulary from native language data and revealed some deficiencies in their communicative competence.

5.2 Literature review

Over the last two decades, learner corpora have become a rich resource for research on vocabulary teaching and learning (Granger, 2003). The comparison of a learner corpus with a native corpus makes it possible to identify distinguishing patterns of use from native language data, including patterns of under-, over-, and misuse in learner lexis, lexico-grammar, and discourse (Hunston, 2002). The last few years have witnessed a marked increase in studies of spoken learner corpora that analyze learners' use of vocabulary in learner speech (Aijmer, 2009; Crossley & Salsbury, 2011; De Cock, 2007, 2011; Götz & Schilk, 2011; Shirato & Stapleton, 2007). These researches on learner corpora of English as a foreign language have highlighted the pragmatic properties of learner lexis and investigated their role in speech (Paquot & Granger, 2012). Among these, Shirato and Stapleton compared the spoken British National Corpus with a small learner corpus (43,651 orthographic words) in Japan and found that Japanese learners of English differed markedly in many spoken language features such as 'discourse markers, some interactive words and terms for marking vagueness and hedges' (2007, p. 394). Spoken learner corpora consisting of texts produced by Chinese EFL learners have also attracted much scholarly attention in the last decade. Based on the subcomponents of the Louvain International Database of Spoken English Interlanguage (LINSEI), He and Xu (2003) investigated the types, ranges, frequency and interactional functions of discourse markers produced by Chinese advanced EFL learners and highlighted their functions in developing speaking fluency. In addition, many studies made use of the COLSEC and investigated features of Chinese learners' spoken English (e.g., Wei, 2004; Xu & Xu, 2007),

including features of lexical chunks, discourse patterns and pragmatic strategies used in interaction. It was found that Chinese learners' use of floor-claiming utterances, reactive tokens, discourse management chunks and conversation fillers differed markedly from that of English native speakers.

The development of spoken language corpora provides great opportunities for the analyses of vocabulary in spoken interaction (Aijmer & Stenström, 2005). The notion of vocabulary has been 'expanded beyond single words' (James & Edward, 2004, p. 242). In addition to single words, multi-word clusters 'make up a good percentage of English core vocabulary' (Schmitt, 2000, p. 224). Based on frequency of occurrences, McCarthy (1999) listed a basic spoken vocabulary (the top 2000 single words) in the Cambridge and Nottingham Corpus of Discourse in English (the CANCODE spoken corpus), which consisted of five million words of conversational English from Britain and Ireland. McCarthy (1999) then divided single words into nine broad categories: *modal items, delexical verbs, interactive words, discourse markers, basic nouns, general deictic items, basic adjectives, basic adverbs* and *basic verbs*. McCarthy commented that the types of basic spoken vocabulary 'compose interaction from a lexical viewpoint' (1999, p. 236). In terms of multi-word clusters, McCarthy and Carter (2002) investigated multi-word clusters retrieved from the CANCODE spoken corpus and identified their functions in interaction:

1. Discourse marking function, such as *you know*, *I mean*, and *and then*. These clusters are used as a topic launcher and signal a transition in a conversation.
2. Vagueness and approximation function, such as *a couple of* and *and something like that*. These clusters exhibit pragmatic integrity and play central interactive roles. They make statements less assertive and help the conversation go smoothly.
3. Face and politeness function, such as *do you think* and *what do you think*. These clusters indicate the speaker's politeness and mark the speaker' efforts to save face for the receiver.

In order to get a systematic comparison between Chinese English language learners and English native speakers, the present study adopted the framework based on McCarthy's (1999) division of single words and McCarthy and Carter's (2002) classifications of word clusters, scrutinizing single words and multi-word clusters respectively. It aimed to discover specific features of basic spoken vocabulary of Chinese learners

in face-to-face interaction, by comparing basic spoken vocabulary in the COLSEC with the BNC. More specifically, the present study aimed to address the following research question: In comparison with a native speaker corpus, what lexical items are underused or overused in face-to-face interaction in the COLSEC?

5.3 Research methods

To make the native speaker corpus as the control data comparable to the learner data in the COLSEC corpus, the present study selected the broadcast discussion and conversation component of the spoken BNC as the reference corpus (hereinafter referred to as the BNC D&C), while other genres such as speeches, lectures and broadcast news in the spoken BNC were excluded. As mentioned earlier, the COLSEC was composed of conversations and context/topic-governed discussions (Wei, Li, & Pu, 2007). The two specific parts of the BNC spoken corpus were chosen, because face-to-face conversations and context-governed discussions in the BNC resembled the genres covered in the COLSEC. With the aid of the BNC Indexer (Lee, 2001), the BNC D&C was selected and composed of 207 texts, totaling 4,972,408 tokens. Of these, the conversation part was composed of 153 texts with about 4.21 million tokens, while the broadcast discussion was composed of 54 texts with about 0.76 million tokens.

The corpus exploration software package, Wordsmith Tools Version 5.0 (Scott, 2004), was used to retrieve recurring single words and multi-word clusters and list their frequencies. Taking into account the sizes of the two corpora in this study, the frequencies were normalized at a rate per 100,000 words. SPSS Statistics 20 was used (Larson-Hall, 2010) to check whether there were any significant differences of normalized frequencies between the COLSEC and the BNC D&C. Given that the COLSEC contains data produced by both Chinese university students and examiners, a step was taken to make the COLSEC and the BNC D&C comparable: Wordsmith was set to use 'part of file' to exclude examiners' utterances indicated by <interlocutor> in the COLSEC, leaving behind the utterances produced by test takers in the COLSEC, with a total of 615,512 tokens. The procedure for extracting the words and clusters was to generate rank-order frequency lists of single words and two-, three-, four-, five-, and six-word clusters for the two corpora.

With respect to authenticity of face-to-face language data, it should be admitted that the COLSEC was not parallel to the BNC D&C. Instead of reflecting a large body of natural discourse, the topics for conversations

and group discussions in the COLSEC were carefully selected and controlled. However, it was still of great value to compare test performance data with native speaker norm as long as the methodological procedure was carefully designed. The present study confined itself to focusing on the lexical items least affected by the controlled topics in the COLSEC and lexical items specific to face-to-face interaction. In terms of the analyses of single words, the present study confined itself to two specific categories in McCarthy's (1999) division: interactive words and discourse markers. Unlike other categories of single words, interactive words and discourse markers in the COLSEC were minimally influenced by the controlled topics, since they mainly dealt with the realm of discourse and interpersonal communication. In terms of the selection of multi-word clusters, a number of high-frequency multi-word clusters had neither semantic nor pragmatic integrity, such as *as is* and *my, in the*. Moreover, due to the specific testing condition, many of the high frequency clusters in the COLSEC were artifacts, such as *my number is, my major is,* etc. Therefore, clusters displaying semantic unity and pragmatic integrity were chosen. Examples include *I think, there are,* and *a lot of*. The retrieval of frequency of single words and clusters was only the preliminary step. Further inferential analyses were carried out to find the meaning and significance in the frequency of occurrences.

5.4 Findings and discussions

This study systematically compared the occurrence rate of top words and clusters in the frequency list in the COLSEC with those in BNC D&C. The sequential explanatory analyses were adopted in that some features were scrutinized on the quantitative analyses of the whole data while others were illustrated with examples and excerpts.

5.4.1 Analyses of single words

With regard to interactive words, there are a number of items in the core word list that represent speakers' stance, such as *whatever, slightly* and *basically*. They are absolutely central to communicative wellbeing, creating and maintaining appropriate social relations (McCarthy, 1999). Table 5.1 presents and examines nine interactive words listed by McCarthy (1999): *just, really, quite, actually, whatever, pretty, basically, slightly,* and *literally.*

Table 5.1 shows that all the nine interactive words occurred much less frequently in the COLSEC, as compared with the BNC D&C. In particular, *actually* occurred almost seven times more frequently in the BNC

Table 5.1 Interactive words in the COLSEC and the BNC D&C

Word	Corpus	Raw frequency	Normalized frequency	Likelihood ratio
just	COLSEC	2418	392.8	66.376**
	BNC D&C	21893	466.1	
really	COLSEC	548	89	576.046**
	BNC D&C	10425	222	
quite	COLSEC	292	47.4	198.097**
	BNC D&C	4751	101.2	
actually	COLSEC	101	16.4	568.710**
	BNC D&C	4483	95.4	
whatever	COLSEC	32	5.2	169.744**
	BNC D&C	1368	29.1	
pretty	COLSEC	28	4.5	71.599**
	BNC D&C	792	16.9	
basically	COLSEC	9	1.5	63.513**
	BNC D&C	462	9.8	
slightly	COLSEC	1	0.2	47.690**
	BNC D&C	228	4.9	
literally	COLSEC	0	0.0	28.570**
	BNC D&C	116	2.5	

Note: Asterisk ** indicates a significant difference at the level of 0.001.

D&C than in the COLSEC. The most frequent cluster in the concordance of *actually*, in both the COLSEC and the BNC D&C, was '*Actually, I…*' However, the usage of *actually* was largely confined to '*Actually, I…*' in the COLSEC while concordances of *actually* in the BNC D&C showed much more diversified usages. For example, there was almost no collocation of *you actually* in the COLSEC while the cluster *you actually* ranked second-highest for two-word clusters in the BNC D&C, with its related clusters *do you actually, have you actually, you actually have,* etc.

With regard to discourse markers, McCarthy listed four single words occurring in the top 2000 in CANCODE spoken corpus: *right, well, so* and *anyway* (1999, p. 242). Table 5.2 lists the frequencies of *right, well, so* and *anyway* in the COLSEC and the BNC D&C.

Unlike the underuse of all nine interactive words in the COLSEC discussed above, Chinese learners tended to use conjunction *so* more frequently than English native speakers, which was also evidenced by another two conjunctions in the COLSEC: *and* and *because*. A possible explanation was that test takers who took part in the group discussions of the CET-SET wanted to emphasize or re-iterate the logics by using

Table 5.2 Discourse markers in the COLSEC and the BNC D&C

Word	Corpus	Raw frequency	Normalized frequency	Likelihood ratio
right	COLSEC	1547	251.3	6887.561**
	BNC D&C	59741	1271.9	
well	COLSEC	1167	189.6	4167.344**
	BNC D&C	39605	843.2	
so	COLSEC	5451	885.6	1422.616**
	BNC D&C	22631	481.8	
anyway	COLSEC	518	84.2	7079.862**
	BNC D&C	42611	907.2	

Note: Asterisk ** indicates a significant difference at the level of 0.001.

more conjunctions. On the other hand, the remaining three discourse markers were used more frequently in English native speakers' utterances: *anyway* occurred almost 11 times more frequently in the BNC than in the COLSEC; *right* five times, *well* four times. Shirato and Stapleton found that Japanese EFL learners 'used more backchannels, such as *mm*, *uhh, ahh, mhm* or *eeto* instead of *well*' (2007, pp. 403–404). In the present study, it was also found that Chinese EFL learners used *mm* much more frequently in the COLSEC and its usage deviated remarkably from the native speaker norm. Besides as a turn-initiator, *mm* in the COLSEC mainly functioned as a sign of hesitation or a pause in face-to-face interaction.

5.4.2 Analyses of multi-word clusters

In line with the classifications of functions of multi-word clusters by McCarthy & Carter (2002), multi-word clusters functioning as 'discourse markers', 'vagueness and approximation', and 'face and politeness' in both corpora were examined.

Discourse marking function was the most distinguished function in the cluster aspect of the COLSEC. Table 5.3 presents the multi-word clusters functioning as discourse markers in the COLSEC, as compared with the BNC D&C.

In terms of multi-word cluster functioning as discourse markers, Chinese EFL learners overused certain lexical items (such as *I think* and *in my opinion*). In particular, the overwhelming *I think* in the COLSEC mainly encoded discourse marking function. As a topic launcher, *I think* co-occurred frequently with *and, so, because* in the COLSEC 980, 1070 and 230 times respectively. The discourse marker *I think* followed the

Table 5.3 Multi-word clusters of discourse marking function in the COLSEC and the BNC D&C

Word cluster	Corpus	Raw frequency	Normalized frequency	Likelihood ratio
I think	COLSEC	9585	1557.2	14487.993**
	BNC D&C	12045	256.4	
you know	COLSEC	1015	164.9	742.739**
	BNC D&C	16948	360.8	
of course	COLSEC	320	52	13.171**
	BNC D&C	1950	41.5	
in my opinion	COLSEC	275	44.7	1059.851**
	BNC D&C	17	0.4	
as far as I am/I'm concerned	COLSEC	41	6.7	80.473**
	BNC D&C	35	0.7	
as far as I know	COLSEC	16	2.6	14.142**
	BNC D&C	35	0.7	

Note: Asterisk ** indicates a significant difference at the level of 0.001.

question '*what's your opinion*' directly 43 times in the COLSEC, marking the turn-taking order of the conversation. Clearly, *I think* was the most frequent item in the COLSEC, as evidenced by the result that *I think* ranked the top (with a normalized frequency of 1557.2) in the two-word cluster list. Only ten words in the single-word frequency list in the COLSEC occurred more frequently than *I think* in the two-word cluster list. Therefore, *I think* was a prominent feature of Chinese learners' spoken English. One possible reason contributing to its high frequency was that the questions like '*what's your opinion?*' were repeated by a large number of test takers in the learner corpus. Nevertheless, non-native speakers seemed to neglect the large varieties of discourse markers they could choose from (Schiffrin, 1987). In addition, the clusters such as *in my opinion, as far as I'm concerned, as far as I know* were commonly-used topic launchers among Chinese EFL learners. As can be seen from the comparison of normalized frequencies in Table 5.3, Chinese EFL learners tended to use these clusters more frequently than English native speakers, a difference that was statistically significant at the level of $p < 0.001$.

In the BNC D&C, on the other hand, the clusters functioning as discourse markers varied greatly. The high-frequency two-word clusters in the BNC D&C were listed as follows: *you know, I think, I mean, I thought, you see, of course,* and *I see*. These two-word clusters and the other clusters in the BNC D&C: *as you know, you know what I mean,* and *do you know what I mean* not only checked the common knowledge between the speakers, but also functioned as the topic launcher.

Table 5.4 shows the common clusters functioning as expressions of vagueness and approximation in the COLSEC, as compared with the BNC D&C.

The following extracts illustrate the vagueness function of *to some extent* in the COLSEC and the BNC D&C:

COLSEC- [Discussing the role of examination in school education]
<Interlocutor> *Do you think examinations can motivate students to study hard and efficiently?*
<sp1> *Yes, **to some extent**, examinations are very important because if we live or study without examinations we don't have spirit to continue our study.*
BNC- [conversation (face-to-face)]
How long does a game take, as a general rule David?
*Ah, maybe it's about three hours. I think that, **to some extent**, depending on the standard of the players.*

Chinese EFL learners used a set of vague items to express vagueness and approximation and they tended to overuse *and so on, to some extent* and *etc.* as compared with English native speakers, a difference that was statistically significant at the level of $p<0.001$. Chinese EFL learners came to realize the importance of vague language, for 'the absence of vagueness in the conversation can make utterances blunt and pedantic' (McCarthy & Carter, 2002, p. 22). Nevertheless, they seemed to be constrained by this very limited set of items. Few of the seven items listed by McCarthy and Carter (2002): *a couple of, and things like that, or something like that, that sort of thing, this that and the other all the rest of it, all this/that sort of thing* occurred in the COLSEC,

Table 5.4 Clusters of vagueness and approximation function in the COLSEC and the BNC D&C

Word cluster	Corpus	Raw frequency	Normalized frequency	Likelihood ratio
and so on	COLSEC	129	21	116.150**
	BNC D&C	278	5.9	
to some extent	COLSEC	38	6.2	76.497**
	BNC D&C	31	0.7	
etc.	COLSEC	20	3.2	21.456**
	BNC D&C	37	0.8	

Note: Asterisk ** indicates a significant difference at the level of 0.001.

nor did the 56 variants listed of vagueness by Aijmer (2002) occur in the COLSEC.

In terms of clusters of face and politeness function, Table 5.5 presents the five items listed by McCarthy and Carter (2002). As indicated in Table 5.5, *do you think* had similar normalized frequency in the COLSEC and the BNC D&C. The difference of *what do you think* in the two corpora however showed statistical significance (p <0.05), while the remaining three items *do you know, I don't know if/whether,* and *I was going to say* appeared much more frequently in the BNC D&C (p <0.001). Moreover, two distinctive items emerged from the COLSEC: *I agree with you* and *I cannot agree with you more*. *I agree with you* (with a normalized frequency of 27.8 in the COLSEC and 0.5 in the BNC D&C), in most cases, was used to mark the speaker's initiation of his or her 'own talk' and indicate the speaker's understanding of the conversational topic. The following extract illustrates the cluster of *I agree with you* of this function.

COLSEC- [Discussing the role of examination in school education]
<*sp1*> *So I do believe that examination is important, but it's not everything. What about you?*
<*sp2*> ***I agree with you.*** *And I think study is the most important thing, though examination is just a tool to check our ability. But study, as a student, study is the first thing. I also spend time enjoying myself, just do some exercises or go out with my friends.*

Table 5.5 Clusters of face and politeness function in the COLSEC and the BNC D&C

Word cluster	Corpus	Raw Frequency	Normalized frequency	Likelihood ratio
do you think	COLSEC	164	26.6	2.364
	BNC D&C	1098	23.4	(p=0.124)
what do you think	COLSEC	53	8.6	4.239*
	BNC D&C	294	6.3	
do you know	COLSEC	17	2.8	169.646**
	BNC D&C	1107	23.6	
I don't know if/whether	COLSEC	5	0.8	112.32**
	BNC D&C	604	12.9	
I was going to say	COLSEC	0	0.0	12.314**
	BNC D&C	50	1.1	

Note: Asterisk * indicates a significant difference at the level of 0.05.
Asterisk ** indicates a significant difference at the level of 0.001.

While in other cases, even if the speaker didn't agree with the other speakers, he/she pretended to agree with the others. In order to keep a smooth and polite progression of the talk, test takers in the CET-SET first used *I agree with you* to make the statement less assertive, and then went on to give their different opinions. The following extract illustrates how test takers try to protect the face of group members in the CET-SET:

COLSEC- [Discussing which is more important for college students, knowledge or experience]

<*sp1*> *as we know, as knowledge is changing is changing quickly, and so we just er need more knowledge to fulfill our mind. Er that will be beneficial our future career.*

<*sp2*> *Yes,* ***I agree with you****. But I think, for me, I'm a student. The experience is the most important to me, so I think going- going out is my best way to benefit from the college life.*

There were nine instances of '*I agree with you, but…*' in the COLSEC, which suggested that the face and politeness function of *I agree with you* in Chinese learners' spoken English. Table 5.6 presents the clusters functioning direct denial or disagreement in both corpora.

Although those items only took up a very small portion of multi-word clusters in the COLSEC, it was worthwhile to scrutinize this unique feature, since *I don't think so/no I don't think so* occurred four times as frequently in the COLSEC as in the BNC D&C, and nine times as in the case of *I don't agree with you*. Chinese learners seemed to use more intrusive interruption compared to English native speakers in turn taking,

Table 5.6 Multi-word clusters marking direct disagreement in the COLSEC and the BNC D&C

Word cluster	Corpus	Raw frequency	Normalized frequency	Likelihood ratio
I don't think so	COLSEC	111	18	93.554**
	BNC D&C	252	5.4	
no I don't think so	COLSEC	34	5.5	38.314**
	BNC D&C	60	1.3	
I don't agree with you	COLSEC	22	3.6	42.896**
	BNC D&C	19	0.4	

Note: Asterisk ** indicates a significant difference at the level of 0.001.

indicating relative deficiency of face and politeness function in the COLSEC. The specific testing condition in the CET-SET probably contributes to the relatively high frequency of direct denial or disagreement. Another possible reason is politeness and discourse marking is a higher level of language production, which requires a level of automaticity that allows learners to process their utterances in real time. Therefore, Chinese learners with limited linguistic competence may focus most of their effort on the production of words and organization of sentences, failing to control their appropriate language use in face-to-face interaction.

In summary, the present study investigated the basic spoken vocabulary in face-to-face interaction in the CET-SET. By comparing high-frequency single words and word clusters in the COLSEC and the BNC D&C, the analyses discussed the categories, usages and functions of the core spoken vocabulary in face-to-face interaction. It was found that Chinese EFL learners tended to underuse lexical items in the following categories: interactive words, interjection in discourse markers, and clusters of vagueness and approximation function. On the other hand, Chinese EFL learners tended to overuse conjunction and hesitation in discourse markers. The analyses also revealed that Chinese learners were confined to a limited set of multi-word clusters and used them repeatedly, neglecting the diverse usages by English native speakers as indicated in the BNC D&C.

5.4.3 Discussions

The results of the present study indicated considerable differences in both single words and multi-word clusters between Chinese EFL learners and English native speakers in face-to-face interaction. As for the single words, one of the most remarkable findings in the present study was the relative deficiency in Chinese learners' use of interactive words, however, they tended to overuse conjunctions (e.g. *and, so, because*) and the interjection *mm* compared to English native speakers, in contrast to their underuse of other interjections and discourse markers. The conventional view of the words in a language is 'that they either have lexical meaning or are confined to syntactic functions in the sentence' (Sinclair & Renouf, 1990, p. 154). Hence usages which are of pragmatic and communicative importance are often overlooked by language learners. These interactive words and discourse markers are in frequent daily use, but largely at a subliminal level among native speakers. Because of their prominent status in face-to-face interaction, these words of semantically empty function deserve particular attention in foreign language teaching and learning. Therefore, the relatively

infrequent use of interactive words and discourse markers in the COLSEC has some pedagogical implications. I suggest that interactive words and discourse markers that enhance communicative competence be introduced at an early stage of language learning, especially in the spoken English class. The semantically empty words should be taught in classroom dialogues and group discussions, only then can learners 'discover, understand, and begin to internalize the expressions of that language' (Carter, 1998, p. 51).

As for the multi-word clusters functioning as discourse markers, on the one hand, Chinese learners overused a set of topic launchers such as *I think, in my opinion, as far as I'm concerned,* and *as far as I know* compared to native speakers of English; on the other hand, they rarely used a large variety of topic launchers which they may choose from. The limited repertoire of clusters functioning as discourse marker by Chinese EFL learners suggest that clusters specific to face-to-face interaction need to be included in the syllabus. Therefore, the addition of high-frequency topic launchers from the native spoken corpus to the lexical syllabus such as *you know, I mean, you see, I see, as you know,* and *you know what I mean* would seem to help considerably enhance naturalness of utterance produced by Chinese EFL learners in speaking tests. Corpus-driven syllabuses and teaching materials, to a large extent, maintain a great amount of authentic language features. In this regard, the inclusion of authentic language features from native speakers' corpora can provide a good opportunity for EFL learners to observe and interact in real discourse (Flowerdew, 2009; Gaviola & Aston, 2001).

Chinese EFL learners used a set of lexical items to express vagueness and approximation (e.g. *to some extent, and so on, etc.*). However, they seemed to be constrained by this limited set, ignoring the diverse range of vague language commonly used by English native speakers such as *a couple of* and *and something like that.* Shirato and Stapleton commented that 'among all of the strings used as interactive units, those encoding vagueness and approximation displayed the largest differences' (2007, p. 405) between Japanese EFL learners and English native speakers. The absence of a large variety of vague language in the COLSEC showed that Chinese learners tended to underuse vague terms in face-to-face interaction as well. How well one can use vague language is a manifestation of one's pragmatic competence (Fraser, 2010). However, language course books 'do not exhibit many examples of vague language, even though it is always pragmatically highly significant' (Carter, 1998, p. 45). Therefore, language teaching should raise learners' awareness of the importance of vague language in communication as early as possible (Zhang, 2011).

Besides, learners should be exposed to a large variety of vague language and then become capable of the strategic use of vague language.

5.5 Conclusion

Communicative competence has become an important part in foreign language teaching, with the emphasis on interaction. From a communicative point of view, the present study discussed the categories, usages and functions of basic spoken vocabulary of Chinese EFL learners as indicated in the COLSEC. The overall findings showed that Chinese learners tended to underuse lexical items representing interactive functions in the following categories: interactive words, interjection in discourse markers, and clusters of vagueness and approximation function. On the other hand, Chinese learners tended to overuse conjunction and hesitation in discourse markers. The findings also showed that Chinese learners were confined to a limited set of multi-word clusters and used them repeatedly.

The present study suggests pedagogical implications for vocabulary instruction in Chinese EFL context. It is suggested that interactive words, discourse markers and clusters of politeness and vagueness functions that enhance communicative competence should be introduced at an early stage of language learning. Moreover, material designers may use more authentic and natural language by adopting corpus-driven syllabuses and teaching materials.

The limitation of the present study perhaps lies in the selection of target corpus. First and foremost, all materials in the COLSEC were collected from oral tests. In terms of authenticity of informal face-to-face language data, the COLSEC was not parallel to the BNC D&C. The performances of the test takers were far different from the speakers' spontaneous talk or communication, though it could be argued that spoken tests are one of the major contexts for non-English major Chinese learners to speak English. As the most significant development in assessing Chinese university students' speaking ability, the CET-SET set a specific goal to encourage Chinese EFL learners to 'participate more actively in interactive communication' (Jin, 2009, p. 52). Meanwhile, test takers' skills on discourse management and interaction with group members are rewarded according to the scoring criteria. Therefore, the present study tries its best to offset the drawbacks by confining itself to focusing on the categories least affected by the controlled topics of the target corpus. Therefore, if a corpus of Chinese learners' real, naturally-occurring spoken English is established in the future, its comparison with that of

a native speaker corpus would be an important part of the continuing analyses of this fascinating subject.

References

Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. New York: Routledge.

Aijmer, K. (2002). *English discourse particles: Evidence from a corpus*. Amsterdam, Netherlands: John Benjamins.

Aijmer, K. (2009). A corpus study of 'I don't know' and 'dunno' in learner spoken English. In A. H. Jucker, D. Schreier & M. Hundt (Eds.), *Corpora: Pragmatics and discourse* (pp. 151–166). Amsterdam, Netherlands: Rodopi.

Aijmer, K., & Stenström, A. (2005). Approaches to spoken interaction. *Journal of Pragmatics, 37*(11), 1743–1751.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Philadelphia, PA: John Benjamins.

Campoy-Cubillo, M. C., Bellés-Fortuño, B., & Gea-Valor, M. L. (2010). *Corpus-based approaches to English language teaching*. New York: Continuum.

Carter, R. (1998). Orders of reality: CANCODE, communication, and culture. *ELT Journal, 52*(1), 43–56.

Crossley, S., & Salsbury, T. L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Language Teaching, 49*(1), 1–26.

De Cock, S. (2007). Routinized building blocks in native speaker and learner speech: Clausal sequences in the spotlight. In M. C. Campoy & M. J. Luzón (Eds.), *Spoken corpora in applied linguistics* (pp. 217–233). Bern, Switzerland: Peter Lang.

De Cock, S. (2011). Preferred patterns of use of positive and negative evaluative adjectives in native and learner speech: An ELT perspective. In A. Frankenberg-Garcia, L. Flowerdew & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 198–212). London: Continuum.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics, 14*(3), 393–417.

Fraser, B. (2010). Pragmatic competence: The case of hedging. In G. Kaltenböck, W. Mihatsch & S. Schneider (Eds.), *New approaches to hedging* (pp. 15–34). Bingley, UK: Emerald.

Gaviola, L., & Aston, G. (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal, 55*(3), 238–246.

Götz, S., & Schilk, M. (2011). Formulaic sequences in spoken ENL, ESL, and EFL. In M. Hundt & J. Mukherjee (Eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 79–100). Amsterdam, Netherlands: John Benjamins.

Granger, S. (2003). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538–546.

He, A., & Xu, M. [何安平, & 徐曼菲]. (2003). Small words in Chinese EFL learners' spoken English [中国大学生英语口语 Small Words 的研究]. *Foreign Language Teaching and Research [外语教学与研究], 35*(6), 446–452.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

James, F. B., & Edward, J. K. (2004). *Vocabulary instruction: Research to practice*. New York: Guilford Press.

Jin, Y. (2009). The national college English testing committee of China. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 44–59). New York: Routledge.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. London: Routledge.

Lee, D. Y. W. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics, 29*(3), 250–278.

McCarthy, M. (1999). What constitutes a basic vocabulary for spoken communication? *SELL: Studies in English Language and Linguistics, 1*, 233–249.

McCarthy, M., & Carter, R. (2002). This, that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *Teanga, 21*, 30–52.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130–149.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Scott, M. (2004). *Wordsmith Tools*. Oxford: Oxford University Press.

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research, 11*(4), 393–412.

Sinclair, J. M., & Carter, R. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Sinclair, J. M., & Renouf, A. (1990). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140–158). Harlow, UK: Longman.

Wei, N. [卫乃兴]. (2004). A preliminary study of the characteristics of Chinese learners' spoken English [中国学习者英语口语语料库初始研究]. *Modern Foreign Languages [现代外语], 27*(2), 140–149.

Wei, N., Li, W., & Pu, J. [卫乃兴, 李文中, & 濮建忠]. (2007). Design principles and annotation methods of the COLSEC corpus [COLSEC语料库的设计原则与标注方法]. *Contemporary Linguistics [当代语言学], 9*(3), 235–246.

Xu, J., & Xu, Z. [许家金, & 许宗瑞]. (2007). Discourse management chunks in Chinese college learners' English speech: A spoken corpus-based study [中国大学生英语口语中的互动话语词块研究]. *Foreign Language Teaching and Research [外语教学与研究], 39*(6), 437–443.

Yang, H., & Wei, N. [杨惠中, & 卫乃兴]. (2005). *Construction and data analysis of a Chinese learner spoken English corpus* [中国学习者英语口语语料库建设与研究]. Shanghai, China: Shanghai Foreign Language Education Press.

Zhang, G. (2011). Elasticity of vague language. *Intercultural Pragmatics, 8*(4), 571–599.

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

6

Features of Formulaic Sequences Used by Chinese EFL Learners in Performing a Story Retelling Assessment Task

Lei Wang and Chan Chen

6.1 Introduction

In the study of second language acquisition (SLA), many researchers have focused their attention on the ways that learners can best acquire the target language (TL). In recent years increasing attention has been paid to the mastery of formulaic sequences (FSs) or chunks in SLA (Lewis, 1993; Nattinger and DeCarrico, 1992; Willis, 1990; Wray, 2000). FSs are those ready-made lexical sequences that can be used without breaking the components into individual parts. Such patterns of language are usually perceived, learned and used as meaningful sequences that are processed as a whole, resulting in reduced learning burden and increased fluency.

There exists a growing literature on formulaic language, its role in communication and its contribution to fluency in second language learning contexts (Schmitt, 2004; Wray, 2002). This gives much evidence from large corpus-based samples that illustrate the extent to which such forms of language play a significant role in the output of language users, especially in spoken language where processing constraints indicate a greater reliance on "fixed" as opposed to creative expressions. However, empirical studies are much needed to indicate the fact in the learning process. Social and interactive contexts decide when and where speakers can rely more on the use of formulaic language or on linguistic creativity while processing language. There are contexts in which they can be more creative by breaking rules and contexts in which creativity can be based on more creative uses of language.

In *Formulaic Language and the Lexicon* (2002), Alison Wray argues that formulaic knowledge has implications for language use at all levels. She

says that mature native speakers of a language apparently produce and interpret "ready made surface structures" (p. 13) for nearly all communicative functions, retrieving sometimes quite lengthy strings from memory as single lexical units, while using our "live grammar and lexicon" (p. 33) sparingly to mainly stitch the pre-casts together.

The present study explores features of formulaic sequences used by Chinese English learners in a fairly-controlled story retelling test. Story retelling, old as it is, is still favoured by EFL practitioners as an important method for assessing learners spoken English. It is hypothesised that when learners are given the authentic language input, and required to retell, they may try to memorise as many as possible of those prefabricated formulaic sequences that they hear from the original text and reproduce some of them in their retelling. At the same time, the learners' previous linguistic knowledge will contribute to constructing the new discourse of story retelling. The research questions of the present study are: To what extent do learners try to memorise the formulaic sequences and use them in story retelling? What are the features of FSs in performing a story retelling assessment task?

6.2 Studies on formulaic sequences and story retelling as an assessment method

6.2.1 Formulaic sequences and language learners

Formulaic language has in recent years become widely recognised as a crucial aspect of second language competence. People started to observe unexpected levels of fixedness in language in the mid-nineteenth century when John Hughlings Jackson took an interest in the ability of aphasic patients who could fluently utter rhymes, prayers, and routine greetings. Some phrases and expressions have become conventionalised as more or less unanalysed composites of form and function. These multi-word chunks have been called various names, including lexical phrases, lexical chunks, formulaic language, ready-made (complex) units, formulaic sequences, etc. (Bolinger, 1976; Coulmas, 1979; Cowie, 1998; Ellis, 1996; Lewis, 1993; Nattinger & Decarrico, 1992; Schmitt, 2004; Widdowson, 1990; Wray, 1999; Yorio, 1980). Some empirical studies have shown that FSs play a crucial role in L1 and L2 child language acquisition (Fillmore, 1976; Hatch, 1978; Pawley & Syder, 1983; Peters, 1983; Vihman, 1982; Weinert, 1995).

It is becoming increasingly apparent that language is largely formulaic in nature, and that the competent use of formulaic sequences is an important part of fluent and natural language use (Cowie, 1998, Nattinger and DeCarrico, 1992, Pawley & Syder, 1983, Schmitt, 2004,

Wray, 2002). It has also been suggested that FS plays an important role in language acquisition. Following the early lead of child language researchers such as Peters (1983), "usage-based" models of language have recently been developed which see first language learning as a process in which rote-learned, formulaic chunks are gradually subject to analysis and abstraction (Tomasello, 2003). Ellis (2003) has proposed that a similar model might be applied to second language acquisition.

Formulaic sequences are believed to play an important role in language production and fluency. Studies comparing formulaic language between written and spoken corpora suggest that formulas are more frequent in spoken language (e.g., Biber, Johansson, Leech, Conrad, & Finegan, 1999; Brazil, 1995; Leech, 2000). Oral speech is constructed in real time, which imposes greater working memory demands than writing, hence the greater need to rely on formulas. It is easier for us to look something up from long-term memory than to compute it in speech (Kuiper, 1996). The research on conversational talk (Pawley & Syder, 1983: 191) shows that "fluent and idiomatic control of a language rests to a considerable extent on knowledge of a body of 'sentence stems' which are 'institutionalized' or 'lexicalized'." The appropriate use of FS in particular registers contributes to the native-likeness in language learning and therefore is an indicator of high language proficiency.

Psycholinguistists and language acquisition specialists are interested in what determines the use of formulaic phrases, whether they are stored as wholes in the learner's lexicon or are just one-time imitations heard by the learner. From a psycholinguistic perspective, formulaic language is generally believed to offer processing advantages. This is because such sequences can be memorised as single units and processed with greater speed and ease than the same words processed creatively by a rule-based system (Schmitt & Carter, 2004: 4–5). Access to prefabricated language, therefore, enables the user to bypass syntactic/discoursal processing requirements, thus avoiding potential overload of working memory. FSs are retrieved and processed as a whole in spoken and written language and allow learners to save processing efforts and "produce language that is phraseologically similar to that of native speakers and to produce language without undue hesitation or disfluency" (Hunston & Francis,1999: 10–11). Using formulas and memorised patterns can in return become a learning strategy to enable learners to find and apply some rules of the target language.

In line with this is a widespread acceptance in the field of second language acquisition that language instruction needs to ensure that learners develop a rich repertoire of formulaic sequences as well as a rule-based competence. As reviewed by Nesselhauf (2005),

psycholinguistic evidence indicates that the human brain is much better equipped for memorising than for processing, and the availability of a large number of prefabricated units reduces the processing effort and thus makes fluent language possible (cf. Aitchison, 1987; Partington, 1998: 20; Pawley & Syder, 1983)

6.2.2 Studies on formulaic sequences in China

The research studies on formulaic sequences in China are mostly based on Chinese EFL learner corpora of different levels. Quite a number of them focus on the relationship between learners' competence in formulaic sequences and their language proficiency. Zhang (2004) compares the use of lexical bundles among advanced learners, intermediate learners and beginners to examine the relationship between language proficiency, fluency, idiomaticity and the use of lexical bundles. The findings show that advanced learners use a significant number and variety of FSs while intermediate learners' use of FSs lacks variety and beginners tend to build sentences by grammatical rules, thus resulting in unidiomaticity. Ding and Qi (2005) indicate that the learners' ability to use formulaic language is a better predictor of oral and written English than grammatical accuracy.

Some Chinese studies used larger-size corpora and investigated the structures and functions of FSs. Wei (2007) analyses the structural and functional characteristics of FSs contained in the College Learners Spoken English Corpus (COLSEC) (totalling 723,000 running words), using Altenberg's framework (1998), that is, full clauses, clause constituents and incomplete phrases. The research reveals that learners use much more types of sequences which are closely tied with the expression of propositional meaning, but much fewer types of sequences which basically perform pragmatic functions. Such preference may result in a lack of pragmatic quality in discourse construction. Similarly, Liu & Liu (2009), based on COLEC (College Learners' English Corpus, totaling 500,000 running words), investigated the FSs in college students' writing. The findings show that the structures of FSs in students' writing are various, which is in consistence with Wei (2007). Both studies present detailed descriptions of the characteristics of lexical patterns of students' spoken or written corpora, but give inadequate explanations of students' preference and errors in their use of FSs.

6.2.3 Formulaic sequences in story retelling as assessment task

Story retelling is regarded as a post-reading or post-listening recall in which readers or listeners tell what they remember orally (Morrow,

1996). It has been proved to be an effective learning and teaching activity both in the first language acquisition and second language learning. Morrow (1985, 1986) carried out three different studies to understand instructional benefits of story retelling. In all three studies, students who had opportunities for story retelling tasks showed significant improvement in their oral language complexity, comprehension of story, and sense of story structure, because story retelling can draw the learners' attention and enhance their already-established language system. In second language acquisition or EFL environment, story retelling in the target language could offer learners a good opportunity to mimic and process the original story they have read or heard by organising and explaining the source story to others. As an important way to improve their speaking ability, it also provides data to gain insight into a reader's or a listener's comprehension process.

In the principle for oral assessments, Skehan (1996) mentions a good retelling should be accurate, covering the major idea units in the source message, and carrying all the necessary supporting details to the extent that the image created is the same as that in the original story. As a communicative task, story retelling requires a range of language skills. The present study investigates learners' use of formulaic sequences in story retelling in a test situation, to understand what is happening when learners are retelling a story. In performing a story retelling, short-term memory is used for processing language when learners try to recall what they have heard and reproduce the story as accurately as possible. Bolinger (1976) emphasises the role of memory in recognition, and argues that language itself is much more memory-based than has been generally thought. He also uses the word "prefabs" to refer to the formulaic sequences, believing those prefabs can do almost as much remembering as they do putting together. Skehan (1998) points out that language is more memory-based and idiomatic in nature and use than is often realised.

To what do learners pay attention when retelling a story? Van Pattern (1990) holds that in the mental process from language input to oral output, the learners focus more on extracting meaning and organising ideas using the target language than language form. Wang's (2004) survey confirmed that many of the students who participated in the Spoken English Test for English Majors –Band 4 (STEM4) put the best of their attentions on the content, including the keywords or phrases that are associated with the main content of the story, when they were doing the story retelling task.

6.3 Research methodology

Formulaic sequences could be of various kinds. Some are salient ones like *Go away!*, others are flexible with slots to fill in like *for __ days*. According to Schimitt (2004), "formulaic sequenences lie on a continuum of transparency/opagueness, with idioms at the obscure end, but with many sequences being quite transparent at the other end (*my point (here) is that ___*)"(p. 6). He makes a speculation based on intuitions that flexible formulaic sequences are widely used in discourse, simply because they are adaptable to a wide range of situations. The present study will focus on how flexible learners could be in using formulaic sequence. The constraints should be on the content of the story. That is, what could be used to fill out those slots cannot go beyond the semantic information from the story. Features of FS will demonstrate its function while retelling is used as assessment.

6.3.1 Data used for the study

The data used for analysis in the present study are story retellings by 42 learners in Spoken English Test for English Majors –Band 4 (STEM4). STEM4 is composed of three tasks: retelling a story, talking on a given topic and carrying out a dialogue (pair work). In the story-retelling task, students listen to the recording of a story (about 300 words) twice, and then immediately afterwards (i.e., without any preparation time) they are given three minutes to retell the story. While listening, they can take notes on a piece of paper. The retelling is recorded and later sent to the authorised institutions for grading.

The 42 learners' recordings were collected during the STEM4 in 2009. The participants were chosen randomly; their recordings were transcribed by the authors without paying any attention to the linguistic errors made by the test takers. Transcripts of the source story in a written text is presented as the source text (ST) (see Appendix 6.1) and that of the output stories by the test takers are presented as Learner Texts (LTs). Paralinguistic, pausal features and hesitation fillers are omitted. Punctuation is also designed primarily to assist reading rather than to show all the pauses. A small corpus was set up based on the transcribed learners' retellings (see Table 6.1).

6.3.2 Identification of formulaic language

Based on the definition of formulaic sequences by Wray (2002: 9): "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being

Table 6.1 Composition of the corpus used for analysis

Type	No. of texts	No. of Words
Source Text (ST)	1	377
Learners Texts (LTs)	42	12227

subject to generation or analysis by the language grammar", 40 formulaic sequences were identified in the ST and their frequencies of use in the LTs were counted with the help of the computer software AntConc 3.3.4, a freeware concordance program.[1] Just as Schmitt (2004) puts it: "modern concordancers are good at identifying contiguous sequences, but we do not yet have software which can identify flexible formulaic sequences automatically from corpora"(p. 7), the software was used as an aid to retrieve some basic formulaic sequences. With the help of the function of "N-grams" in AntCon 3.3.4, we tried 2, 3, 4 and 5 grams and finally decided on the 40 sequences for analysis for the present study. Variant sequences produced by learners will also be listed to know if they have constructed LTs by using their own syntactic knowledge.

6.3.3 An interview

Semi-structured interviews were conducted soon after the STEM4 with 28 test takers in four groups, randomly selected. The interview was conducted in both English and Chinese for better communication. Six questions (see Appendix 6.2) were asked, with the aim of identifying the reasons for test takers' use of formulaic sequences in the story-retelling task. The questions centred around: whether test takers used the formulaic sequences from the source text as a result of their memory, or read from the formulaic sequences they noted down while listening to the story, or reproduced some sequences out of their own grammatical knowledge, based on the information they got from the story. As a semi-structured interview, we used the list of questions mainly as guidelines to ensure that the same general areas of information were collected from each interviewee; but still allowed some degree of freedom and adaptability in data collection.

6.4 Results

6.4.1 Features of formulaic sequences in learner's retelling

The source text is a short story about how an owner of a small supermarket won the hearts of local people with not only the various kinds of commodities that he sells, but also his understanding, love, and trust

for the local people. It is a fairly easy story with 377 words. If it is used as listening or reading material in a test, such a passage would be very easy for this group of second year English majors. However, because the task is to retell the story immediately after listening to it twice under examination conditions, no one would expect that learners to recite it purely based on their memory.

We identified manually 40 formulaic sequences, all of which are simple phrases, from the source text. It is assumed that the learners would use some of those FSs in their story retelling. It is also assumed that the participants might have noted down some of them while listening to the story. All the 40 FSs are listed in the order of their appearance in the story and the frequencies of uses by the subjects are shown in Table 6.2.

As shown in Table 6.2, four FSs are used by more than half of the subjects. Among the high frequencies are some commonly used and simple ones: *run away* (90.48%), *(a) middle-aged woman* (73.81%), *three years ago* (73.81%), *Why don't you* (66.67%) and *a basketball match (54.76%)* respectively. The sequence *(a) middle-aged woman* enjoys a high percentage of uses because it indicates the main character of the story. *Run away* is a phrase that Chinese learners learned in their early years of English study when they studied the verb *run*. Many participants could use ___ *years ago* (73.81%), a slot sequence that requires a numeral to fill in. Vague words such as *several* and *many* were used by those participants who did not remember the exact number of years.

In contrast, those relatively difficult ones like *felt someone pat her on the shoulder* (0%), *Pointing to* (0%), *pleaded with her for* (2.38%), *at a low price* (2.38%), *with an understanding smile* (4.76%), *without knowing what to do* (4.76%), *dropped into* (7.14%), *… in her mother's eyes* (9.52%); etc. are used by less than five students. Some of them are usually structures that are different from the Chinese equivalents. For example, no one used "*felt someone pat her on the shoulder*". The low frequency in the use of the FSs in the LTs indicates that, instead of making use of those ready-made FSs, learners must have tried some other ways to express the meaning of the original story.

At the listening stage, some of the 40 FSs may have been noted down by the learners on the pieces of paper provided. However, the fact that not a single lexical phrase was used by every learner indicates that it is impossible for them to depend on the pure use of FSs to complete their story retelling. Comparing the source text with the learners' responses can help us to understand how flexible the learners are when applying their linguistic knowledge to convey of the information of the original story. Many learners were able to construct their own phrases and

Table 6.2 Use of FLs in learners' texts (LTs)

No.	Formulaic Sequences in ST	Freq.	%	No.	Formulaic Sequences in ST	Freq.	%
1	*owned a small supermarket*	16	38.10	21	*a basketball match*	23	54.76
2	*All the people nearby*	6	14.29	22	*laughed at*	17	40.48
3	*like to go shopping*	11	26.19	23	*with my bare feet*	4	9.52
4	*big chain stores*	9	21.43	24	*the money in her pocket*	3	7.14
5	*everything from toothpaste to television*	8	19.05	25	*barely enough for a meal*	5	11.90
6	*closed down*	18	42.86	26	*Seeing no hope*	10	23.81
7	was very curious	15	35.71	27	*in her mother's eyes*	4	9.52
8	*dropped into*	3	7.14	28	*without knowing what to do*	2	4.76
9	*(a) middle-aged woman*	31	73.81	29	*started to weep*	7	16.67
10	*walked out of*	1	2.38	30	*felt someone pat her on the shoulder*	0	0.00
11	*Why don't you*	28	66.67	31	*turned around*	10	23.81
12	*at low prices*	1	2.38	32	*run away*	38	90.48
13	*with a smile*	7	16.67	33	*dreamed of*	7	16.67
14	*because of a pair of sport(s) shoes*	15	35.71	34	*with an understanding smile*	2	4.76
15	*three years ago*	31	73.81	35	*Your son can't wait*	15	35.71
16	*work(ed) very hard to raise the family*	9	21.43	36	*But I can wait*	15	35.71
17	*one of her two children…*	4	9.52	37	*pay me later*	13	30.95
18	*rush(ed) in*	12	28.57	38	*was deeply touched*	11	26.19
19	*pleaded with her for*	1	2.38	39	*not only could*	2	4.76
20	*Pointing to*	0	0.00	40	*but also understanding, love and trust*	4	9.52

Note: Freq = Frequencies used by 42 Chinese EFL learners %= the percentage of users.

expressions, though sometimes the meaning would be slightly different from the source text. Below we present some samples.

Sample 1

ST: *Mr. Smith owned a small supermarket. All the people nearby like to go shopping there.*

LTs: (1) *Mrs. Smith owns a small supermarket, and everyone like shopping there.*

(2) *Mrs. Smith owns a small supermarket and people like going shopping there.*

(3) *Mr. Smith owned a small supermarket. All the nearby neighbours are shopping there.*

The FS "like to go shopping" was exactly used by 11 subjects (26.19%). In the LTs as shown above, this FS could be replaced by three possibilities, though there might be some slight differences in meaning. However, the grammatical rules guarantee the correctness of the sentences and ultimately make it successful in communication.

Sample 2

ST: *Several months ago, a few big chain stores were opened in town. They sold everything from toothpaste to televisions and prices were pretty low.*

LTs: (1) *Recently, there is a big supermarket opened in the town and it sells toothpaste, television and the price is very low.*

(2) *Several months ago, there was a chain store opened in their town and sold a large scale of commodities from toothpaste to television.*

(3) *There are several chain stores nearby, and they are not only sell toothpaste, but also television and so on.*

In the ST, the phrase "*everything from … to…*" is used to exemplify different items sold in the store. According to the LTs produced, only eight participants used the exact phrases. For the rest, we found three basic variant types. The basic meaning of the original sentence is well expressed in (1), which is quite a simple sentence. A new expression that may be deviated from the idiomatic expression is coined in (2), and in (3) another more popular chunk "*not only…, but also…*" is used, which can convey much the same meaning as the original one.

Sample 3

ST: *Suddenly she felt someone pat her on the shoulder.*

LTs: (1) *Someone/Mr. Smith pat her on the shoulder/pat on her should.*
(2) *There was a pat on her shoulder.*

The Fs "*felt someone pat on the shoulder*" is a difficult one for Chinese learners. No one in the sample used the exact phrase structure of the source text. When we search the key word "pat" in the LTs, it was found that over half of the participants (52.38%) have used the word. But the phrase structures they adopted are "*pat sb. on the shoulder*" or *"there was a pat on her shoulder*", in which the idea of "felt" is gone. Others simplified it incorrectly by saying "*pat sb. shoulder*" or simply "*pat sb.*", which may be the result of literal translation of the same expression in Chinese.

Sample 4

ST: *Seeing no hope in her mother's eye, Tommy ran away, ….*
LTs: (1) *When the boy thought there was no hope…*
(2) *Knowing that there was no hope …*
(3) *Tommy/He saw/know no hope…*

Although no participant said exactly "*Seeing no hope in her mother's eye…*", the idea of the phrase was well expressed in the LTs. About 23.81 % of the participants could say "Seeing no hope", but only 9.52% mentioned "*in her mother's eyes*". Some participants also tried to rephrase it by "*there be*" pattern which is more common in spoken language. Thus, the use of the formulaic sequence seemed more difficult for the learners.

Sample 5

ST: *She really wanted to buy her son a nice pair of sport shoes, but the money in her pocket was barely enough for a meal.*
LT: (1) *She really wanted to buy her son shoes but the money in her pocket is rarely enough for the food.*
(2) *But the woman had no money then because she had only the money to buy the meal.*
(3) *The mom really wanted to buy a nice pair of sports shoes for Tommy but the money in her bag is bare enough.*
(4) *The woman felt very sorry, but he rarely can't afford, she has little money.*

As shown above in Sample 5, instead of following the original phrase '*the money in her pocket was barely enough for a meal',* learners used different sentence and phrase structures to convey similar meanings, including

some ungrammatical ones. It is obvious that learners remembered the content of the story and tried to reproduce the idea linguistically.
While Table 6.2 demonstrates the use of FSs by the learners, the above five samples, on the other hand, indicate that although at the listening stage learners can capture some FSs, including the understanding of the meanings and functions of FSs, it is difficult in the production state to reproduce the exact same FSs. Many learners are linguistically competent enough to convey the original meaning without using the FSs.

6.4.2 Learners' comments on their use of formulaic language in story retelling assessment task

The main purpose of the focus-group interview was to understand what our learners were actually doing when completing the story retelling task. The interview was a fairly controlled one in which interviewers asked each interviewee the six questions. The subjects' answers were noted down and sorted out afterwards. The following are some of the main points we found that are relevant to the discussion of the research topic.

When Question 1 "*What do you pay attention to while listening to the story? (content, words and phrases or lexical chunks)*" was asked, most interviewees told us that they paid more attention to the content of the story, because their teachers have told them that the assessment criterion usually break the whole story into 25 theme units and the omission of one will result in the loss of four marks out of 100. For this purpose alone, many of them focused more on the information relating to the name of the character, the date of the major events and so forth. Such responses found resonance in the answers given for the second question, which is about what students have written down in their notes, whether names, figures, dates or phrases and lexical chunks. Nearly every interviewee said that they have taken some notes that included mostly key words like nouns and verbs. To write down as fast as they could those words and phrases were their priority. Some applied the skills they have learned in note-taking, for example, the use of abbreviations.

Questions 3 and 4 were designed to know if the interviewees have remembered or noted down the words and phrases in the original story, and if they follow their notes and use them in their reproduction of the story. It was found that learners tended to first try their memories and then refer to their notes. They said that if they could remember the original words and phrases they would use the exact phrase because it was safer. However, if they couldn't, they would rely on their linguistic

knowledge to restructure what they wanted to express. One interviewee told us that when she couldn't say "*dropped in*", she simply used "*visit*" instead. Another interviewee said when she couldn't remember how the word "*pat*" was used in the story; she just used "*pat on the shoulder*". The learners would use some formulaic sequences like "*…not only … but also…*" even if they did not write them down in their notes. Generally speaking, all interviewees expressed the idea that whenever they could they would use the phrases that they got from the source text because it would be more idiomatic and appropriate.

Question 5 focused on test takers' belief in using formulaic language. Do they agree that the more chunks they use, the more fluent they are? Most of our interviewees believe that using the words and phrases appearing in the ST is more convenient, safer, more idiomatic and economical and it would be better to echo all the words of the story, including all the lexical chunks. But the fact is that it is impossible to remember all of them, even with the help of the notes; they had to turn to their linguistic knowledge especially when they wanted to connect their ideas and make coherent sentences in their story retelling.

6.5 Discussion

With the nature of retelling in mind, that is, to comprehend a story or event first and then convey the same information or create the same image in listener's mind, we should say retelling as an assessment task is more integrative than other forms of oral assessment. Story retelling can help reveal learners' ability to make inferences as they organise, integrate, and classify information from the source story, using the target language. Story retelling also provides learners an opportunity to analyse stories and build their oral language as they acquire related vocabulary (Scheinkman, 2004).

Features of FS in learners' story retelling are significant in that it could inform us what, in essence, is the true reflection of learners' oral English competence. From the corpus analysis, we do find the use of FS facilitated learners to achieve a better performance in the retelling assessment task. Those frequently used ready-made chunks, such as *"run away (90.48%)", "the middle-aged woman (73.81%)", "three years ago (73.81%)", "why don't you… (66.67%)",* and *"a basketball match (54.76%)"*, assisted test takers in completing the retelling assessment task, allowing them to handle the task with less effort, and more fluently reproduced texts could therefore be expected. Large FS like "*because of a pair of sport(s) shoes*" made it easy for them to construct longer sentences. Retelling

is different from free conversation or monologue. In retelling learners consciously recall what they have heard in terms of contents and language. A better performance in a retelling assessment task does not usually involve the expressions of novel ideas and personal opinions and attitudes; nevertheless, speakers do have to fit together the information they got from the original story and express it in grammatically correct utterances in which the FSs played a very important role.

The story in this particular test is not difficult for this group of English majors to understand, if it is used as a listening input. However, in the process of story retelling, they did not use all the FSs they have heard. There are 14 FSs that were used by less than ten per cent of the subjects. Simple sequences like *"walked out of…" and "at low prices… "* were only used by one learner respectively, which is quite surprising. Such a phenomenon demonstrates well that receptive skills are not equal to productive ones and that to understand FSs does not guarantee its uses (Ding & Qi, 2005). Although the subjects are long-time learners of English (more than ten years), they can only use some simple FSs in the target language production; and those more difficult ones, especially those with no Chinese equivalents, are less favoured or used.

The analysis also indicates that learners used different strategies in performing retelling. Some might pay more attention to follow the ST, focusing more on the original language features. For example, they noted down more FSs. Others might depend more on their ability to memorise the content and restructure the target language according to their grammatical knowledge of English. Toward some difficult FSs, some of them used avoidance strategy in communication. When they were not particularly sure about the usage of a certain phrase, they would rather choose a similar sentence structure, to play it safe. This is in accordance with Ding et al.'s (2005) study, in which they found that students' command of lexical chunks in retelling was also shown by their inadequate use of the chunks they had heard from the listening material.

To achieve a good performance in story retelling, various kinds of grammatical knowledge have to be exercised by the learners apart from memorising those ready-made FSs. With the understanding of the content as the supporting framework, learners can make free changes with the phrase and sentence patterns. The inadequate use of some FSs in LTs suggested that learners are able to communicate even though they do not remember those FSs exactly. Here grammatical knowledge plays a significant role. In other words, in the process of retelling, the source input provides the test takers with a "scaffold" with which they

can creatively reconstruct a new discourse. Though the application of FSs does contribute to the fluency and closeness to the ST, it is fairly important to have a good knowledge of the accurate form, that is, the grammatical well-form-ness of the target language, in order to achieve a better result in a test situation.

In completing the retelling assessment task, learners noticed the gap between what they had listened to, comprehended and remembered, and what they were able to express in the target language. In this case, the use of the FSs may have played a less important role than the grammatical knowledge that they have learned and that is retrievable from their long-term memory. Learners' linguistic knowledge must have played a very important role in producing a fluent piece of story. In our data, we found many examples that can demonstrate learners' use of their grammatical and lexical knowledge to reproduce the story. Without previous knowledge of both FSs and linguistic rules, it is hard to reproduce a successful piece of work. Retelling is a task that can demonstrate the task takers' ability to use both the language by using linguistic rules and the FSs they have gained from the input. The new oral discourse is not a word-for-word production of the original; instead it should be a novel and creative one.

Therefore, based on the output of the retellings of English learners, we can study the relationship between learners' use of FSs and their potential linguistic creativity. If retelling does not require the repetition of the exact words from the original, it means that there will be much room for linguistic creativity, because in the process of retelling, learners can unavoidably and creatively use the target language. In doing a similar research, Ding & Qi (2005) found most students can understand the content of the story, but when they started to retell they only made use of a few lexical chunks that appeared in the ST.

The follow-up interview served well as a means of knowing what learners did while performing the retelling task, whether they preferred to use the FSs or not. We learned from our interviewees that they paid more attention to the content in listening, and made efforts to write down the key words and phrases. However, when they started to retell the story, they had to rely on their linguistic ability to make the retelling sound as close to the original story as possible.

Retelling is a task that involves the processes of absorbing, saving, internalising, understanding and expressing. It is a useful technique for checking learners' understanding and language reproduction ability. Unlike listening comprehension tests that use multiple-choice questions, story retelling requires test takers to reproduce large segments of

text and think about the sequences of ideas or events and their relative importance. To study this special type of output from Chinese EFL learners gives us an opportunity to know what the real competence of our learners is in terms of both their use of FSs and their potential linguistic competence. We believe that high quality texts reproduced are the results of constant use of those prefabricated FSs and learners' grammatical knowledge.

6.6 Conclusion

In the research, firstly, we looked at the general features of the FSs produced by Chinese EFL learners in a story retelling assessment task. Secondly, learners' comments on their own production in the follow-up interviews reveal part of the reasons why they cannot "remember" all the lexical chunks. The findings of the analyses suggest that learners, even facilitated by the input from the source being readily available (in this case, the original story), need to exercise not only their ability to remember the ready-made formulaic sequences, but also their general linguistic knowledge and capacity to construct meaningful and correct sentences using the target language. Memorisation of the ready-made chunks they heard in the ST alone is not sufficient for completing successfully the story retelling task under examination condition. The findings also support the use of story retelling tasks as an appropriate tool in EFL context for measuring learners' ability in comprehension, selection of important information, as well as their ability to use the target language to construct meaningful texts orally. In addition, we argue that story retelling as a pedagogical task can help the learners to develop their integrated English speaking skills for communication.

Appendix 6.1 The story (source text)

Mr. Smith owned a small supermarket. All the people nearby like to go shopping there. Several months ago, a few big chain stores were opened in town, they sold everything from toothpaste to televisions, and the prices were pretty low. Many small shops were closed down, except Mr. Smith's small supermarket. The owner of a chain store was very curious.

One day, he dropped into Mr. Smith's supermarket, and saw a middle-aged woman buying fruit. When the woman walked out of the

supermarket, he stopped the woman politely and asked her, "Madam, why don't you go shopping in the large chain stores? They have many more kinds of fruit at low prices." With a smile, the woman said, "You want to hear? It's because of a pair of sports shoes." Then the woman told her story.

Three years ago after her divorce, she had to work very hard to raise the family. One of her two children, Tommy, was in elementary school. One day, when she was buying food in the supermarket, Tommy rushed in and pleaded with her for a pair of sports shoes. Pointing to the shoes, Tommy cried, "Mummy, I had a basketball match today. They all laughed at me when I played with my bare feet." She really wanted to buy her son a nice pair of sports shoes, but the money in her pocket was barely enough for a meal. "I am sorry, Tommy," she said sadly, "I promise when we have money…" Seeing no hope in his mother's eye, Tommy ran away. Standing there without knowing what to do, she started to weep.

Suddenly, she felt someone pat her on the shoulder. She turned around and saw Mr. Smith, the owner of the supermarket, holding the pair of Adidas that her son dreamed of. "Take them." He said with an understanding smile. "But I don't have money," she shook her head. "Your son can't wait," he said, "But I can wait. Take it, and pay me later."

After hearing the story, the owner of the big chain store was deeply touched. Now he knew why people still liked to go shopping in Mr. Smith's supermarket. Not only could you find all kinds of commodities there, but also understanding, love and trust.

Appendix 6.2 Interview questions

1. What do you pay attention to while listening to the story? (content words and phrases or lexical chunks)
2. What can you usually note down? (name, figures, date, phrases/chunks)
3. Can you use some of the lexical chunks that appeared in the story? For example…
4. If you have remembered or noted down the words and phrases in the original story, do you use them?
5. Do you think by using some lexical chunks, you will sound more fluent?
6. Are you satisfied with your performance? What are the difficulties in your retelling? Why?

Note

1. AntConc3.2.4 is a computer program developed by the Japanese expert Laurence Anthony and this software is available on the Internet (www.antlab.sci.waseda.ac.jp/software.html)

References

Aitchison, J. (1987). *Words in the minds*. New York: Basil Blackwell Inc.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. Cowie (ed.). *Phraseology: Theory, analysis and applications* (pp. 101–122). Oxford, Oxford University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Pearson Education.

Brazil, D. (1995). *A grammar of speech*. Oxford: Oxford University Press.

Bolinger, D. (1976). Meaning and Memory. *Forum Linguisticum, 1*(1), 1–14.

Coulmas, F. (1979). On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics, 3*(3–4), 239–266.

Cowie, A. P. (1998). Introduction. In A. P. Cowie (ed.), *Phraseology: Theory, analysis and applications* (pp. 1–2). Oxford: Oxford University Press.

Ding, Y., & Qi, W. [丁言仁, 戚焱] (2005). Use of formulaic language as a predictor of L2 oral and written performance [词块运用与英语口语和写作水平的相关性研究]. *Journal of PLA University of Foreign Languages* [*解放军外国语学院学报*], *28*(3), 49–53.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition, 18* (1), 91–126.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Fillmore, W. (1976). *Cognitive and social strategies in second language acquisition*. Stanford University. PhD Thesis.

Hatch, E. (1978). *Second language acquisition*. Rowley, Mass.: Newbury House.

Hunston, S., & Francis, G. (1999). *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Kuiper, K. (1996). *Smooth talkers: The linguistic performance of auctioneers and sportscasters*. Mahwah, NJ: Erlbaum.

Leech, G. (2000). Same grammar or different grammar? Contrasting approaches to the grammar of spoken discourse. In S. Sarangi & M. Coulthard (eds.), *Discourse and social life* (pp. 48–65). London: Longman.

Lewis, M. (1993). *The lexical approach: The state of ELT and the way forward*. Hove, England: Language Teaching Publications.

Liu, X., & Liu, X. [刘晓玲, 刘鑫鑫] (2009). A corpus-based study on the structural types and pragmatic functions of lexical chunks in college English writing [基于语料库的大学生书面语词块结构类型和语用功能研究]. *Chinese Foreign Language* [*中国外语*], *6*(2), 48–53.

Morrow, L. M. (1985). Retelling stories: A strategy for improving children's comprehension, concept of story structure and oral language complexity. *Elementary School Journal, 85*(5), 647–661.

Morrow, L. M. (1986). Effects of structural guidance in story retelling on children's dictation of original stories. *Journal of Reading Behavior, 18*(2), 135–152.

Morrow, L. M. (1996). Story retelling: A discussion strategy to develop and assess comprehension. In L. B. Gambrell & J. F. Almasi (eds.), *Lively discussions: Fostering engaged reading* (pp. 265–285). Newark, DE: International Reading Association.

Nattinger, J., & DeCarrico, J. (1992). *Lexical phrasal and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam & Philadelphia: John Benjamins.

Partington, A. (1998). *Patterns and meanings – using corpora for English language research and teaching*. Ameterdam: John Benjamins

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt (eds.), *Language and communication* (pp. 191–226). New York: Longman.

Peters, A. M. (1983). *Units of language acquisition*. Cambridge: Cambridge University Press.

Scheinkman, N. (2004). Picturing a story. *Teaching Pre K-8, 34*(6), 58–59.

Schmitt, N. (ed.) (2004). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action – an introduction. In N. Schmitt (ed.), *Formulaic sequences acquisition, processing, and use* (pp. 1–22). Amsterdam: John Benjamins.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38–62.

Skehan, P. (1998). *A cognitive approach to language learning*. Shanghai: Shanghai Foreign Language Education Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

Van Pattern, B. (1990). Attending to content and form in the input experiment in consciousness. *Studies in Second Language Acquisition, 12*(3), 287–301.

Vihman, M. M. (1982). A note on children's lexical representations. *Journal of Child Language, 9*(1), 249–253.

Weinert, R. (1995). The Role of Formulaic Language in Second Language Acquisition: A Review. *Applied Linguistics, 16*(2), 181–205.

Wei, N. [卫乃兴] (2007). Phraseological characteristics of Chinese learners' spoken English: Evidence of lexical chunks from COLSEC [中国学生英语口语的短语学特征研究— COLSEC语料库的词块证据分析]. *Modern Foreign Languages* [现代外语], *30*(3), 280–291.

Widdowson, H. G. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.

Willis, D. (1990). *The lexical syllabus: A new approach to language teaching*. London: Collins COBUILD.

Wray, A. (1999) Formulaic language in learners and native speakers. *Language Teaching*, 32(4), 213–231.
Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21*(4), 463–489.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Yorio, C. A. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly, 14*(4), 433–442.

7

Assessing Incidental Vocabulary Learning by Chinese EFL Learners: A Test of the Involvement Load Hypothesis

Chanchan Tang and Jeanine Treffers-Daller

7.1 Introduction

One of the most challenging tasks for learners of a Second Language (L2 learners) consists in developing a vocabulary large enough to be able to read and write fluently and take part in conversations on a range of topics. According to Adolphs and Schmitt (2003) learners need 2000–3000 of the most frequent English word families to be able to take part in everyday conversations, whilst they need 5000 word families to begin to read authentic texts (Schmitt, 2007). For unassisted comprehension of written texts it is assumed learners need around 8000–9000 word families, and a vocabulary of 6000–7000 word families for spoken text (Nation, 2006). Many researchers have indicated that L2 learners worry about the formidable task of learning thousands of words (see for example Jones, 1995; Kim, 2008; Lawson & Hogden, 1996), particularly in contexts where learners have few opportunities to go to the country of the target language and/or have little knowledge about the target language culture, as is the case for many Chinese learners of English (Shao, 2014). For teachers it is equally challenging to find ways to help students acquire a wide range of words within the limited class time. Researchers can help address this issue by providing evidence regarding the effectiveness of different approaches to vocabulary learning and teaching.

There are of course many different ways to build up a vocabulary. Hunt and Beglar (1998) outlined three approaches to enhance vocabulary learning—incidental learning, explicit instruction, and independent strategy development. Nation (1990) has shown that intentional vocabulary learning, in particular instruction, does aid in the learning

of words, especially in the earlier stages of language learning. However, because of the limited amount of time that is available in class, only a few words can be taught by direct instruction (Nagy & Herman, 1987). Instead, the large majority of words are assumed to be acquired while the learner is reading a text or listening to a message, and focuses on the content instead of on learning words, that is through incidental vocabulary learning (Hulstijn, 2003). However, more recent work in this field shows that vocabulary gains from reading or writing are very limited. According to Nation and Wang (1999) at least ten exposures are needed if learners are to be successful at learning unknown words. More recently, Pellicer-Sanchez (2012) has shown that effects of frequency of exposure of new words are significant from three to five repetitions onwards, whilst unknown words that are repeated eight times begin to be read like known words. Because L2 learners often have limited exposure to target language input, and their input is generally limited to classroom contexts, they are unlikely to make large vocabulary gains by repeated exposure alone.

As there is considerable evidence that vocabulary take-up from reading is rather limited, researchers need to focus their attention on how incidental vocabulary learning can be promoted in the process of reading, for example by encouraging learners to use dictionaries, or providing glosses or asking learners to engage in different post-reading tasks (Hulstijn, Hollander & Greidanus, 1996; Ko, 2005). More research needs to be done to explain the reasons for the different degrees of effectiveness of tasks (Anderson, 1995; Joe, 1995, 1998; Paribakht & Wesche, 1997). One possibility is that the effectiveness of each task is determined by the depth of processing of vocabulary items by learners, but operationalizing depth of processing is difficult. For this purpose, Laufer and Hulstijn (2001) put forward the Involvement Load Hypothesis (ILH), which is based on the idea that incidental vocabulary can be promoted by involving learners in different post-reading tasks which require learners to engage with the words in the text in a variety of ways. Tasks differ from each other with respect to the degree of processing depth needed to carry out the task (see Section 2 for details).

Laufer and Hulstijn (2001) and Hulstijn and Laufer (2001) call for further tests of the ILH, although few researchers (Kim, 2008, 2011; Yaqubi et al., 2010) have so far responded to their invitation. The current study aims to contribute to the discussion by focusing on two points which the authors mention in their papers as particularly relevant for tests of the ILH, namely the relative importance of different components of

the ILH and the differences between the effects of input-oriented and output-oriented tasks on vocabulary learning and retention.

The participants in our study are Chinese learners of English who are at A2 level in English. So far the ILH has not been tested with students who have a relatively low level of proficiency in English. The participants in Kim's (2011) study were adult L2 learners of English, divided over two groups: one group were enrolled on a pre-university intensive course and had TOEFL scores between 470 and 520 and the other group were university students who had a TOEFL score above 520. While Kim (2011) found no interaction effects between language proficiency and task type in their study of the ILH, they also call for including a wider range of proficiency in future studies testing the ILH, and the current study therefore fills an important gap in our knowledge by focusing on the lower end of English language ability.

7.2 Incidental vocabulary acquisition

In the domain of L1 and L2 pedagogy, the term incidental vocabulary acquisition is understood to mean "learning without an intent to learn, or as the learning of one thing, e.g. vocabulary, when the learners' primary objective is to do something else, e.g. to communicate" (Schmidt, 1994, cited in Laufer & Hulstijn, 2001, p. 10). In other words, incidental vocabulary acquisition means that learners focus on understanding the meaning of spoken or written information while reading or listening and not on vocabulary learning *per se*. In such a process, new words are acquired "as a by-product of other cognitive exercises involving comprehension" (Gass, 1999, p. 319). In practical terms, incidental vocabulary learning can be operationalized as a learning process with absence of any forewarning of subsequent retention tests (Hulstijn, 2003).

Paribakht and Wesche (1997) were among the first to show that vocabulary learning can be promoted through a combination of reading and enhancement activities. They found that words practised through exercises were retained better than words for which the meaning was inferred from the context. Hence, asking learners to carry out tasks could be an effective tool for vocabulary learning as this might stimulate learners to process words more deeply. Although depth of processing remains difficult to measure objectively, it is likely that the nature of processing activities in which learners engage affects their retention of information: more elaborate processing activities will lead to better retention. Laufer and Hulstijn (2001) were the first to apply Craik and

Lockhart's (1972) depth of processing hypothesis to vocabulary learning by proposing the Involvement Load Hypothesis (ILH), which claims that each task can induce a certain amount of "involvement load", and that the effectiveness of a task is determined by the "involvement load" it induces. Put simply, the more learners engage with the words they learn (for example by focusing on the spelling, the meaning or aspects of the way the words are used), the better they will retain them.

The motivational-cognitive construct of involvement consists of three basic components: *need, search* and *evaluation*. *Need* is a motivational construct, concerned with the "need to achieve" (Laufer & Hulstijn, 2001, p. 14), whilst *search* and *evaluation* belong to cognitive dimensions, concerned with noticing and attending to form-meaning relationship (Schmidt, 1994, cited in Laufer & Hulstijn, 2001). *Need* refers to the motivation of learning target words and the drive to comply with task requirements. *Search* is the attempt to find the meaning of an unknown L2 word or the attempt to find a suitable L2 word form for a particular L1 concept. *Evaluation* refers to whether or not learners are required to compare the target words with other words. Tasks can of course induce these involvement factors to different degrees. For the purpose of the ILH, the authors suggest that there are three possible levels of involvement for each: none, moderate and strong. All three involvement factors may not be at work simultaneously during a reading-based task, or in other words, a task can induce any one, two, or all three of the components of involvement for each word. The involvement load of a task is defined as the combination of the three involvement factors, which can be absent or present, moderate or strong (see Section 3 for more details).

So far, few researchers have attempted to directly test the ILH, although some researchers have tested aspects of it. Yaqubi et al. (2010) are among the few who tested ILH, but they did not find that tasks that induced a higher involvement load led to higher scores on the post-test. Instead, they claim that output-oriented tasks lead to better results regardless of the degree of involvement load of the tasks. Kim (2011), on the other hand, found moderate support for the ILH in initial vocabulary learning because learners acquired words more effectively through tasks that induced a higher degree of involvement, as tested in an immediate post-test. In addition, Kim found strong support for the ILH in retention of vocabulary in a delayed post-test. Importantly, the author suggests that the effects of different tasks may not be visible immediately but only at a later stage, and calls for further investigation of the long-term effects of tasks with different involvement loads.

We aim to first of all test the central claim of the ILH that tasks with a higher involvement load will be more effective than those with a lower involvement load. Second, we hope to contribute to the discussion about the relative contribution of different components of involvement to vocabulary acquisition. There is some evidence that the different components of involvement do not have an equal impact on students' vocabulary retention. Laufer and Hulstijn (2001) suggest that *search* may be less important than the other two components, and Kim (2011, p. 125) found some evidence that "strong evaluation induces much greater involvement in processing a word than the moderate evaluation and the other two components." It is therefore important to investigate the contribution of different components of involvement in greater detail. Third, we will look into differences between input-oriented tasks and output-oriented tasks, and test Laufer and Hulstijn's claim that these two task types are equivalent as long as the involvement load of the tasks remains constant. This is relevant because Yaqubi et al. (2010) found that output-oriented tasks and input-oriented tasks with the same involvement indices were not equally effective. Our fourth research question relates to the differences between initial vocabulary learning (which is measured with an immediate post-test) and vocabulary retention (to be measured with a delayed post-test), as we wanted to find out whether a higher involvement load leads to better vocabulary retention in the longer term, as Kim (2011) suggested. In Hulstijn and Laufer (2001), one of the few studies which directly tested the ILH, the superior results of the students who were required to write compositions may well have been due to the fact that they had more time to spend on the task, as Kim (2011) points out. This is an important point that Hill and Laufer (2003) explored in great depth in a follow-up study on incidental vocabulary learning. In the current study, we will carefully control for time-on-task, to ensure that differences in vocabulary learning and retention cannot be explained by differences in time-on-task.

For a variety of reasons it is particularly difficult for learners with Chinese as their L1 to learn English words. First of all, Chinese and English belong to different word families (Sino-Tibetan and Germanic), which means there are virtually no cognates between the languages which could facilitate vocabulary learning (Larrañaga, Treffers-Daller, Tidball & Gil Ortega, 2012). Second, while learning vocabulary in another language generally involves learning a new way to map meaning onto form, this is particularly complex for learners whose L1 uses a logographic script and who need to learn words in an alphabetic script (see also Cheng & Yang, 1989, who investigated differences in

processing of characters and words). Third, many Chinese learners are not very familiar with incidental vocabulary learning, because the key vocabulary learning strategy in EFL classrooms in China and Taiwan is rote learning of vocabulary lists (Li, 2004; Smith, Kilgarriff & Sommers, 2008) and there are few opportunities for learning words from meaning-focused input, that is listening and reading in the classroom (Nation, 2007). For this reason a study into incidental vocabulary learning and assessment among Chinese learners is very much needed.

7.3 Methodology

7.3.1 Participants

This experiment was conducted in a secondary vocational school in China. The participants were 230 students (male and female) in six intact classes in the second term of Grade One. These six classes were used as six groups in the experiment and each carried out one of the six different tasks specified below. The majority (83 per cent) of participants were aged 18, 14 per cent were 17 years old and 3 per cent were 16 years old. At the time of data collection they had learned English for three years in junior school. The participants' proficiency level was roughly equivalent to the A2 level on the Common European Framework of Reference for Languages, which provides widely-used guidelines used to describe achievements of learners of foreign languages (Council of Europe, 2001). This was established by comparing students' performance on their high school examinations to the CEFR descriptors. Since they had been allocated to classes based on the results of the high school exams, the overall English level of the participants in these six classes was quite similar. The first author also obtained access to the students' most recent mid-term English examinations, the average scores of which were within a range of three points (the total score of the paper was 100 points), which lends support to our assumption that students were at similar levels of language proficiency. The experiment was carried out towards the end of the second term in Grade One, a period during which the participants were moving to Grade Two.

7.3.2 Choice of reading passage and target words

Since the participants had almost completed the second term of Grade One and would move on to Grade Two after the summer holiday, the text was taken from the textbook for the first term of Grade Two (see the appendix to this chapter). The first author selected the passage so that its level of difficulty was appropriate but still challenging for the

learners, to ensure the participants could understand the general meaning and at the same time acquire new vocabulary items incidentally. To examine the difficulty of the vocabulary in each text, several reading passages were analysed with the help of the VocabProfile program (http://www.lextutor.ca/vp/eng/), which provides information about the frequency layers to which the words in a text belong. In addition, the first author consulted the teachers from the school as well as the Word Bank (vocabulary glosses) provided after the reading passage in the original textbook. On the basis of the information obtained from all these sources, the most suitable reading passage from among the first chapters of the book for Grade Two was chosen for the experiment. It is a narrative about an event that took place at an airport. The passage contains 222 words of which ten were unknown according to the students' teachers. Prior to the main data collection we also carried out a pilot study to investigate whether these were unknown. Ten students from the same level took part in this pilot. These students did not take part in the main study. The ten words were indeed unknown, except for *down*, because some students in the pilot had partial knowledge of this word: the students selected "antonym of up" as the meaning of *down*, wich is possible, but not appropriate in the context. In other words, 95.5 per cent of the words in the selected passage were known, which made this a suitable but sufficiently challenging reading task.

We selected eight words as the target words in the experiment from among the ten originally chosen. The eight target words were: *airline, backup, frightening, kick, luggage, screen, spread,* and *stare at*. Two of the ten were excluded from the analysis: *down*, for reasons mentioned above and *point at* because the teachers told us that using two phrasal verbs in this study was too complex at this level. We decided to include words from different categories to ensure there would be some variation in level of difficulty among the target words. According to Ellis and Beaton (1993), part of speech is an important determinant of the learnability of words, with nouns being the easiest to learn, followed by adjectives, whilst verbs and adverbs are the most difficult to learn for L2 learners. The target words consisted of three nouns, two verbs, two adjectives and one phrasal verb, which means that the most and the least difficult word categories were included in the study. We also consulted VocabProfile to determine the frequency levels of these eight target words. Four of these (*frightening, kick, screen,* and *spread*) belonged to the 2000 word frequency band. The words *airline* and *stare at* were found to belong to the 3000 word frequency band, whilst *backup* and *luggage* were from the 4000 and the 7000 bands respectively. All ten new

words were printed in **bold** to increase the saliency of these words to participants. According to Sharwood Smith (1993), if a word is salient in the input to the learners, there is a greater chance for it to be selected and processed by the L2 learner. Input enhancement using bold typeface is one way in which input saliency can be increased.

7.3.3 Task design

We designed six tasks with different involvement loads to investigate the effects of the tasks, the amount of involvement load in each and the impact of the different components of involvement (*need*, *search* and *evaluation*) on vocabulary acquisition and retention. The degree of involvement required by the different tasks was expressed in an Involvement Index: absence of a factor was counted as 0, a modest involvement with a factor as 1, and strong involvement with factor as 2. The tasks carried out by the six groups are described below. As the aim of the study was to measure students' incidental acquisition of words from the reading passage (with or without additional activities), in line with the methodologies used in Hulstijn and Laufer (2001) and Kim (2011), students were not informed they would be tested on their knowledge of the words after reading the passages and undertaking the different tasks.

Group 1 reading only (-need, -search, -evaluation)
(see appendix 7.1, Task 1)

The students involved in this task were only asked to read the given passage. The Chinese translation equivalents were provided in the text just after each new English word and there were no post-reading activities. This means the learners do not feel the need to learn the words, nor did they need to search for the translation equivalents or compare the meanings of the words to other words. This task was classified as scoring zero on the Involvement Index (0+0+0).

Group 2 reading + comprehension questions (-need, -search, -evaluation)
(see appendix 7.1, Task 2)

The reading passage given to Group 2 was the same as the one used in Group 1, with the Chinese equivalent of the ten new words in the text. The difference from Task 1 was that there were comprehension questions which students had to answer after reading the text, but these were irrelevant to the target words. Since the new words were glossed in the text and they were irrelevant to the comprehension questions, the learners did not need to learn the words nor to search or evaluate

the words' meanings. Therefore, the Involvement Index of Task 2 is also 0 (0+0+0).

Group 3 reading + comprehension questions (+need, +search, -evaluation) (see appendix 7.1, Task 3)

In this task, the comprehension questions were designed with relevance to the target words, so the participants needed to know the meaning of the target words in order to complete the task, and the factor *need* was clearly present. In addition, the ten new words were glossed in the end of the passage alphabetically, so the factor *search* was triggered. Since the glossary listed the word meaning that was relevant for the context, the *evaluation* factor is absent. So the Involvement Index of Task 3 is 2 (1+1+0).

Group 4 reading + comprehension questions (+need, -search, +evaluation) (see appendix 7.1, Task 4)

The reading passage and comprehension questions were exactly the same as in Task 3. However, unlike Task 3, the new words were glossed in the margin rather than at the end, so there was no need to search for the meaning. Moreover, because different meanings of one word were provided, the participants needed to compare between them in order to choose one that was most suitable for the given context. For this reason, a modest *evaluation* was triggered. So the Involvement Index of Task 4 is 2 (1+0+1).

Group 5 reading + comprehension questions (+need, +search, +evaluation) (see appendix 7.1, Task 5)

Task 5 was designed to involve all three involvement load components. It shared the same reading passage and comprehension questions as Tasks 3 and 4, so a moderate *need* was present. Students also had to evaluate the word meanings, as in Task 4, so the *evaluation* component was induced. The difference was that the glossary was located at the end of the passage according to alphabet instead of being in the margin, so the participants needed to search the word meaning. Therefore, the Involvement Index of Task 5 is higher, namely 3 (1+1+1).

Group 6 reading + sentence production (+need, -search, ++evaluation) (see appendix 7.1, Task 6)

Task 6 shared the same reading passage with the previous tasks and also shared the same kind of glossary with Task 4 (with the word meaning

glossed in the margin and several options to choose from). The post-reading activity for Task 6 differed from the previous tasks in that students needed to create sentences rather than answer multiple-choice comprehension questions. In this task, in order to produce new sentences, participants were required to make a decision about additional words which would combine with the new word in an original text. Therefore, strong *evaluation* was induced. Therefore, the Involvement Index of Task 6 is 3 (1+0+2). Among all the tasks, Task 6 was the only one that was output-oriented.

The involvement load of each task is listed in Table 7.1. It shows that the involvement load is lowest for Tasks 1 and 2, moderate for Tasks 3 and 4, and highest for Tasks 5 and 6.

7.3.4 Procedure

The experiment was carried out in June 2011. Six classes were given six different reading tasks during their normal class time. They were only told to read the passage and complete the post-reading activities, except for Group 1, for which there were no post-reading activities. Students were not informed they would have to complete a vocabulary

Table 7.1 Involvement load of six tasks in the present study

Tasks	Involvement Components			Involvement Index
	Need	Search	Evaluation	
1. Reading with glosses in text but no comprehension questions afterwards	–	–	–	0
2. Reading with glosses in text but irrelevant to the comprehension questions	–	–	–	0
3. Reading with glossary in the end relevant to the comprehension questions	+	+	–	2
4. Reading with glosses in margin relevant to the comprehension questions	+	–	+	2
5. Reading with glossary in the end which consists of several options and relevant to comprehension questions	+	+	+	3
6. Reading with glosses in margin and make sentences afterwards	+	–	++	3

test afterwards. Each task took 17 minutes to complete. After the completion of the tasks, the task paper was collected and the participants were given a vocabulary test. They were required to provide the Chinese equivalents for these English words within eight minutes. Their answers were scored afterwards. The delayed post-test was held seven days after the immediate post-test, as in the studies of Hulstijn and Laufer (2001) and Yaqubi et al. (2010). All participants received the same vocabulary test again but the order of the items differed from that in the immediate post-test. This test also took eight minutes. The two post-tests were scored by the first author. The following scoring method was adopted in this study: Zero points were given for items which were not translated or were wrongly translated. One point was given for items which were semantically appropriate, such as the superordinate, synonym, but not the best possible translation for the target item. Two points were assigned to a complete correct translation. The maximum score that could be obtained on both tests was 16.

Forty-five of the 230 participants who took part in the study had to be excluded from the analysis because they either attended only one test or they did not provide consistent birthdays in two tests, which made it impossible for us to allocate the two test papers to the same student. Complete data sets were obtained from 185 students.

7.4 Results

Students obtained mean scores of 9.97 (SD 4.69) in the immediate post-test and to 6.17 (SD 4.13) in the delayed post-test (out of a maximum of 16 points). We used non-parametric tests to investigate whether these differences were significant because the scores from the immediate post-test and the delayed post-test were not normally distributed. The differences between the two post-tests were significant in a Wilcoxon's Signed Ranks test ($Z=-10.77$, $p < .001$).

The overall results of the study are displayed in Figures 7.1 and 7.2. It shows that students in groups 1 and 2 obtained the lowest scores, followed by those in groups 3 and 4, whilst students in groups 5 and 6 obtained the highest scores. The rank order of the groups is the same for the immediate and the delayed post-tests. The results of the Kruskal Wallis test reveal that the differences between six groups are significant in the immediate post-test ($\chi^2 -= 56.02$, df = 5, $p < .001$) as well as the delayed post-test ($\chi^2 = 65.21$, df = 5, $p < .001$).

In order to get more evidence about the impact of the involvement load of different tasks on vocabulary retention we regrouped the

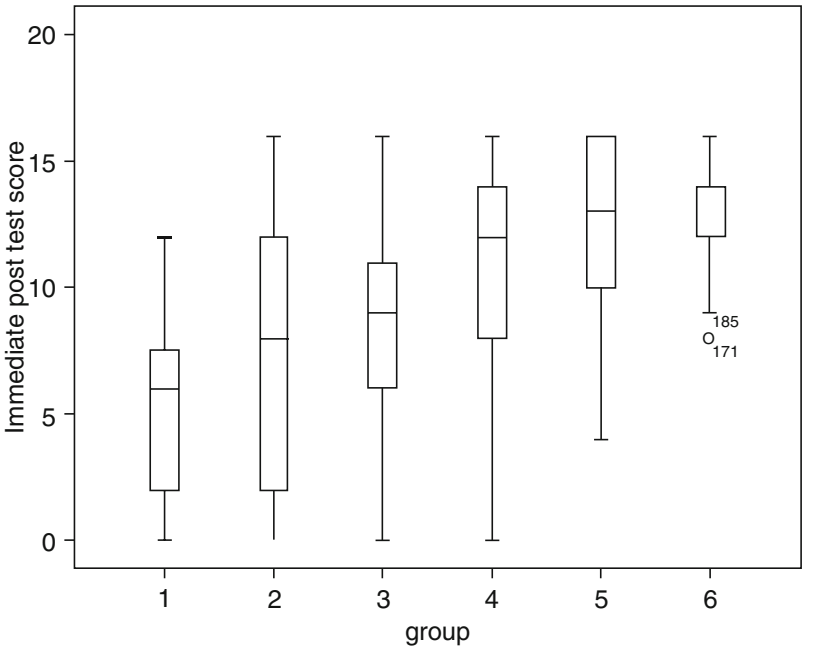


Figure 7.1 Immediate post test scores across groups

original six groups into three new groups according to their scores on the Involvement Index. The first two groups were combined to make Group A, as they both have an Involvement Index of 0. Group B consists of groups 3 and 4 because they share an Index of 2. The last two groups were combined to make Group C, since the Involvement Index of each is 3. As Table 7.2 shows, Group C performed best both in the immediate post-test and in the delayed post-test, whilst group A obtained the lowest mean score in two tests. The differences between these three new groups are significant in the immediate post-test (Kruskal Wallis, χ^2 =41.61, df = 2, p< .001) as well as the delayed post-test (χ^2 = 56.98, df =2, p< .001). In addition, post hoc comparisons indicate that all three groups are significantly different from each other (Table 7.3).

Further evidence regarding the importance of post-reading activities can be obtained from a comparison of the results of group 1 with those of all the other groups, because all groups except group 1 engaged in additional post-reading activities. After correcting for multiple

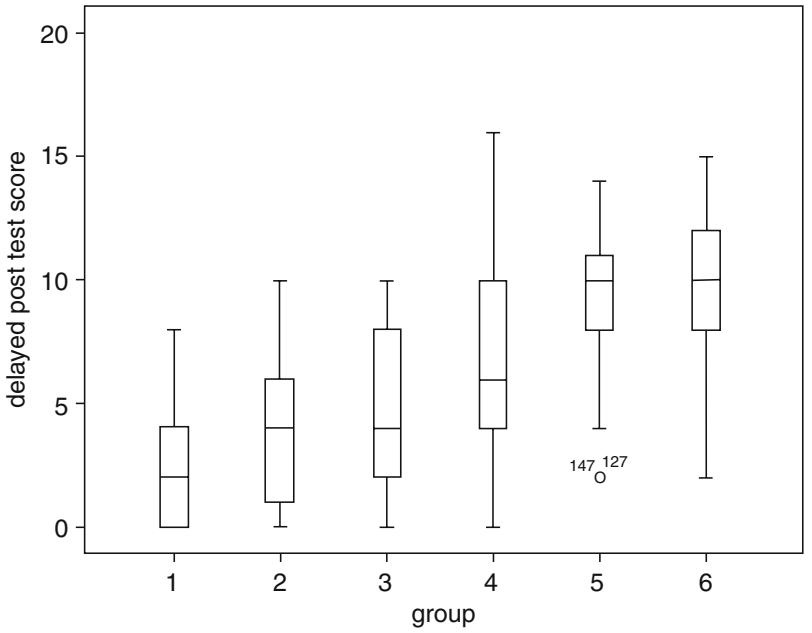


Figure 7.2 Delayed post test scores across groups

Table 7.2 Median of scores among the three groups based on the involvement load index

Groups based on the Involvement Load Index	Tasks	Involvement index	N	Median	
				Immediate	Delayed
Group A	Task 1 & Task 2	0	54	7.00	4.00
Group B	Task 3 & Task 4	2	67	10.00	6.00
Group C	Task 5 & Task 6	3	64	13.50	10.00

Table 7.3 Post hoc comparison of intergroup differences based on involvement load index

Group	Immediate post-test		Delayed post-test	
	z	*p* (adjusted)	*Z*	*p* (adjusted)
A–B	–3.05	.007	–3.01	.008
A–C	–6.43	.000	–7.44	.000
B–C	–3.60	.001	–4.71	.000

Table 7.4 Intergroup differences between group 1 and the other five groups (post hoc comparisons following Kruskal Wallis test)

	Z	*p* **(adjusted for multiple comparisons)**	*r* **(effect size)**
Groups 1–2	–2.442	ns	–
Groups 1–3	–2.060	ns	–
Groups 1–4	–4.693	<.001	.63
Groups 1–5	–5.273	<.001	.80
Groups 1–6	–6.214	<.001	.83

comparisons with the Bonferroni correction, the post hoc comparisons carried out on the immediate post-test show that Task 1 is not significantly different from Task 2 or Task 3; however, it is significantly different from Tasks 4, 5 and 6 (see Table 7.4). Effect sizes were computed manually for the Kruskal Wallis test, following the procedure outlined in Field (2013, p. 227 and p. 248).

The results for the delayed post-test were very similar in that group 1 was significantly different from groups 4, 5 and 6 but not from the other groups. For reasons of space these results are not reported in detail here.

We investigated the relative importance of the three components (*need, search* and *evaluation*) on vocabulary acquisition by regrouping participants into two groups according to the presence or absence of each component: ±*need,* ±*search* and ±*evaluation* (see Table 7.5 for details). We named each group based on the name and the status of the components. For example, Group NA referred to the group which performed the task where the *need* factor was absent; Group NP meant that the group which performed the task where *need* was present. The groupings based on the factors *search* and *evaluation* were created in similar ways. As we can see in Table 7.5, the median values in the factor-present group tended to be higher than in the factor-absent group in both the immediate and the delayed post-tests.

For *need*, the differences between the two groups were significant in the immediate post-test (Mann Whitney U test, U = 1773.5, p <.001) and the delayed post-test (U= 1599.50, p <.001), but for *search* the differences were not significant in either post-test. The grouping based on *evaluation* did result in significant differences in the immediate post-test (U =1794.50, p <.001) as well as the delayed post-test (U =1596.00, p < .001).

Table 7.5 Median of scores based on classifications of involvement components

Components	Groups	N	Tasks	Median	
				Immediate Post-test	Delayed Post-test
Need	Group NA	54	Task 1, 2	7	4
	Group NP	131	Task 3, 4, 5, 6	12	8
Search	Group SA	87	Task 1, 2, 4, 6	12	6
	Group SP	98	Task 3, 5	10	8
Evaluation	Group EA	84	Task 1, 2, 3	8	4
	Group EP	101	Task 4, 5, 6	13	8

Table 7.6 Effect sizes of group differences based on the classification according to *need*, *search* and *evaluation*

Group	Immediate post-test	Delayed post-test
	r	r
Group Need	.394	.434
Group Search	ns	ns
Group Evaluation	.499	.541

Thus, there are significant differences between the groups in their scores on the immediate post-test and the delayed post-test, if the grouping variable is *need* or *evaluation*, but not when *search* is chosen as the group factor. The effect sizes displayed in Table 7.6 for the three groupings indicate that *need* and *evaluation* are relevant factors in initial vocabulary acquisition and retention whilst *search* is not. A stronger effect size was found for *evaluation* than for *need* in the immediate post-test, and the same was true for the delayed post-test. This means that *evaluation* has a stronger impact on scores than *need*. The relative weight of each of the components of the involvement load model is therefore as follows: *evaluation* > *need* > *search*. Effect size differences appear to be slightly higher for the delayed post-tests than for the immediate post-tests for *need* as well as *evaluation*, but these differences are not significant.

Finally we focused on the impact of differences in the involvement load of tasks on the degree of vocabulary loss between the immediate

Table 7.7 Vocabulary loss between the immediate and the delayed post tasks

Vocabulary Loss	Mean	SD
Task 1	–2.91	2.234
Task 2	–3.90	3.833
Task 3	–4.10	2.857
Task 4	–4.51	3.280
Task 5	–3.13	2.617
Task 6	–3.85	2.949

post-test and the delayed post-test. Contrary to our expectations, all groups lost a roughly equal number of words between the immediate post-test and the delayed post-test (see Table 7.7). A Mann Whitney U test was used to investigate whether the differences between the six conditions with respect to the number of items lost after one week were significant, but these analyses did not reveal any significant differences.

7.5 Discussion

First of all it is important to note that all six treatments led to some acquisition of target words, which confirms that incidental vocabulary acquisition through reading is possible. However, different tasks had different effects on vocabulary learning and retention: the results revealed that scores increased from task 1 to task 6, which lends support to the ILH in that tasks with a higher involvement load are more effective for vocabulary acquisition than those with a lower involvement load, and confirms earlier findings of Paribakht and Wesche (1997) that practising words in post-reading activities supports vocabulary learning. Although Laufer and Hulstijn (2001) have argued that the depth of processing tasks would necessarily require a longer amount of time to complete, this study has shown that even when the time-on-tasks is controlled across tasks, the tasks with higher involvement load led to better vocabulary retention both in the immediate post-test and in the delayed post-test. This confirms earlier findings of Kim (2008; 2011), who also controlled for time-on-task, but contrasts with Yaqubi et al. (2010), who found students who carried out a task to which they had allocated a high Involvement Index obtained lower scores than students who performed tasks with a lower Involvement Index. We think the poor results of Yaqubi et al.'s task with a high Involvement Index might

be due to the fact that students had to look up the meanings of words in a dictionary. This might have been too difficult and time-consuming, and might have distracted from reading and understanding the text. In addition, it is not clear whether students brought the same dictionaries or different dictionaries, which is a confounding factor. Finally, it is not clear whether the authors controlled for time-on-task, which is essential to be able to evaluate the differences in type-of-task.

The fact that we found no significant differences between Task 1 and Task 2 shows that the nature of post-reading activities does matter. If questions about the text are unrelated to the target items (as in Task 2), students do not feel the need to learn the words, and the two tasks are equally (in)effective. However, the absence of significant differences between Task 1 and Task 3 indicates that answering questions that are relevant for the target items does not necessarily increase students' engagement with the words to a sufficient degree. It is only when they need to *evaluate* the words in the text against other words (as in Tasks 4, 5 and 6) that they need to process the words more deeply and this increases their chances of remembering the words in a post-test.

No differences were found in initial vocabulary learning or retention among students who carried out Task 5 or Task 6, which were equivalent with respect to the Involvement Index but differed from each other because Task 5 was more input-oriented and Task 6 more output-oriented. This confirms Laufer and Hulstijn's (2001) prediction that input- and output-oriented tasks will be equally beneficial for vocabulary acquisition if the involvement load is kept constant across tasks. Our results therefore provide little support for the findings of Yaqubi et al.'s study (2010), who concluded that tasks that are equivalent from the perspective of involvement but differ from each other because of their orientation towards input or output do not necessarily lead to the same results.

Our study provides clear evidence that the three components (*need, search* and *evaluation*) differ significantly from each other with respect to their impact on incidental vocabulary acquisition. The effect sizes revealed that the largest proportion of the variance was explained by *evaluation*, followed by *need*, with *search* in third position. Other researchers also concluded that *evaluation* is the most important component of involvement. Kim (2011) claims that this is particularly the case for initial vocabulary learning, but in our study this was also found to be the case for vocabulary retention. Although *need* is the second most important factor among the three, we found it is difficult to control or manipulate. For instance, Task 1 and Task 2 were designed *not* to trigger

need, but we could not ascertain that they did not trigger any need to learn the words among the learners. Maybe some participants in Task 2 felt a strong need to learn the words just because they were curious about the target words, although the post-reading tasks did not require any need to comprehend the target words. Hence, *need* remains hard to measure as its role depends to some extent on learners' motivation and attitudes towards the task.

Our study also revealed that there was a decrease in the mean scores among all groups from the immediate post-test to the delayed post-test, which is to be expected as learners often forget some newly learned words after a few days. Therefore, the results indicate that reinforcement of newly learned words is still needed if students are to remember them in the longer term, regardless of the amount of involvement load of the vocabulary learning task. The fact that effect sizes were slightly higher for the delayed post-test is interesting in the light of Kim's (2011) comments about the importance of investigating the long term effects of tasks with different involvement loads on the acquisition of L2 vocabulary.

Our experiment confirms Laufer and Hulstijn's (2001) suggestion that the impact of *search* on incidental vocabulary acquisition might be lower than that of the other two components, because *search* was found to have no significant effect on incidental vocabulary acquisition in our study. A possible reason for the lack of a significant effect of *search* may be the manipulation of the construct itself in the current study. In Laufer and Hulstijn's (2001) study, *search* was triggered by consulting a dictionary or teachers. In the current study, however, *search* was operationalized by referring students to a glossary at the end of the text with only L1 equivalents. We operationalized the presence or absence of *search* by the location of the glossary, that is *search* was absent when a marginal glossary was provided, but present when the glossary was presented at the end with words in alphabetical order. This kind of *search* was relatively limited by comparison with the approach suggested by Laufer and Hulstijn, and it may explain why in our study *search* was found to have little impact on vocabulary learning and retention.

7.6 Pedagogical implications

In the present study, we have seen how certain reading tasks can contribute to vocabulary acquisition, which may have important implications

for L2 teaching and learning in general, but in the Chinese context in particular. The study demonstrates, first of all, that learning vocabulary through reading is possible and feasible, but reading with enhancement activities tends to be more effective. It is particularly important that we have shown that incidental vocabulary learning works in the Chinese context, because there is less awareness in China of the potential of incidental vocabulary learning and learners tend to rely on rote learning to enhance their vocabularies (Li, 2004). As we have shown, task 1 (where translation equivalents were given in the text) and task 2 (with post-reading questions which are irrelevant to the target words) have a low involvement load because learners do not need to engage with the new words at all. Post-reading activities which are relevant to the target words, such as those in tasks 3, 4, 5 and 6 which require learners to engage with the meaning of the new words in a variety of ways, are much more beneficial for vocabulary acquisition. Teachers should therefore be aware of the importance of the involvement load of tasks they develop. Aiming at designing tasks with a high involvement load will not necessarily limit teachers' choice of task types. As the result from the comparison between Task 5 and Task 6 suggests, there is no significant difference between input-oriented tasks and output-oriented tasks both in initial vocabulary learning and in retention. As long as a high involvement load can be induced, teachers have many options in designing reading tasks.

Because learners are likely to forget some of the vocabulary items they have learned after a certain amount of time has elapsed, it is necessary for teachers to provide repeated exposures and additional tasks to maintain the initial vocabulary gains. For instance, teachers can create opportunities for students to encounter the same words in different contexts and to process the words several times in doing various post-reading tasks. Finally it may be beneficial for teachers as well as learners to investigate to what extent practice with incidental vocabulary learning changes students' own vocabulary learning strategies (Schmitt, 1997) and their perceptions of their own learning.

7.7 Conclusion

The present study set out to investigate the effects of different tasks on incidental vocabulary acquisition. In an attempt to test the Involvement Load Hypothesis (ILH), an experiment was conducted among Chinese students in a secondary vocational school whose

proficiency was estimated to be at A2 level on the CEFR. To the best of our knowledge, this was the first test of the ILH among students with a relatively low level of proficiency. The results showed that the students learned more words in reading tasks with a higher involvement load than in tasks with a lower involvement load both in the immediate post-test and the delayed post-test, which is in accordance with Hulstijn and Laufer's (2001) finding that tasks with higher involvement load lead to better vocabulary learning and retention. While through mere reading students can learn a certain number of words, this method is far from effective. We also found that the three components of involvement construct did not carry the same weight. *Evaluation* turned out to be the most important of the three and *search* was the least important. Students carrying out output-oriented tasks did not outperform those doing input-oriented tasks with the same involvement load. Thus, our study does not support the findings of Yaqubi et al. (2010) on this point.

As with Hulstijn and Laufer (2001), the current study focused on learning and retention of word meanings only. It remains to be seen whether tasks with a higher involvement load will lead to better learning of other aspects of word knowledge, as described in Nation (2001). For example, we do not know whether learning of derivational suffixes, formulaic sequences or collocations improves in tasks with a higher involvement load. Hence, it is recommended that future research should address the effects of involvement load on the learning of other aspects of vocabulary knowledge. Further research should also look into the long-term effects of tasks with different involvement loads, as the current study suggests the different effects of tasks persisted after one week, but we do not know if these effects would be measurable later on and which types of activities help support vocabulary retention in the longer term.

Appendix 7.1 Tasks 1–6

Reading passage:

When the computer is down

The most frightening words in the English language are "Our computer is down". You hear these words more and more when you are on business. The other day I was at the airport, where I was waiting for a ticket to Washington. But the girl in the ticket office said, "I'm sorry, our computer is down. That's the reason why we can't sell tickets."

I looked down at the computer and every passenger was just standing there drinking coffee and staring at the black screen. Then I asked her, "What do all you people do?"

"We give the computer the information about your trip, and then it tells us whether you can fly or not."

After the girl told me they had no backup computer, I said, "Let's forget the computer. What about your planes? They are still flying, aren't they?"

"I couldn't tell without asking the computer."

"Are there any other airlines that are flying to Washington within the next hours?"

"I wouldn't know," she said, pointing at the dark screen, "Only 'IT' knows. 'IT' can't tell me."

By this time there were quite a few people standing in lines. Word soon spread to other travellers that the computer was down. Some people went white, some people started to cry and still others kicked their luggage…
Target words: frightening staring at screen backup airline
spread kick luggage

Task 1

Gender: Birthday:
Directions: read the following passage for fun and get the general meaning in 20 minutes.

When the Computer Is Down

The most **frightening** (可怕的) words in the English language are "Our computer is **down** (停止运行的)". You hear these words more and more when you are on business. The other day I was at the airport, where I was waiting for a ticket to Washington. But the girl in the ticket office said, "I'm sorry, our computer is down. That's the reason why we can't sell tickets."

I looked down at the computer and every passenger was just standing there drinking coffee and **staring at** (盯着) the black **screen** (屏幕). Then I asked her, "What do all you people do?"

"We give the computer the information about your trip, and then it tells us whether you can fly or not."

After the girl told me they had no **backup** (备用的) computer, I said, "Let's forget the computer. What about your planes? They are still flying, aren't they?"

"I couldn't tell without asking the computer."

"Are there any other **airline**s (航线) that are flying to Washington within the next hours?"

"I wouldn't know," she said, **pointing at** (指向) the dark screen, "Only 'IT' knows. 'IT' can't tell me."

By this time there were quite a few people standing in lines. Word soon **spread** (传播，散布) to other travellers that the computer was

down. Some people went white, some people started to cry and still others **kick**ed (踢) their **luggage** (行李)…

Task 2

Gender: Birthday:
Directions: read the following passage and complete the comprehension questions in 20 minutes.

See the reading passage in **Task 1**

Reading comprehension

1. Where was the writer the other day?
   A. at home; B. at an airport; C. in a hotel; D. in a computer store
2. Which city was the writer taking the plane to?
   A. Washington B. Paris C. Tokyo D. London
3. Why the writer was going to that place? According to the passage, the most possible answer should be
   A. Visiting his/her friends.
   B. Visiting his/her family.
   C. Travelling.
   D. On business.
4. Where did the girl work?
   A. She was working in the ticket office.
   B. She was working in the school.
   C. She was working in the bank.
   D. She was working in the restaurant.
5. According to the passage, which of the following words were heard more and more in the English language?
   A. The tickets were sold out.
   B. Contact us during working hours.
   C. The airline has been cancelled.
   D. Our computer is down.
6. Why the girl told the writer that they couldn't sell tickets?
   A. Because the computer was down.
   B. Because the tickets were sold out.
   C. Because the writer had no money.
   D. Because the airline had been cancelled.
7. What were the other passengers drinking while standing in the line?
   A. They were drinking cola.
   B. They were drinking coffee.
   C. They were drinking fruit juice.
   D. They were drinking water.

8. According to the girl, were the planes still flying?
   A. Yes.
   B. No.
   C. She couldn't tell without asking the computer.
   D. She was unwilling to tell.

Task 3

Gender: Birthday:
Directions: read the following passage and complete the comprehension questions in 20 minutes.

When the Computer Is Down

The most **frightening** words in the English language are "Our computer is **down**". You hear these words more and more when you are on business. The other day I was at the airport, where I was waiting for a ticket to Washington. But the girl in the ticket office said, "I'm sorry, our computer is down. That's the reason why we can't sell tickets."

I looked down at the computer and every passenger was just standing there drinking coffee and **staring at** the black **screen**. Then I asked her, "What do all you people do?"

"We give the computer the information about your trip, and then it tells us whether you can fly or not."

After the girl told me they had no **backup** computer, I said, "Let's forget the computer. What about your planes? They are still flying, aren't they?"

"I couldn't tell without asking the computer."

"Are there any other **airline**s that are flying to Washington within the next hours?"

"I wouldn't know," she said, **pointing at** the dark screen, "Only 'IT' knows. 'IT' can't tell me."

By this time there were quite a few people standing in lines. Word soon **spread** to other travellers that the computer was down. Some people went white, some people started to cry and still others **kick**ed their **luggage**…

Reading comprehension

1. According to the passage, "Our computer is down" are the most___words in the English language.
   A. exciting
   B. terrible
   C. helpful
   D. cheerful

Vocabulary glosses:
airline n. 航班
backup adj. 备用的
down adj. 停止运行的
frightening adj. 可怕的
kick v. 踢
luggage n. 行李

point at 指向
screen n. 屏幕
spread v. 传播，散布
stare at 盯着

2. What was the airline that the writer was taking?
   A. The airline flying to Washington.
   B. The airline flying to Paris.
   C. The airline flying to Tokyo.
   D. The airline flying to London.
3. What was the matter with the computer?
   A. It was working actively.
   B. It was breaking into pieces.
   C. It was fine.
   D. It stopped working.
4. When the computer was down, the screen turned to be
   A. black B. green C. red D. yellow
5. What could the girl in the ticket office do for the passengers without asking the computer?
   A. She could sell a ticket.
   B. She could write out a ticket.
   C. She could answer the passenger's questions.
   D. She could do nothing.
6. If there had been a backup computer, which of the following situation would NOT happen?
   A. The girl could do nothing.
   B. The girl could sell a ticket.
   C. The girl could answer the passenger's questions.
   D. Everything would continue working.
7. Which of the following statement is NOT mentioned?
   A. Some people went white.
   B. Some people quarrelled with the girl.
   C. Some people started to cry.
   D. Some people kicked their luggage.
8. The last paragraph suggests that
   A. A modern computer won't be down
   B. Computers can take the place of humans
   C. Sometimes a computer may bring suffering to people
   D. There will be great changes in computers.

Task 4

Gender: Birthday:
Directions: read the following passage and complete the comprehension questions in 20 minutes.

When the Computer Is Down

The most **frightening** words in the English language are "Our computer is **down**". You hear these words more and more when you are on business. The other day I was at the airport, where I was waiting for a ticket to Washington. But the girl in the ticket office said, "I'm sorry, our computer is down. That's the reason why we can't sell tickets."

frightening adj. 可怕的
down adj. 情绪低落
adj. 停止运行的
prep. 向下

I looked down at the computer and every passenger was just standing there drinking coffee and **staring at** the black **screen**. Then I asked her, "What do all you people do?"

stare at 盯着
screen n. 屏幕
n. 纱窗
v. 掩藏，遮蔽
v. 放映，播放

"We give the computer the information about your trip, and then it tells us whether you can fly or not."

After the girl told me they had no **backup** computer, I said, "Let's forget the computer. What about your planes? They are still flying, aren't they?"

backup n. 增援，援助
adj. 备用的

"I couldn't tell without asking the computer."

"Are there any other **airline**s that are flying to Washington within the next hours?"

airline n. 航班

"I wouldn't know," she said, **pointing at** the dark screen,
"Only 'IT' knows. 'IT' can't tell me."

point at 指向

By this time there were quite a few people standing in lines. Word soon **spread** to other travellers that the computer was down. Some people went white, some people started to cry and still others **kick**ed their **luggage**…

Reading comprehension (see the reading comprehension in Task 3)

spread v. 展开，铺开
v. 传播，散布
v. 扩散，蔓延
kick v. 踢
v. 踢球得分；
射门得分
luggage n. 行李

Task 5

Gender: Birthday:
Directions: read the following passage and complete the comprehension questions in 20 minutes.

When the Computer Is Down

The most **frightening** words in the English language are "Our computer is **down**". You hear these words more and more when you are on business. The other day I was at the airport, where I was waiting for a ticket to Washington. But the girl in the ticket office said, "I'm sorry, our computer is down. That's the reason why we can't sell tickets."

I looked down at the computer and every passenger was just standing there drinking coffee and **staring at** the black **screen**. Then I asked her, "What do all you people do?"

"We give the computer the information about your trip, and then it tells us whether you can fly or not."

After the girl told me they had no **backup** computer, I said, "Let's forget the computer. What about your planes? They are still flying, aren't they?"

"I couldn't tell without asking the computer."

"Are there any other **airline**s that are flying to Washington within the next hours?"

"I wouldn't know," she said, **pointing at** the dark screen, "Only 'IT' knows. 'IT' can't tell me."

By this time there were quite a few people standing in lines. Word soon **spread** to other travellers that the computer was down. Some people went white, some people started to cry and still others **kick**ed their **luggage**…

Vocabulary glosses:	luggage n. 行李
airline n. 航班	point at 指向
backup n. 增援, 援助	screen n. 屏幕
adj. 备用的	n. 纱窗
down adj. 情绪低落	v. 掩藏，遮蔽
adj. 停止运行的	v. 放映，播放
prep. 向下	spread v. 展开，铺开
frightening adj. 可怕的	v. 传播，散布
kick v. 踢	v. 扩散，蔓延
v. 踢球得分; 射门得分	stare at 盯着

Reading comprehension (see the reading comprehension in Task 3)

Task 6

Gender: Birthday:
Directions: read the following passage and complete the comprehension questions in 20 minutes.

See the reading passage in **Task 4**
Making sentences with the following words.
1. frightening
2. stare at
3. screen
4. backup
5. airline
6. spread
7. kick
8. luggage

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24(4),* 425–438.

Anderson, J. R. (1995). *Cognitive Psychology and Its Implications* (4th ed.). New York: Freeman.

Cheng, Ch. M., & Yang, M.-Y (1989). Lateralization in the visual perception of Chinese characters and words. *Brain and Language, 36(4),* 669–689.

Chen, X., Ramirez, G., Luo, Y., Geva, E., & Ku, Y.-M. (2012). Comparing vocabulary development in Spanish- and Chinese speaking ELLs: The effects of metalinguistic and sociocultural factors. *Reading and Writing, 25(8),* 1991–2020.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf).

Craik, F. I. M., & Lockhart, R. S. (1972). Level of processing: A framework for memory research. *Journal of verbal learning and verbal behaviour, 11(6),* 671–684.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43(4),* 559–617.

Field, A. (2013). *Discovering statistics: Using IBM SPSS statistics* (4th ed.). London: Sage.

Gass, S. (1999). Discussion: Incidental vocabulary learning. *Studies in Second Language Acquisition, 21(2),* 319–333.

Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *International Review of Applied Linguistics, 41(2),* 87–106.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (eds). *The Handbook of Second Language Acquisition* (pp. 349–381). Oxford: Blackwell.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal

glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal, 80(3),* 327–339.
Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51(3),* 539–558.
Hunt, A., & Beglar, D. (1998). Current research and practice in teaching vocabulary. *The Language Teacher, 22(1),* 7–25.
Joe, A. (1995). Text-based tasks and incidental vocabulary learning. *Second Language Research, 11(2),* 149–158.
Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition. *Applied Linguistics, 19(3),* 357–377.
Jones, F. R. (1995). Learning an alien lexicon: A teach-yourself case study. *Second Language Research, 11(2),* 95–111.
Kim, Y. J. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58(2),* 285–325.
Kim, Y. J. (2011). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 61*(*Suppl. 1),* 100–140.
Ko, Myong Hee (2005). Glosses, comprehension, and strategy use. *Reading in a Foreign Language, 17(2),* 125–143.
Larrañaga, P., Treffers-Daller, J., Tidball, F., & Gil Ortega, M. (2012). L1-transfer in the acquisition of manner and path in Spanish by native speakers of English. *International Journal of Bilingualism, 16(1),* 117–138.
Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: the construct of task induced involvement. *Applied Linguistics, 22(1),* 1–26.
Lawson, M. J., & Hogden, D. (1996). The vocabulary-learning strategies of foreign-language students. *Language Learning, 46(1),* 101–135.
Li, X. (2004). *An analysis of Chinese learners' beliefs about the role of rote learning in vocabulary learning strategies*. Unpublished doctoral dissertation. University of Sunderland, Sunderland.
Nagy, W., & Herman, P. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. Mckeown & M. Curtis (eds.). *The nature of vocabulary acquisition* (pp. 19–35). Hillsdale, NJ: Lawrence Erlbaum.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. London: Newbury House.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63(1),* 59–81.
Nation, P. (2007). The four strands. *Innovation in Learning and Teaching, 1(1),* 1–12.
Nation, I. S. P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language, 12(2),* 355–380.
Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary development. In J. Coady & T. Huckin (eds.). *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174–200). Cambridge: Cambridge University Press.
Pellicer-Sanchez, A. (2012). Automaticity and speed of lexical access: Acquisition and assessment. PhD thesis, University of Nottingham.

Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N. C. Ellis (ed.). *Implicit and explicit learning of languages* (pp. 165–210). London: Academic Press.

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (eds.). *Vocabulary: Description, acquisition and pedagogy* (pp. 199–227). Cambridge: Cambridge University Press.

Schmitt, N. (2007). Current perspectives on vocabulary teaching and learning. In J.Cummins & C. Davison (eds.). *International handbook of English language teaching*: *Part II* (pp. 827–841). New York: Springer.

Shao, X. (2014). A study of Chinese college students' English reading anxiety. *American Journal of Educational Research 2(5)*. Retrieved from: pubs.sciepub.com/education/2/5/10/

Sharwood Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition, 15(2),* 165–179.

Smith, S., Kilgarriff, A., & Sommers, S. (2008). Making better wordlists for ELT: Harvesting vocabulary lists from the web using WebBootCat. Retrieved from http://www.kilgarriff.co.uk/Publications/2008-SmithKilgSommers-DAE-BetterWordlists.doc

Yaqubi, B., Rayati, R. A., & Allemzade Gorgi, N. (2010). The involvement load hypothesis and vocabulary learning: The effect of task types and involvement index on L2 vocabulary acquisition. *The Journal of Teaching Language Skills, 2(1),* 145–163.

8

Chinese Users' Perceptions of the Use of Automated Scoring for a Speaking Practice Test

Xiaoming Xi, Jonathan Schmidgall and Yuan Wang

8.1 Introduction

The use of automated scoring in large-scale testing for high-stakes purposes has generated a lot of debate and controversy in the last decade, and researchers have described a variety of ways in which the use of automated scoring may impact validity (e.g., Clauser, Kane, & Swanson, 2002; Weigle, 2010; Xi, 2010). Validity frameworks that tie automated scoring into the overall validity argument for using an assessment (Williamson, Xi & Breyer, 2012; Xi, 2012) highlight the many ways in which the use of automated scoring may impact the validity argument for the entire assessment, including test takers' interactions with assessment tasks; the accuracy and generalizability of assessment scores; test takers' interpretations and uses of assessment scores; and the overall consequences for test takers, the educational system, and the broader society.

In this study we investigate Chinese users' perceptions of automated speech scoring in the context of an on-line practice test for TOEFL iBT Speaking, and in imaginary scenarios of different uses of automated speech scoring. The TOEFL Practice Online (TPO) test, which uses retired TOEFL iBT test forms, has been designed to help prospective TOEFL iBT test takers become familiar with and better prepared for the TOEFL iBT test. SpeechRater v1.0, an automated scoring system, was deployed for the TPO Speaking test in 2006 to provide instantaneous score feedback on users' responses to the TPO Speaking test. SpeechRater-generated scores are expected to be used by the test takers to help them self-evaluate their readiness to take the TOEFL iBT Speaking test.

Xi, Higgins, Zechner & Williamson (2008) reported on the extensive development and evaluation efforts that support the use of SpeechRater

v1.0. However, a few critical areas of future development and research were also identified to support enhanced versions of SpeechRater. User perceptions of and reactions to SpeechRater were among the areas where research is most needed in order to strengthen the validity argument for using SpeechRater. SpeechRater is the first operational automated system for scoring spontaneous speech, which is different than highly predictable speech such as "read-aloud" speech or highly constrained spoken responses elicited from very simple tasks (e.g., "How many months are there in a year?"). Due to a lack of familiarity with speech recognition and processing technologies, its users may have deeply-rooted suspicions about whether SpeechRater can do an adequate job in evaluating human speech. In particular, research evidence is required to show the extent to which SpeechRater changes the way users interact with the tasks, and interpret and use the scores. In the automated scoring literature, user perceptions of automated scoring systems are an underexplored area, especially how the awareness of the use of automated scoring changes the way users interact with the test and how information about scoring impacts how scores are used to make different types of decisions intended by the test developer, ranging from low-stakes purposes such as for practice to high-stakes decisions such as admissions or licensure. This project will potentially advance our knowledge in this area.

8.2 Previous literature

Prior validation studies on automated scoring systems tend to emphasize one or more of three approaches: (1) demonstrating the correspondence (in both agreement and reliability) in item-level scores between automated scoring systems and human scorers, (2) examining the relationship between automated scores and criterion measures (e.g. Bridgeman, Powers, Stone, & Mollaun, 2012), and (3) understanding the construct represented within the scoring processes that automated scoring systems use (Yang, Buckendahl, Juszkiewicz, & Bhola, 2002). With few exceptions empirical studies focus on item-level evaluations and restrict their scope to one of these three approaches to validation. However, literature on the consequences of the use of automated scoring systems has been scarce, although it is widely accepted that consequences of test use should be evaluated to support the validity of an assessment. Furthermore, few, if any, empirical studies have undertaken comprehensive analyses combining different types of validity evidence to support the use of an automated scoring system, although some have

advocated the incorporation of multiple approaches into formal validity arguments for the overall assessment in a manner more consistent with a more comprehensive validation strategy (e.g. Bennett & Bejar, 1998; Clauser, Kane, & Swanson, 2002).

Xi et al. (2008) used an argument-based approach to synthesize different types of validity evidence for automated scoring. In their study, they integrated and evaluated the existing evidence to support the use of SpeechRater v1.0 in a low-stakes practice environment. The argument-based approach provided a mechanism for them to articulate the strengths and weaknesses in the validity argument for using SpeechRater v1.0 and put forward a transparent argument for using it for a low-stakes practice environment. Using this framework, they identified gaps in the existing research for SpeechRater v1.0. Specifically, the areas of research pursued included improving the prediction accuracy for the whole test-taking population and for test takers with different native language backgrounds, and expanding the construct coverage of the scoring model. Furthermore, the researchers identified a need to explore alternative criterion measures other than human scores to validate the scores provided by SpeechRater. Another potential area of investigation included users' perceptions of and interactions with this system, and the impact of users' perceptions on their decision-making based on the scores. This area is the focus of this study.

A number of researchers and theorists have warned that the use of automated scoring has the potential to both weaken face validity and produce negative washback effects. Herrington and Moran (2001) speculated that writing for a nonhuman audience may have an impact on performance, and argued that the limitations of automatic scoring may narrow the range of what can be assessed and affect the composition of the essays that are to be assessed. While making the case for the validity and reliability of an automatic essay scoring system, Landauer, Laham, & Foltz (2003) emphasize that validity arguments based on correlations with human scores must be conceived and presented with authenticity (and face validity) in mind. An automated scoring system that is open to coaching and/or forgery would significantly undermine its validity. Powers, Burstein, Chodorow, Fowles & Kukich (2001) challenged users to "stump" e-rater if e-rater were to be used to score essays and found that users with more expertise were more likely to "fool" e-rater into giving them higher scores than human raters. Algorithms have been developed in e-rater to flag responses that may show these abnormal patterns, and e-rater has been used in conjunction with human scoring

in actual applications to minimize chances of test takers gaming the system (Deane, 2013).

In an exploratory case study, Herrington and Moran (2006) interviewed ten community college students after they had completed an essay graded by an automatic scoring system. Students were not clearly informed about the scoring procedure prior to writing the essay, and only two students reported being aware that their essay would be evaluated by an automated scoring system. Six students reported that they preferred that their essay be read by people and only two students preferred the computer. After students were shown descriptors that characterized performance at various score levels, seven of ten believed that the computer would only be able to assess surface features, as opposed to the entire range of performance. While this evidence is based on a very small sample, it reflects fears about the limitations of automatic scoring.

Despite these fears and challenges, very little research has been conducted to investigate user perceptions regarding the validity and usefulness of automated scoring. In a rare study, Brent and Townsend (2006) piloted an automated essay-grading system for their sociology classes and asked students about their scores. The tasks were low-stakes class assignments, and students were encouraged to contact their professors if they believed their scores were unfair. The researchers reported that fewer than 5% of their students believed that their grades were unfair, and half of those students were concerned with "minor problems" (e.g., a phrase that was not properly recognized). Most students appreciated the immediate feedback (92%) and detailed comments (88%) that the scoring system provided, but over one-third of the students (35%) "disliked the initial grading." The study results suggest that students may be able to recognize and appreciate some of the benefits of automatic scoring despite their skepticism.

Kelly (2001) surveyed users of GRE and GMAT scores in order to investigate the consequences of automated scoring on score interpretation and use. At the time of Kelly's survey, GMAT writing scores were produced by a combination of human and computer scoring, and GRE writing scores were produced by human scoring. Kelly surveyed faculty and graduate admissions officers in the US ($N = 58$) in order to determine whether the use of automated scoring had (in the case of GMAT writing) or would have (in the case of GRE writing) an impact on score interpretation and use. Kelly found that a large percentage of score users (GMAT 64%; GRE 46%) did not think that scores produced by a combination of

human and computer scoring were treated any differently (or would be treated any differently, in the case of GRE Writing) than those produced solely by computer scoring. However, this finding was largely explained by indifference towards both GMAT and GRE writing scores by score users; admissions officers indicated that the scores were not given much weight during the admissions process. Thus, score users expressed a greater degree of skepticism towards the utility of these scores regardless of the method used for scoring. When questioned on the utility of a human/computer-scoring paradigm, most GMAT users (71%) were not concerned about issues of fairness or accuracy, although for GRE writing a large percentage of users (53%) believed that automated scoring might fail to capture important aspects of the construct (e.g., creativity). Kelly suggested that one possible explanation for the disparity between perceptions of GMAT and GRE Writing users could be that GRE users were asked to make a hypothetical judgment regarding a scoring method with which most users were unfamiliar.

8.3 Factors that may impact user perceptions of computer scoring

User perceptions of automated scoring may be very complex to unpack. The scarce empirical literature on this topic was not guided by or framed in any coherent conceptual framework that hypothesizes the factors influencing user perceptions of automated scoring, how their perceptions may impact the ways in which they interact with test tasks and interpret and use test scores, and how the particular application of automated scoring may impact the validity argument for the use of the assessment.

In this study, user perceptions of automated scoring are hypothesized to be affected by several sets of factors that are interwoven with users' beliefs about automated scoring. The first set of factors is related to features of the automated scoring system, such as the constructs the system can assess and the extent to which it can assess the full construct. The second set of factors are contextual, including the intended uses of the scores, the nature of the tasks to which automated scoring is applied, and the specific application of automated scoring (i.e., whether the automated score is being used as the sole score, a "check" score, or a second score). On the users' side, their beliefs and perceptions about computer scoring are expected to impact their behaviors, including the ways they interact with tasks scored by machine, and interpret and use the scores. Their general beliefs about computer scoring, which can be naïve and unstable, are prone to be influenced by their additional knowledge

about the specific automated scoring system that is being used, that is their interpretations about how the particular automated scoring system works and its strengths and limitations. Their evolving beliefs and perceptions, which are also mediated by the contextual factors discussed as the second set of factors, may lead to behavioral changes such as adapting their test-taking strategies to computer scoring to gain higher scores. These changing beliefs and perceptions are also manifest in the way they perceive the accuracy and utility of automated scoring.

In this study, we have designed survey and interview questions to include the two sets of factors discussed above and the user variables that interact with these factors.

8.4 Research questions

The research questions to be addressed are as follows:

1. What are the general attitudes and beliefs about human and computer scoring of speaking for a sample of Chinese TPO users?
2. How does awareness of the scores being produced by a machine impact the way Chinese users interact with the speaking tasks? How do the stakes of an assessment and the way computer scoring are used (sole rater vs. in combination with human scoring) impact the likelihood of users to try to "trick" the computer?
3. How does awareness of their scores being produced by a machine impact the way Chinese users interpret and use the scores? How does awareness of the way the automated score is used (sole rater vs. in combination with human scoring) impact Chinese users' confidence in scores?
4. How do Chinese users interpret the score report and the information about the limitations of SpeechRater and the intended use of the scores?
5. How could the SpeechRater score report and other related information be enhanced and modified to satisfy Chinese users' needs?

8.5 Methodology

8.5.1 Instruments

8.5.1.1 Survey

A survey that addresses the research questions above was designed. Both Likert-scale questions and open-ended questions were included. A small-scale pilot study was conducted with more than 200 TPO users

from various native language backgrounds taking a draft survey and three participating in phone interviews about their survey responses. The survey was revised based on results from the pilot study. Much of the original survey was retained, but several questions were adapted in order to provide clarification. A slightly shortened version of the survey, containing the questions discussed in this study and their results, is provided in Appendix 8.1.

8.5.1.2 Interview protocol

Based on the survey responses and interviews in the pilot study, we developed interview protocols to elicit more in-depth information from a small group of survey responders. Participants' survey responses were reviewed before the interviews to design questions that addressed specific issues that may have come up in their responses. During each interview, a description of how SpeechRater and automated speech scoring work was provided (Appendix 8.2).

8.5.2 Participants and data collection procedures

8.5.2.1 Survey participants

Mass e-mail invitations were sent to recent users of the TPO Speaking in China (N = 413). New Oriental Education and Technology Group, the largest provider of private educational services in China, delivers the TPO product on a regular basis using local servers to a massive number of Chinese students preparing for the TOEFL test. We received 195 responses from New Oriental-delivered TPO test users, and 47 responses from regular TPO users in China for a total of 242 responses.

As compensation, a predetermined number of respondents received a free TPO code which provided access to an additional TPO test, a value of close to US$40. In order to receive a free TPO code, the respondent was required to submit a valid survey response before the advertised deadline and was advised that the first 50 people to complete the survey by the deadline would be provided a free code once it is confirmed that they made reasonable efforts to respond to the survey questions carefully.

After checking the validity of the data (e.g., we removed responses completed within less than five minutes and those which had unreasonable response patterns, e.g., inconsistent responses to the first three survey questions regarding their general attitudes towards the accuracy of computer scoring, contrasted with human scoring), we retained 227 responses for our analyses. Thus, the response rate based on valid responses was approximately 55%.

Among these respondents, 39% planned to study Science, 25% Business, 21% Social Science and Education and 13% Humanities. The survey sample and the entire sample of TPO Chinese users invited by e-mails were roughly comparable in their scaled Speaking score distributions, as shown by the similar proportions of individuals in each of the four quartiles across the two samples (see Appendix 8.3).

8.5.2.2 Interview participants

In-depth follow-up interviews were conducted with a smaller representative sample of the users in China (N = 35). We also conducted two focus groups in Beijing and Wuhan respectively, but due to space constraints, the results are not reported in this paper. The results from the focus groups were similar to those from the interviews.

All users who completed the interview received a free TPO code as compensation. TPO users were interviewed within one month of taking a TPO Speaking test, and completed the online survey in advance of their interview. Table 8.1 summarizes the number of participants from different cities in China. The lead author conducted most of the interviews in Chinese but also trained a Chinese research assistant to assist with the interviews.

8.5.3 Data analyses

8.5.3.1 Survey data

Responses to the Likert-type questions were aggregated across all participants and were summarized as they pertained to the five research questions. Means, standard deviations, and frequency counts of responses to each survey question were produced. Given the low frequency counts in some extreme categories, five-point Likert-type scales were reduced to three points, corresponding to *negative*, *neutral*, and *positive* to facilitate group comparisons. The survey responses were compared across different academic disciplines (Science vs. Non-science: Social science,

Table 8.1 Number of interview participants from different cities in China

	Onsite interview	**Phone interview**	**Total**
Beijing	4	12	16
Wuhan	2	8	10
Shanghai	0	9	9
Total	6	29	35

Business, Humanities), and age groups (age ≤18 vs. age >18) using Chi-Square analyses to see if there were group differences in acceptance and perceptions of SpeechRater across the three categories. Answers to open-ended questions that pertain to each research question were also summarized.

8.5.3.2 Interview data

The interviews were transcribed, translated into English, coded and summarized. The summary addressed five themes, corresponding to the five research questions. A set of broad conceptual categories within each theme was identified and developed. Subsequently, two coders who are Chinese speakers were trained on the coding scheme by the lead author. Coders examined a common set of seven (20%) interviews independently. Twenty-eight interview questions (including some survey questions and a few follow-up questions) were coded. The coders agreed perfectly on the two- or three-point scale items (e.g. Questions 5, 6, and 8 in Appendix 8.1); all the discrepancies were associated with the five-point Likert-scale questions. Among all the questions, the exact inter-coder agreement was 90.3%, and the adjacent agreement was 96.9%. Discrepancies in coding were resolved through discussion. The lead coder coded the rest of the interviews.

8.6 Results

1. **What are the general attitudes and beliefs about human and computer scoring of speaking for a sample of Chinese TPO users?** Survey questions 1–4 solicited general attitudes and beliefs towards human and computer scoring of speaking. Most respondents (85.5%) indicated that they believed that humans are comparatively more accurate raters of speaking (Question 1). When asked about their general preference for computer scoring vs. an expert human rater, 73.1% of the respondents indicated a preference or strong preference for human scoring while only 12.3% preferred or strongly preferred computer scoring (Question 3).

   Question 2 asked users to indicate their confidence in computer scoring in an absolute sense (i.e., not comparing it to human scoring), provided that it was endorsed by experts, and respondents more frequently selected response categories indicating low confidence in computer scoring (40.9%; with 37.4% indicating *not very confident* and 3.5% *not confident at all*, versus 18.5% who selected *confident* or *very confident*). When asked which type of scoring would make them

Table 8.2 Participants' perceptions of human versus computer scoring

Survey question	Human scoring	About the same	Computer scoring
1. "Which is more accurate?"	85.5%	6.2%	8.4%
3. "Which do you prefer?"	73.1%	14.5%	12.3%
4. "Which makes you more anxious?"	40.6%	32.5%	26.9%

more anxious, slightly more respondents indicated that they were more anxious about human scoring (40.6%) than computer scoring (26.9%) (Question 4). Responses to questions that compared perceptions of computer and human scoring are summarized in Table 8.2, below.

In the interviews, when asked about their prior experience with other automated speech scoring systems, none indicated having used any automated speech scoring systems other than SpeechRater.

For each of the these questions, Chi-square analyses showed that there were no significant differences in the frequency counts of the three collapsed categories across gender or academic discipline groups (see Appendix 8.4 for the results).

2. **How does awareness of the scores being produced by a machine impact the way Chinese users interact with the speaking tasks? How do the stakes of an assessment and the way the computer score is used (sole rater vs. in combination with human scoring) impact the likelihood of users to try to "trick" the computer?**
Prior to taking the test a number of TPO users in China were not aware that the TPO used computer scoring (23.8%), and among those that did, only 20.6% indicated in their survey responses that they consciously changed the way they responded during the test with computer scoring in mind (Questions 5 and 6). Among those that indicated that they changed their response strategies because of computer scoring ($N = 34$), many tried hard to pronounce words very carefully (23 of 34); others tried to keep speaking even if they made little sense (19 of 34); some tried to organize their speech very carefully (14 of 34); and some spoke as quickly as they could (16 of 34). Follow-up interviews also suggested that 26.9% ($N = 7$) of those who were aware of the use of SpeechRater for scoring the TPO test ($N = 26$) attempted to change their strategies. These users reported focusing more on their fluency ($N = 4$), pronunciation ($N = 3$), organization

(*N* = 2), vocabulary and grammar (*N* = 1) and putting less emphasis on content or logic (*N* = 3) (some indicated using more than one strategy). Of those who indicated that they did not change their strategies (*N* = 25; three interviewees did not answer this question directly), they cited reasons such as "the purpose of taking the TPO is for practice rather than getting a high score" (*N* = 5); "don't want to waste an expensive practice test" (*N* = 4); "I don't know how – not clear about the scoring criteria" (*N* = 3); and "there is no use tricking a practice test" (*N* = 2) (some provided more than one reason).

Questions 16 and 18 asked users to consider two different scoring scenarios and the likelihood that they would try to "trick" the computer scoring system into giving them higher scores using unspecified strategies. Under the scenario where a computer scoring system and human rater would be used together, approximately 51.0% of users indicated that they would be *somewhat likely*, *likely* or *very likely* to try to "trick" the computer versus 48.9% who indicated *not likely at all* or *not very likely* (see Table 8.3). Interviews with users suggested that many did not fully understand the scenarios as presented in the survey, and when clarified, only 23.6% of them indicated that they were *somewhat likely*, *likely* or *very likely* to "trick" the computer under the combination of computer and human scoring paradigm, which was a substantial drop from what was found in the survey data (51.0%).

Under the scenario where only a computer would be used to score a high-stakes speaking test, approximately 57.3% of users indicated that they would be *somewhat likely*, *likely* or *very likely* to employ strategies to try to "trick" the computer (versus 42.7% who indicated *not likely at all* or *not very likely*). The interview data showed that a higher proportion of users, 67.6% would be *somewhat likely*, *likely,*

Table 8.3 Participants' likelihood to trick the computer in survey data versus interview data

		Likelihood to trick the computer				
		Not likely at all	**Not very likely**	**Somewhat likely**	**Likely**	**Very likely**
Check on human score	Survey	22.9%	26.0%	27.3%	16.7%	7.0%
	Interview	5.9%	70.6%	11.8%	11.8%	0.0%
Sole score	Survey	18.5%	24.2%	27.8%	19.8%	9.7%
	Interview	0.0%	32.4%	2.9%	47.1%	17.6%

or *very likely* to "trick" the computer under the computer-score-only scenario. The interview data also provided some insights into the reasons for their answers. The key reason for not being likely to trick the computer under the combined computer and human scoring scenario was that it was perceived as difficult to trick because there is always human scoring involved. When the computer is the only scorer, participants provided reasons such as "feasibility – can learn the tricks through practice; computer is easy to trick", "if others trick and I don't, it's unfair to me", and "practice may familiarize people with the software and help them adapt to it".

Group comparisons by age group and discipline for questions 16 and 18 using Chi-square analyses revealed that the frequency counts in the three collapsed categories were not significantly related to age or discipline (See Appendix 8.4 for the results).

3. **How does awareness of their scores being produced by a machine impact the way Chinese users interpret and use the scores? How does awareness of the way the computer score is used (sole rater vs. in combination with human scoring) impact Chinese users' confidence in scores?**

After reading the SpeechRater FAQs (see http://www.ets.org/s/toefl/pdf/toefl_tpo_faq.pdf), users were asked to indicate whether they believed SpeechRater could give them an accurate score (Question 10). A total of 21.1% indicated they were either *confident* (20.7%) or *very confident* (0.4%) in the accuracy of scores, only a slight increase from the 18.5% of respondents who indicated that they had confidence in the speaking scores provided by a computer in the general attitudes section of the survey. An additional 36.6% of the respondents reported being *somewhat confident* in the accuracy of SpeechRater scores. Thus, a total of 57.7% of respondents indicated that they were at least *somewhat confident* in the accuracy of SpeechRater scores.

Close to half of the survey respondents (47.6%) indicated that they would use scores from SpeechRater to evaluate whether they were ready for the TOEFL iBT, while 12.8% indicated they would not, and 39.6% responded that they were not sure (Question 14). The follow-up interviews provided additional insights into their rationales for using or not using TPO to assess readiness for taking the official test (24 of the 35 participants provided reasons). The most frequent rationales cited included general trust in the product and in its provider ($N = 7$), and the belief that SpeechRater scores were the most viable (or sole) source of feedback when a human rater was not available ($N = 3$). As for those who did not use SpeechRater scores

to gauge readiness, some cited general doubts about the TPO scores given different test conditions (N = 5), and others expressed doubts about the accuracy of the SpeechRater scores or about the convergence between human and computer scoring (N = 2). Interestingly, quite a few (N = 5) indicated they registered for the TOEFL iBT test before taking the TPO given the test seat capacity constraint, so they took TPO just for practice.

The survey provided several hypothetical scenarios regarding how computer scoring may be used, including as the sole scoring mechanism or in conjunction with human scoring. After reading the description of each scoring scenario, survey respondents indicated their confidence in the accuracy of the scores produced by the scenario (Questions 17 and 19).

With regard to their confidence about the accuracy of scores, 77.5% were at least *somewhat confident* about the accuracy of the scores under the condition in which the computer score is used as a check on human scoring (*somewhat confident*, 40.5%; *confident*, 34.4%; *very confident*, 2.6%). This overall percentage dropped to 68.6% when the computer score is used as the sole score (see Table 8.4). The interview data revealed a much bigger gap in reported confidence levels for these two conditions. An overwhelming majority of the interviewees (96.9%) expressed that they would be at least somewhat confident about the resultant scores under the condition that computer scoring is used as a check on human scoring. In comparison, only 23.8% expressed the same levels of confidence about the accuracy of scores under the computer-score-only scenario.

With regard to questions 10, 17 and 19, the Chi-square tests showed that there was no difference in the response patterns across

Table 8.4 Participants' confidence in score accuracy under different uses of computer scoring in survey data versus interview data

		Confidence in accuracy of scores				
		Not confident at all	**Not very confident**	**Somewhat confident**	**Confident**	**Very confident**
Check on human score	Survey	3.1%	19.4%	40.5%	34.4%	2.6%
	Interview	0.0%	3.1%	9.4%	65.6%	21.9%
Sole score	Survey	2.2%	29.2%	38.7%	27.7%	2.2%
	Interview	19.0%	57.1%	4.8%	19.0%	0.0%

the three collapsed categories by age or the field in which respondents planned to study (see Appendix 8.4 for the results).

4. **How do Chinese users interpret the score report and the information about the limitations of SpeechRater and the intended use of the scores?**

   As part of the survey, respondents read about the limitations of SpeechRater in the FAQs. A total of 63.0% of respondents indicated that they did not have trouble understanding the SpeechRater FAQs (*Understood most of it*, 45.8%; *Understood completely*, 17.2%) (Question 9). Several of the FAQs included discussions of the limitations of SpeechRater, including FAQ #4: "How is SpeechRater automated scoring different from human rater scoring?" The limitation of SpeechRater discussed in this FAQ pertained to the fact that SpeechRater did not evaluate all of the aspects of speaking that a human rater would. When asked whether this fact bothered or concerned them, 25.6% indicated that they were at least somewhat bothered by it; 59% were either *not bothered at all* or *a little bothered* by it (Question 11).

   The survey also presented information about the TPO scaled score range, and approximately half of the respondents indicated that they either *understood most of it* (31.3%) or *understood it completely* (18.1%) (Question 12). However, subsequent interviews suggested that many users interpreted the scaled score range as a prediction of their score on a TOEFL iBT Speaking test, rather than of the score that a human rater would be expected to give them based on their performance on the TPO Speaking test. Many users also regarded one- or two-point differences in TPO Speaking scores as very important, and ignored information about the range as peripheral or unimportant.

   The perception that TPO scores were intended to serve as predictions of TOEFL iBT scores persisted during interviews with users, and often influenced their positive or negative views regarding SpeechRater's accuracy. Users consistently failed to consider the predicted human score range as an important criterion in evaluating SpeechRater's accuracy and generally cited the wide range as an indication of its imprecision.

5. **How could the SpeechRater score report and other related information be enhanced and modified to satisfy Chinese users' needs?**

   After being presented with a sample TPO score report, the survey users were asked to indicate whether they were satisfied with the information provided. More than half (55.5%) indicated that they were satisfied (either *mostly satisfied* or *completely satisfied*) and only

18.5% indicated that they were *not satisfied at all* or *a little satisfied*. All respondents were asked to indicate whether each of the following four items would be useful to enhance the score report: "Feedback on pronunciation, grammar, etc." (66.1%, *Very useful*); "Suggestions for improvement" (68.7%, *Very useful*); "Scores on specific speaking items" (55.1%, *Very useful*); and "An audio recording of test responses" (55.9%, *Very useful*). Interviews found that most users were particularly interested in computer-generated feedback (93.8%). In addition, when asked about the utility of receiving exemplar responses to the TPO Speaking test they had just completed, almost all were excited about the prospect. They were also overwhelmingly positive about the prospect of having access to benchmark responses at different score levels and a rationale of why a certain response received a certain score.

8.7 Discussion

The data from the surveys and interviews, especially the comparison of the two sets of data, offer some insights into how users' perceptions of automated scoring could be unpacked. We hypothesized that user perceptions of automated scoring are impacted by their knowledge about the specific scoring system, the intended use of the assessment, the particular application of automated scoring as well as their general attitudes towards and beliefs about automated scoring. The results of the study show that these factors interacted in complex ways to influence the way the users approach test tasks, and interpret and use scores from assessments that use automated scoring. The study results also revealed that users' beliefs about automated scoring were naïve and unstable, and susceptible to influences such as additional understandings about the contextual factors for using automated scoring and about the automated scoring system itself.

Regarding the **first research question** and users' general attitudes and beliefs, a preference for human scoring over computer scoring persisted amongst the Chinese TPO users. The majority of the respondents also believed that human raters were more accurate than computer raters and indicated lower confidence in computer scoring of speaking. This sentiment was most strongly apparent in the survey questions about general attitudes and perceptions, but was also evident in subsequent discussions during interviews. The interviews revealed their lack of experience with computer scoring systems, which might be one of the reasons for their distrust of computer scoring of speech.

Regarding the **second research question** and users' interactions with test tasks because of computer scoring, when asked about whether they changed their strategies to get higher scores in a practice test because of computer scoring, most did not report changing the way they responded to the tasks as shown in both the survey and the interview data. For those who did report adapting their strategies to SpeechRater scoring, many of the strategies they reported, if actually used, may lead to performances that are negatively evaluated by human raters, such as "trying to keep speaking even if they made little sense," and "speaking as quickly as they could." However, others, such as "trying hard to pronounce words very carefully," and "trying to organize their speech very carefully," seemed to be legitimate strategies one would use in responding to any speaking tasks. In spontaneous speech production responding to speaking test tasks, second language speakers may choose to devise strategies to optimize the use of their limited cognitive capability, in order to maximize the overall quality of their speech. They are likely to make conscious decisions to focus their cognitive resources on certain aspects of their speech at the expense of others, for example, prioritizing accuracy at the expense of fluency. These decisions are typically driven by their interpretations of the scoring criteria, such as what gets valued. Their awareness of the use of automated scoring may lead them to make different interpretations of the scoring criteria than if a human rater were used, and prompt them to use different strategies. However, in the survey we did not gather information on how the participants used multiple strategies in combination and prioritized different qualities of speech in response to SpeechRater scoring, as it would have been difficult to elicit this kind of complex information through a survey.

The interview data provided some insights into the reasons why most of them did not change their strategies because of computer scoring, mostly related to their understanding that there is no use tricking the computer to get a higher score on a practice test.

When respondents were asked about their willingness to use strategies to "trick" the computer when the computer is used to score, and the score is going to be used to make important decisions about them, slightly more respondents indicated that they would be *somewhat likely*, *likely,* or *very likely* to do so to get a higher score than under the combined human and computer scoring scenario with the awareness that the scores are to be used for high-stakes decisions.

However, the interview data showed a much wider difference in users' willingness to "trick" the computer under the two computer-scoring

conditions described. Although some possibility still existed, devising strategies to "trick" the computer when it is used along with human raters to score speaking was beyond their imagination. This much greater gap could be explained mainly by the additional understanding they gained during the interviews about how SpeechRater and automated scoring of speech work, and what it means to have computer scoring used in conjunction with human scoring. In particular they may have developed a much better understanding of the limitations of computer scoring when used alone as well as its potential benefits in providing quality control on human scoring. This additional knowledge may have had an impact on how they think they would interact with an assessment that uses automated scoring under different use scenarios. This provides supporting evidence for our hypothesis that knowledge about the specific scoring system may change users' perceptions and the way they interact with an assessment.

Regarding the **third research question**, overall the survey respondents showed a reasonably high level of acceptance of using SpeechRater for scoring TPO despite some reservations. Further, close to half of them reported using the SpeechRater scores to gauge their readiness for taking the official test, which is the primary intended use of the TPO. However, this intended use was possibly compromised by the limited test seats available in China at the time of the study, so some reported taking the TPO just for practice right before taking the official test. Both are legitimate uses of TPO, and the users had reasonable confidence in the SpeechRater scores being used for these low-stakes purposes.

One theme that emerged was that as users were led to reflect more on the information provided to them and became more familiar with the process of computer scoring, their instinctive bias towards automated scoring seemed to decrease and confidence in it seemed to be boosted. This was more apparent in the in-depth interviews than survey results, as the former allowed the researchers to clarify how automated scoring of speech works, and users' misconceptions, if any. Even users extremely skeptical of computer scoring of speaking remained open to the possibility of its utility with further description and clarification. Interestingly, the survey users' confidence in the accuracy of computer speech scoring was comparable to their confidence in SpeechRater after reading the SpeechRater FAQs. This suggests that reading the FAQs may not be as effective as having automated scoring explained to the users by an expert. This points to the need for doing a better job in communicating how SpeechRater works to enhance users' confidence. A combination of text, graphics and video could be used to illustrate the inner workings of SpeechRater in a more appealing and accessible way.

The survey respondents expressed more confidence in the resultant scores when the computer score is used as a check on human scoring than when it is used as the sole score. In the interviews, after the participants gained some understanding of how SpeechRater works, their confidence levels for using the computer score as a check score boosted substantially but decreased considerably for using the computer score as the sole score, with almost all being at least somewhat confident about the scores under the condition of the computer score as a check score. This suggests that users' acceptance and confidence in automated scoring for tests used for high-stakes purposes is dependent on how automated scoring is actually used. The changes of their perceptions of automated scoring during the interviews also suggest that to promote appropriate use of automated scoring, test providers need to be transparent and communicate in accessible language both the limitations of automated scoring and its potential advantages in ensuring test score quality.

Regarding the **fourth research question** about users' interpretation of the SpeechRater score report and relevant information about SpeechRater provided to TPO users, more than half of the survey respondents indicated that they were not bothered by the limitation discussed in the FAQ that SpeechRater evaluates only a subset of the speech qualities in the human scoring rubrics.

The Chinese TPO users expressed satisfaction with TPO score reports, but retained misconceptions and overlooked important information. Although the SpeechRater score is intended to be a proxy for the score a human rater would produce for the same performance, Chinese TPO users often interpreted it as a prediction of what they would receive on a TOEFL iBT test rather than the current TPO Speaking test. In addition, many users overlooked the importance and meaning of the score range in interpreting their scores. Users that were more skeptical of SpeechRater often treated the score range dismissively. Thus, both the meaning and importance of the predicted human score range could be more emphasized and further clarified to encourage users to utilize the information it provides more effectively.

Regarding **the last research question**, most users embraced any suggestion to provide additional information in their TPO score report. The most popular suggestions to promote test preparation and further language study were *feedback on grammar, vocabulary, etc.*; *suggestions for improvement*; and *exemplar responses at different score levels*. Most users indicated that they would take this information very seriously, and some expressed surprise that it would be possible to provide any level of detailed feedback to individual users using SpeechRater. If some of this

diagnostic information could be included in TPO users' score reports, it might also serve to increase confidence in the accuracy of SpeechRater scores as score users begin to see the level of complexity in which SpeechRater operates.

We also found that Chinese TPO users' perceptions of automated speech scoring did not vary by age or academic discipline, which may be considered proxies for their knowledge and acceptance of science and technology. This shows that automated speech scoring may be a very complex phenomenon that is far beyond the understanding of most people. General misunderstandings or misconceptions of automated speech scoring are likely to persist. Automated scoring providers have the responsibility to provide accessible information to the users to improve users' understanding of automated scoring and facilitate appropriate interpretations and uses of automated scores in different contexts.

8.8 Limitations and future research

This study is, to our best knowledge, the first to investigate users' perceptions of automated scoring of speech, and the factors that influence the way they interact with a test, and interpret and use test scores. We used a combination of surveys and interviews. However, the survey was rather long and took 20–30 minutes to complete, which may have had some negative effects on the validity of the responses. The interviews provided opportunities for the users to develop a better understanding of how SpeechRater works, and clarify questions they had regarding the intent of some survey questions, and helped us better understand the patterns observed in the survey data. In some cases, the interview data yielded findings that were different than those from the survey. However, both the survey and interview data were self-reported data, and there was as much as one month of lapse between taking the TPO Speaking test and participating in the interview. So TPO users' reports of their test-taking strategies in particular may be inaccurate reflections of what they actually did while taking the test. Further, in some cases, opinions and perceptions were solicited from imaginary scenarios regarding the use of automated scoring rather than their actual experiences. Although there is evidence that the additional knowledge about SpeechRater and the imaginary use scenarios that users were exposed to during the interviews may have impacted their perceptions, one cannot rule out the possibility that direct interactions with ETS researchers (rather than independent third-party researchers) may have influenced the way they responded to the interview questions.

It is critical to continue this line of research on users' perceptions of automated scoring. In particular, a useful strand of research is to investigate test takers' interactions with test tasks and strategy use in relation to the intended uses of an assessment and the way automated scoring is used. This line of research could be conducted through the use of stimulated recalls of test takers' strategy use right after taking a test, and discourse analysis of the actual speech elicited across different conditions. Additional research on the acceptance of and confidence in automated speech scoring in different use scenarios by diverse stakeholders, including test takers, test users, and English language teachers is also critically needed to advance the field, which has focused too much on the technical quality of automated scoring.

Appendix 8.1 Abbreviated TPO speaking user survey

*Note. (Percentage of respondents selecting each answer choice is included in parentheses. $N = 227$)

1. Do you believe a computer is capable of scoring a speaking test more or less accurately than an expert human rater?
   ( ) Human is much more accurate (43.2%)
   ( ) Human is a little more accurate (42.3%)
   ( ) Computer and human are equal (6.2%)
   ( ) Computer is a little more accurate (5.7%)
   ( ) Computer is much more accurate (2.6%)
2. How confident are you that computer scoring is accurate?
   ( ) Not confident at all (3.5%)
   ( ) Not very confident (37.4%)
   ( ) Somewhat confident (40.5%)
   ( ) Confident (16.7%)
   ( ) Very confident (1.8%)
3. Which type of scoring would you prefer, computer scoring or an expert human rater?
   ( ) Strongly prefer human (21.1%)
   ( ) Prefer human (52.0%)
   ( ) No preference (14.5%)
   ( ) Prefer computer (11.0%)
   ( ) Strongly prefer computer (1.3%)
4. Which would make you more anxious: Having your speaking test scored by a computer, or an expert human rater?
   ( ) Much more anxious with human (12.8%)
   ( ) More anxious with human (27.8%)

( ) About the same (32.6%)
( ) More anxious with computer (22.5%)
( ) Much more anxious with computer (4.4%)

5. The TOEFL Practice Online Speaking test is scored by a computer system (SpeechRater). Before you took the test, did you know this?
   ( ) Yes (76.2%) ( ) No (23.8%)
6. Please think back to your responses to the TOEFL Practice Speaking test. Because a computer system was scoring your Speaking test, did you change how you responded in any way?
   ( ) Yes (20.6%) ( ) No (79.4%)
7. Please click on any strategy that you used because of computer scoring (all that apply).
   ( ) I spoke as quickly as I could. (44.4%)
   ( ) I tried very hard to pronounce words carefully. (63.9%)
   ( ) I tried to keep speaking even if I made little sense. (53.8%)
   ( ) I tried to organize my speech very carefully. (38.9%)
   ( ) Others (Please specify):
8. Before this survey, did you read the Frequently Asked Questions (FAQs) about SpeechRater?
   ( ) Yes, I read FAQs before I took the TOEFL Speaking practice test (25.6%)
   ( ) Yes, I read FAQs after I took the TOEFL Speaking practice test (16.7%)
   ( ) No, I have not previously read the FAQs (57.7%)
9. You have read the FAQs now. Were you able to understand how SpeechRater works to score your speaking practice test?
   ( ) Didn't understand at all (1.8%)
   ( ) Understood a little (11.9%)
   ( ) Understood some of it (23.3%)
   ( ) Understood most of it (45.8%)
   ( ) Understood completely (17.2%)
10. You have read the FAQs. Are you confident that SpeechRater can give you an accurate score?
    ( ) Not confident at all (1.8%)
    ( ) Not very confident (40.5%)
    ( ) Somewhat confident (36.6%)
    ( ) Confident (20.7%)
    ( ) Very confident (0.4%)
11. The FAQs explain that SpeechRater does not evaluate all of the aspects of your speaking that an expert human rater would. Did this fact bother/concern you?

( ) Not bothered at all (12.3%)
( ) A little bothered (46.7%)
( ) Somewhat bothered (25.6%)
( ) Bothered (12.8%)
( ) Very bothered (2.6%)

12. Did you understand the "Scaled Score Range" information about your Speaking section scores provided in your score report?
( ) Didn't understand at all (8.8%)
( ) Understood a little (15.0%)
( ) Understood some of it (26.9%)
( ) Understood most of it (31.3%)
( ) Understood completely (18.1%)
13. Were you satisfied with the information about your Speaking section scores provided in your score report and other related documents (the product details and the FAQs)?
( ) Not satisfied at all (4.0%)
( ) A little satisfied (14.5%)
( ) Somewhat satisfied (26.0%)
( ) Mostly satisfied (48.9%)
( ) Completely satisfied (6.6%)
14. You received scores on the TOEFL Speaking Practice test from SpeechRater. Will you use those scores to evaluate whether you have strong enough speaking skills to take an official TOEFL iBT test?
( ) Yes (47.6%) ( ) No (12.8%) ( ) Not sure (36.9%)
15. How useful would the following information be if it could be included in your score report for the TOEFL Practice Online Speaking test?

	Not Useful	**A little Useful**	**Somewhat Useful**	**Very Useful**
Feedback on your pronunciation, grammar, etc.	(2.6%)	(8.4%)	(22.9%)	(66.1%)
Your scores on specific speaking items	(3.1%)	(11.0%)	(30.8%)	(55.1%)
Suggestions for improvement	(2.6%)	(7.5%)	(21.1%)	(68.7%)
An audio recording of your test responses	(3.1%)	(15.9%)	(25.1%)	(55.9%)

For a speaking test used to make important decisions about test takers, score accuracy is typically checked by having two human raters score the same response. If the difference between the two raters' scores is unacceptable, a third rater is used to score it again.

16. If a well-designed computer system is used to check the accuracy of human scoring and identify the cases for rescoring by another human rater, how likely are you going to use strategies to "trick" the computer, hoping to receive higher scores?
    ( ) Not likely at all (22.9%)
    ( ) Not very likely (26.0%)
    ( ) Somewhat likely (27.3%)
    ( ) Likely (16.7%)
    ( ) Very likely (7.0%)
17. How confident will you feel about the accuracy of the scores?
    ( ) Not confident at all (3.1%)
    ( ) Not very confident (19.4%)
    ( ) Somewhat confident (40.5%)
    ( ) Confident (34.4%)
    ( ) Very confident (2.6%)
    Alternately, imagine that the computer system is the only rater used to score your speaking test. No human raters will be used to score it. The score is going to be used to make an important decision about you.
18. How likely are you going to use strategies to "trick" the computer, hoping to receive higher scores?
    ( ) Not likely at all (18.5%)
    ( ) Not very likely (24.2%)
    ( ) Somewhat likely (27.8%)
    ( ) Likely (19.8%)
    ( ) Very likely (9.7%)
19. How confident will you feel about the accuracy of the scores?
    ( ) Not confident at all (2.2%)
    ( ) Not very confident (29.2%)
    ( ) Somewhat confident (38.7%)
    ( ) Confident (27.7%)
    ( ) Very confident (2.2%)

Appendix 8.2 Description of how SpeechRater works

1. To score speaking it has to understand first what a speaker has said. The recognizer is trained to recognize different accents. For example, a Japanese speaker may mispronounce "road" as "load" and the computer can get trained on this and when it hears the incorrect pronunciation of "load" again it may associate it with "road".

It also relies on the words around it to confirm its guess. If the word before it is "muddy" the computer will have a higher confidence that the word is "road" not "load".

2. After the recognizer transcribes what has been said, we have designed various computer programs to analyze the speech.
   a. For example, *fluency of speech* can be analyzed.
   b. There is also a program that analyzes *pronunciation* and compares your pronunciation to that of high proficiency speakers.
   c. The *grammar* program can analyze the grammatical errors in your response. It does so by comparing the word strings in your response to those in speech by high-level speakers.
      For example, the probability of "I are" being together would be zero in the standard speech database, then the computer would be able to conclude "I are" is an error. (use "strong computer" as another example). Use one of them to save time.
   d. The *vocabulary* module can analyze whether you basically use the same words or use more varied vocabulary.
3. In order to assign scores, we use responses already scored by human raters. The computer assigns a score to each of these aspects I just talked about. But the human scores are different, they are on a 0–4 point scale. Then we try to determine the relationship between the computer scores and the human scores and come up with a formula that can convert the computer scores on different aspects to a single score that is like the score assigned by human raters, so the computer learns from the human raters to assign appropriate scores to a response.

Appendix 8.3 Proportions of individuals in each of the four quartiles of TPO speaking scores for the invited sample versus the actual survey sample

	Invited sample (*n*)	% of invited sample	Actual survey sample (*n*)	% of actual survey sample
Q1 (15–20)	95	23%	39	19%
Q2 (22)	91	22%	43	21%
Q3 (23)	91	22%	57	28%
Q4 (24–30)	136	33%	65	32%
Total	413	100%	204	100%

Note: Some responses in the survey sample did not have a corresponding TPO speaking score, and so were excluded from this table. Thus, the actual survey sample total reported in this table (n = 204) is less than the total included in the analysis (n = 227).

Appendix 8.4 Chi-square analyses of item responses by subgroups (age, discipline; *N*=227)

Question No.	Group	Result
1	Academic discipline	$X^2_{\ 2} = 2.78$, $p = .25$
	Age	$X^2_{\ 2} = 1.59$, $p = .45$
2	Academic discipline	$X^2_{\ 2} = .95$, $p = .62$
	Age	$X^2_{\ 2} = .31$, $p = .86$
3	Academic discipline	$X^2_{\ 2} = 1.57$, $p = .46$
	Age	$X^2_{\ 2} = .17$, $p = .92$
4	Academic discipline	$X^2_{\ 2} = 5.51$, $p = .06$
	Age	$X^2_{\ 2} = 2.89$, $p = .24$
10	Academic discipline	$X^2_{\ 2} = .40$, $p = .82$
	Age	$X^2_{\ 2} = .26$, $p = .88$
16	Academic discipline	$X^2_{\ 2} = 5.78$, $p = .06$
	Age	$X^2_{\ 2} = .37$, $p = .83$
17	Academic discipline	$X^2_{\ 2} = .31$, $p = .86$
	Age	$X^2_{\ 2} = .32$, $p = .85$
18	Academic discipline	$X^2_{\ 2} = 4.59$, $p = .10$
	Age	$X^2_{\ 2} = .03$, $p = .99$
19	Academic discipline	$X^2_{\ 2}$ ($N = 137$) $= .17$, $p = .92$
	Age	$X^2_{\ 2} = .79$, $p = .67$

References

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17.

Brent, E., & Townsend, M. (2006). Automated essay grading in the sociology classroom. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 177–198). Logan, Utah: Utah State University Press.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT Speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*(1), 91–108.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*(4), 413–432.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480–499.

Herrington, A., & Moran, C. (2006). Write placer plus in place: An exploratory case study. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 114–129). Logan, Utah: Utah State University Press.

Kelly, P. A. (2001). *Automated scoring of essays: Evaluating score validity* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3028998)

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education, 10*(3), 295–308.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping e-rater: Challenging the validity of automated essay scoring (Graduate Record Examination Board Professional Report No. 98–08bP). Princeton, NJ: Educational Testing Service.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335–353.

Williamson, D., Xi, X., & Breyer, J. (2012). A Framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

Xi, X. (2010). Automated scoring and feedback systems – Where are we and where are we heading? *Language Testing, 27*(3), 291–300.

Xi, X. (2012). Validity and the automated scoring of performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 438–451). New York: Routledge.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (Research Report No. RR-08–62). Princeton, NJ: Educational Testing Service.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391–412.

9

Project-Based Group Assessment in the Second Language Classroom: Understanding University Students' Perceptions

David D. Qian

9.1 Introduction

The field of education has witnessed substantial changes owing to the paradigm shift in the philosophy of teaching and learning. The traditional teacher-centred classroom now appears to have given way to a more liberal approach, known as the student-centred approach. This paradigm shift has caused tremendous changes in the way teaching, learning and assessing are conducted. In today's classrooms at the tertiary level, where the teacher-centred teaching approach traditionally dominated, the student-centred approach to learning has become increasingly common. Within this context, a large number of less restrictive assessment procedures have been introduced to second-language classrooms to replace traditional tests and examinations (Qian, 2010). Such assessment procedures are generally known as alternative assessments, which are often creative, nonintrusive and task-based. Therefore, they can be seen as an extension of day-to-day classroom activities tapping into higher-order thinking and problem-solving skills (Brown & Hudson, 1998; Richards & Schmidt, 2002). Alternative assessments typically emphasize both processes and products and are often transparent in scoring by pre-established assessment criteria. More importantly, most alternative assessment procedures call upon the classroom teacher to play a critical role in designing, coordinating and conducting the assessment, as well as making effective use of the feedback to help students further improve their language proficiency.

In this context of change, group assessment appears to be a popular assessment method. Furthermore, where this method is chosen for student evaluation, it often takes the form of project work (Brown, Bull & Pendlebury, 1997; Carless, 2005). Because of the popularity of this assessment method and owing to the impact and effects this assessment method has on students' grades, understanding the merits, demerits and potential problems of PBGA has become an important research topic in various academic disciplines. Researchers looking into this mode of assessment are trying to understand teachers' views and students' perceptions as well as the assessment processes and outcomes, so as to identify the benefits and drawbacks as a result of adopting such an assessment approach.

In the context of the Department of English in which the present study is based, PBGA is a popular means of assessment in many courses for undergraduate English majors. A typical PBGA in the English major programme starts with the formation of student groups, at which stage students normally have the freedom to choose their work partners. Once the group is formed, the group members would get together to discuss, according the expectation of the teacher as indicated in the assignment description distributed to the students, the design of the project and division of work. Students tend to choose one of the two ways of completing the project. Some groups would divide the total workload into a number of portions with each individual member being responsible for one part of the project. In this mode, group interaction is kept to the minimum and therefore a capable coordinator is highly desirable so that individual work parts can be seamlessly connected into one project through skilful coordinating and editing of the final product. The other mode for PBGA, which is often more encouraged by the teacher, is for the group to work on all the parts together. This mode tends to require frequent interactions among the group members in exploring, discussing, arguing, and collaborating in the process of creating the project and completing the project work. While this form of learning may be much more time-consuming than the individually-based group work, there is enormous potential for gaining new knowledge through working together and learning from each other. Some PBGA also involves presentations as part of the deliverable.

PBGA in this English major programme are typically assessed by one of the two methods: some teachers assign a group score for all the students in the group while other teachers may choose to give individual

scores when individual work, such as an individual presentation as part of a group presentation, is identifiable in a group project.

9.2 Literature review

As the present study focuses on student voices, the literature review in this section covers mainly research studies that investigated students' views of PBGA. Although students generally regard the method positively (Hall & Buzwell, 2012), their feedback reveals both positive and negative aspects of PBGA.

9.2.1 Students' perceptions in favour of project-based group assessment

Generally speaking, students have seen many advantages in this assessment method (Hall & Buzwell, 2012). For example, they regard workload sharing (so that each member has less work to do) and being able to learn from peers during the process of PBGA as the main advantages that PBGA can offer (Walker, 2001). They agree that PBGA provides more learning opportunities for them, especially in the aspects of seeing the importance of being able to listen to others, learning to communicate and to do teamwork effectively (Bentley & Warwick, 2013; Lima, Carvalho, Flores & Hattum-Janssen, 2007; Livingstone & Lynch, 2000). Through teamwork, they are also able to develop a range of skills for improving team performance. Such skills include generating research ideas, goal-setting, sharing views and collaborating among team members, drawing action plans and organizing, conducting research, and time management (Bourner, Hughes & Bourner, 2001; Burdett, 2003; Lima, Carvalho, Flores & Hattum-Janssen, 2007; Mills, 2003). Burdett (2003) also reports additional benefits of PBGA, namely, well-organized group work can lead to improved learning processes, improved grades, and new friendships (Walker, 2001). According to Pfaff and Huddleston (2003), factors that predicate success in PBGA include the desire to get a good project grade, absence of free-riders, a clear perception of workload, sufficient time for the project work, and adoption of peer evaluation. Generally speaking, students are willing to participate in group work and indeed enjoy working in groups (Walker, 2001). They agree that group work facilitates the development of essential project skills that are needed in their future careers (Mello, 1993). When asked about their preference when having to choose between PBGA and a traditional written exam, research in the Swedish context

indicates that students tend to prefer group assessment (Hellström, Nilsson & Olsson, 2009).

9.2.2 Problems with project-based group assessment: students' voices

However, in spite of all the above positive aspects, studies investigating students' views of PBGA has repeatedly reported serious weaknesses in various aspects during the conduct of PBGA (Aggarwal & O'Brien, 2008; Bentley & Warwick, 2013; Burdett, 2003; Gibbs, 2009; Hall & Buzwell, 2012; Lima, Carvalho, Flores & Hattum-Janssen, 2007; Walker, 2001). In particular, students perceive the uneven distribution of workload among peer group members (Bentley & Warwick, 2013; Burdett, 2003) and simplistic and unfair grading (i.e., one grade for all) by teachers (Aggarwal & O'Brien, 2008; Gibbs, 2009; Walker, 2001) as major defects of PBGA. Students complain that the levels of commitment to the assignment often vary, sometimes substantially, among group members, which leads to different levels of contributions from individual members in the same group (Mills, 2003; Walker, 2001). Furthermore, since teachers assign only a single grade for the whole group in many cases, the assessment result becomes the same for those who have contributed a lot as well as those who only made a very limited or even no contribution. This frustrates some major contributors in the group, as their efforts are not properly recognized and fairly rewarded by such a single grade (Barfield, 2003; Gibbs, 2009). In one study (Walker, 2001), students criticized that the marking scheme intended for this group project assignment failed to take into account the different levels of contribution by individual group members and the assessment was therefore unfair. Because of such uneven contributions in group projects which sometimes do not necessarily result in differentiated assessment outcomes, terms like *social loafing* and *free-riding* are thus often used to describe the phenomenon wherein some individual students make little or no contribution to the project at all but still receive the same grade as everyone else in the group. Research indicates that social loafing has been a major source of unhappiness and dissatisfaction with PBGA for many students during their group work (Aggarwal & O'Brien, 2008; Hall & Buzwell, 2012; Piezon & Ferree, 2008; Walker, 2001). In addition, difficulty in coordinating group work and finding mutually convenient time slots for group project meetings are also reported to be problems for conducting group work (Burdett, 2003; Livingstone & Lynch, 2000). This is especially true when some group members are part-time students but full-time workers (Barfield, 2003).

9.2.3 Summary of previous research findings

In summary, PBGA as an innovative form of alternative assessment has shown its popularity and vitality in many aspects. In general, students do not have objections to being assessed by PBGA although they still have reservations about some problematic practices in the process of implementing PBGA. Research has reported numerous positive elements about PBGA, such as providing opportunities for peers to learn from each other, recognizing the need to listen to others when working as a group, developing better communication skills and enhancing teamwork ability, nurturing basic research skills and being better able to understand, design and develop all parts of group projects in the process. Main negative aspects include social loafing, unfair grading if the teacher does not see the importance of assigning individually differentiated grades within a group, and difficulty in coordinating project work when group members do not have the same level of commitment or availability, difficulty in arranging project meetings when some group members also have other commitments to attend to while being students. These challenges invariably cause dissatisfaction among group members, especially among those who are more devoted members and aspire to attain high grades. These are thus seen as the main sources of unhappiness for doing PBGA.

9.2.4 Research questions for the present study

In reviewing the existing research on students' perceptions of group assessment, with a particular reference to project-based group assignment, two noteworthy issues have been revealed: 1) while there is substantial attention to issues related to PBGA, these studies were mostly conducted with research participants from academic disciplines other than language studies; 2) furthermore, little published research on this topic is actually based on Hong Kong classrooms. Having realized these research gaps, the present study aims to investigate the following research issues:

1. How do Hong Kong students generally perceive the adoption of project-based group assessment in English-as-a-second-language (ESL) classrooms?
2. Assuming students may change their perceptions in the course of study at university, how do senior-year language students' perceptions of PBGA compare with those of their freshmen counterparts?
3. From the student perspective, what are the main aspects needing improvement for a better implementation of PBGA as an alternative assessment procedure?

9.3 Research method

9.3.1 Research participants

The study was based on the results of a questionnaire survey with two groups of undergraduate students in an English department of a major public university in Hong Kong. One group of 42 students were just completing their first year of study and the other group of 20 students were completing their final year of study as English majors in the same department.

9.3.2 Research instrument

A semi-structured questionnaire was designed for this study. The questionnaire was developed based on the literature review and informal interviews with students. Before being administered to the main sample groups, the draft versions of the questionnaire were first piloted carefully for clarity and accuracy with small groups of participants who had similar backgrounds as those of the participants in the main study. Further details of the main questionnaire items (eight statements) are described in Tables 9.1 and 9.2. A four-point Likert Scale was adopted in the questionnaire to measure students' stances toward PBGA. As the participants were all English majors, both the language of the questionnaire and that of the participants' responses were in English.

There are two sections in the questionnaire. Section 1 collects background information on the respondent. The second section contains questions covering the following aspects:

1. Students' attitude towards the adoption of PBGA
2. Whether PBGA can accurately evaluate group performance
3. Whether PBGA can fairly evaluate individual contributions
4. Whether PBGA can promote students' interest in further learning
5. What are students' preferences between a traditional test and a PBGA for completing a high-stakes assessment
6. Whether the students have had pleasant or unpleasant experiences in PBGA

In the process of completing the questionnaire, the participating students were given the opportunity to ask for clarifications if they could not understand any of the statements or terms used in the questionnaire. In addition, the participants were also given an opportunity to make textual comments in the questionnaire. In the version given to the first-year group, sufficient space was provided at the end of the questionnaire so that the students could put down whatever they thought

Table 9.1 Survey results: first-year students (n=42)

No.	Statement	Strongly Agree No. (%)	Agree No. (%)	Disagree No. (%)	Strongly Disagree No. (%)
1	In general, group project is a good way of assessing students	3 (7.1)	15 (35.7)	19 (45.2)	5 (11.9)
2	Group projects can accurately assess students' learning outcomes as a group	2 (4.8)	10 (23.8)	25 (59.5)	5 (11.9)
3	Group projects can fairly assess learning outcomes of individual students	1 (2.4)	3 (7.1)	28 (66.7)	10 (23.8)
4	Project-based group assessment can promote further learning	8 (19.0)	23 (54.8)	9 (21.4)	2 (4.8)
5	If I can choose, I would prefer to be assessed through an individual project than a group project for an important assessment, such as an end-of-term exam	21 (50.0)	15 (35.7)	4 (9.5)	2 (4.8)
6	If I can choose, I would prefer to be assessed through a traditional test than a group project for an important assessment, such as an end-of-term exam	15 (35.7)	16 (38.1)	7 (16.7)	4 (9.5)
7	My experience with project-based group assessment has generally been pleasant	3 (7.1)	8 (19.0)	20 (47.6)	11 (26.2)
8	My experience with project-based group assessment has generally been unpleasant	10	21	10	1

Table 9.2 Survey results: final-year students (n=20)

No.	Statement	Strongly Agree No. (%)	Agree No. (%)	Disagree No. (%)	Strongly Disagree No. (%)
1	In general, group project is a good way of assessing students	0 (0.0)	8 (40.0)	11 (55.0)	1 (5.0)
2	Group projects can accurately assess students' learning outcomes as a group	2 (10.0)	5 (25.0)	12 (60.0)	1 (5.0)
3	Group projects can fairly assess learning outcomes of individual students	0 (0.0)	2 (10.0)	11 (55.0)	7 (35.0)
4	Project-based group assessment can promote further learning	2 (10.0)	13 (65.0)	4 (20.0)	1 (5.0)
5	If I can choose, I would prefer to be assessed through an individual project than a group project for an important assessment, such as an end-of-term exam	11 (55.0)	6 (30.0)	3 (15.0)	0 (0.0)
6	If I can choose, I would prefer to be assessed through a traditional test than a group project for an important assessment, such as an end-of-term exam	5 (25.0)	2 (10.0)	7 (35.0)	6 (30.0)
7	My experience with project-based group assessment has generally been pleasant	6 (30.0)	8 (40.0)	5 (25.0)	1 (5.0)
8	My experience with project-based group assessment has generally been unpleasant	1 (5.0)	7 (35.0)	6 (30.0)	6 (30.0)

was relevant and important. This design was modified when the questionnaire was administered to the senior-year group of students: A space for qualitative comments was provided under almost all of the eight statements. This modification was made to ensure that the participating students would be able to elaborate more specifically on the position she or he has taken with regard to each particular statement in the questionnaire so that their views could be further expressed in detail without being restricted by the quantitative design of the questionnaire.

9.3.3 Data collection

The finalized versions of the questionnaire were administered to the two groups separately: 42 first-year students completed their questionnaires in hard copy in a lecture room with a valid response rate of 91%; meanwhile, 20 senior-year students received and completed their questionnaires via email with a valid response rate of 74%.

9.3.4 Data analysis

The data analysis is composed of two dimensions: quantitative and qualitative. The quantitative analysis was performed using SPSS 22. In this part of the analysis, descriptive profiles were generated for the overall sample as well as the two cohort groups. In addition, since these were ordinal data, a non-parametric measure, Mann-Whitney U Test, was conducted to compare the means of the two groups' responses to the statements in the questionnaire. A general profile of the quantitative descriptions and comparisons of the means are provided in three tables (Tables 9.1, 9.2, and 9.3).

The qualitative analysis was carried out based on the data collected from the students' comments provided in the questionnaire. An initial glance over the qualitative data revealed that the senior-year group provided a large amount of qualitative data while the data supplied by the first-year group were limited. Therefore, the qualitative analysis had to focus more on the senior-year group.

9.4 Results: students' voices with regard to various aspects of PBGA

According to the questionnaire response data, all the 20 senior-year students have been assessed by PBGA more than five times. This basically suggests they have experienced at least one, or probably more, PBGA each semester. A follow-up discussion with some students in this group indicates that they are indeed experienced in PBGA as they have in fact

Table 9.3 Comparing group means: results of Mann-Whitney U tests

No.	Statement	Whole Sample Mean[1] (N=62)	First Year Group Mean (*n*=42)	Final Year Group Mean (*n*=20)	Mann-Whitney U Value	Sig Level
1	In general, group project is a good way of assessing students	2.37	2.38	2.35	414	0.921
2	Group projects can accurately assess students' learning outcomes as a group	2.27	2.21	2.40	368	0.374
3	Group projects can fairly assess learning outcomes of individual students	1.84	1.88	1.75	370	0.373
4	Project-based group assessment can promote further learning	2.85	2.88	2.80	396	0.678
5	If I can choose, I would prefer to be assessed through an individual project than a group project for an important assessment, such as an end-of-term exam	3.34	3.31	3.40	401	0.746
6	If I can choose, I would prefer to be assessed through a traditional test than a group project for an important assessment, such as an end-of-term exam	2.77	3.05	2.30	275	0.023*
7	My experience with project-based group assessment has generally been pleasant	2.35	2.07	2.95	207	0.001**
8	My experience with project-based group assessment has generally been unpleasant	2.69	2.95	2.15	227	0.002**

Notes:
1. A four-point Likert Scale is adopted: 4=strongly agree, 3=agree, 2=disagree, 1=strongly disagree.
2. * = significant at $p = 0.05$.
3. ** = significant at $p = 0.01$.

been assessed by this method at least ten times or more during their course of study at university. The first-year students, on the other hand, reported that they had on average between three to four PBGAs at the time they were responding to the survey, as they had just completed the first year of their study then. Therefore, the 20 senior-year students were much more experienced than the first-year group in handling PBGA.

The rest of the quantitative results of the survey are summarized in Tables 9.1–9.3. The results include the descriptive statistics (summary of response numbers and means) of the whole sample as well as the profiles of each cohort group. In addition, since the data are ordinal in nature, comparisons between the two groups were made via Mann-Whitney U Tests to determine whether there exist significant statistical differences between each pair of group means, as this may usefully indicate the different positions the two student groups take towards the adoption of PBGA in their university studies.

In the following sections on qualitative analysis, the 62 participants are numbered anonymously from S1 to S62, with S1–S42 being the first-year students and S43–S62 the senior-year students.

Statement 1: In general, group project is a good way of assessing students.
When the results of the two groups of students in Tables 9.1 and 9.2 were combined, their views on this issue appear very mixed: 58% of the students do not perceive PBGA favourably. In addition, the result of Mann-Whitney U Test indicates no significant difference between the two group means concerning this statement. Therefore, it can be concluded that the two groups of students have similar views of PBGA in general, and slightly over half of them perceive PBGA negatively. One student (S44) points out that: "For teachers, it is an easy way to manage because it will be labour-intensive to assess students one by one, but for students, their individual ability may not be reflected in group projects". However, the same student also makes the following positive comment:

> *It (PBGA) is a good exercise for equipping ourselves to future work in society because the ability to work as a team is required in many jobs. We have to learn how to deal with the reality and the real world, and group work sometimes boosts our EQ and trains us to work with different people.*

Another student makes a more balanced comment:

> *With limited time and in teachers' point of view, group project is, indeed, a good and an economical way of assessing students. Group projects also*

> *allow students to conduct a more comprehensive and in-depth research that one person couldn't handle with time and other restrictions. They can also assess students' cooperation skills. Cooperation skills are somehow as important as academic ability. However, if you are talking about an individual student's knowledge about a certain topic, group projects can rarely fairly reflect the true outcomes. (S57)*

Statement 2: Group projects can accurately assess students' learning outcomes as a group.

As shown in Tables 9.1 and 9.2, students' perceptions of this aspect have been basically pessimistic, since the majority (69%) of the 62 students do not perceive PBGA to be an accurate method for group assessment. Furthermore, the result of Mann-Whitney U Test indicates no significant difference ($p > 0.05$) in the two group means related to this statement. Therefore, it can be seen that the two groups of students basically share their views in this aspect of PBGA. Students tend to think that: "Group work may only be completed by some of the group members but not all of them if the allocation of work is unclear" (S44). Some respondents agree that PBGA can assess group work and group performance but they believe that "teachers have no way of finding out whether the project was done by one student or the whole group" (S55). One student (S57) points out that, due to the different views held, sometimes PBGA may fail to accurately assess how well students as a group understand a specific task:

> *In my experience, students in the same group might have very different stances towards a topic and opinions about the way of presenting it. Those conflicts might turn out to affect the outcomes of the project, but it doesn't mean that they didn't have a good understanding of the topic.*

Another student points out: "The teacher could only grade each student based on the finished product but there are a lot of items such as negotiations or planning which cannot be accurately reflected from the products" (S48).

Statement 3: Group projects can fairly assess learning outcomes of individual students.

Unfortunately, as also shown in Tables 9.1 and 9.2, students' views of this aspect are extremely negative: the overwhelming majority (90%) of the 62 students do not deem PBGA to be a fair method for group assessment. Since the result of Mann-Whitney U Test indicates no significant difference ($p > 0.05$) in the two group means, it appears that both

groups agree that PBGA cannot assess individual students' learning outcomes fairly because "The teacher can't accurately assess the division of labour" (S45). "You never really know the share of workload. It could be all one member's work" (S43). In some cases "students' performance can be restricted by many factors in a group, such as the attitude of other group mates" (S46). However, the focus of the negative views in this aspect largely points to the issue of free-riders in group projects because frequently students find that some of their group mates do not make as much effort as they should while others have to work extremely hard on their own in order to finish the whole project for the group because they want a good grade (S52, S53, S56, S60, S62).

Nevertheless, some students are more insightful. A senior-year student contends: "A group project should be seen as a whole project with cohesion but not a combination of individual work because it is understood that individual work is not easily assessed in group projects" (S50).

Statement 4: Project-based group assessment can promote further learning. Students have, as shown in Tables 9.1 and 9.2, more favourable views in this aspect: 46 (74%) students believe that PBGA is conducive to further learning, and the result of Mann-Whitney U Test relating to this statement indicates no significant difference ($p > 0.05$) in the two group means. Therefore, it is safe to state that a great majority of the students believe that PBGA can promote and are therefore conducive to further learning. They reckon group projects motivate students to learn because "Students are more motivated to learn as a group" (S45) as they often need to carry out their own research as a group (S49) and in order to get better results, they "have to research for extra materials" (S56). In this process of learning, they "will try to find more relevant resources such as literature paper or reference books to extend their topics in the project. Thus, further learning is promoted in this way" (S52); students can also help each other and learn from each other in the process (S50).

However, one student (S43) has reservations about this statement as she believes that only when group members are mutually supportive can this assessment method promote further learning. Furthermore, there are students who take more negative stances as they "cannot find any reasons why assessment can initiate one to learn" (S58). On the contrary, group projects sometimes might restrict students' performance or learning due to discrepancies in the personalities and abilities among group members (S48). One student (S57) has made a good point in his argument:

> *I personally can't see how project-based group assessment can directly promote further learning. I would say interesting topics promote further*

> *learning. As long as students find a topic interesting and worth learning, they would spend more time and effort in understanding it.*

Statement 5: If I can choose, I would prefer to be assessed through an individual project than a group project for an important assessment, such as an end-of-term exam.

As revealed by Tables 9.1 and 9.2, the situation concerning this aspect does not look encouraging: An overwhelming majority (85%) of the 62 students show that they would prefer to be assessed through individual projects rather than group projects when the stakes are high. Since the result of Mann-Whitney U Test indicates no significant difference ($p > 0.05$) in the two group means, it appears that the overwhelming majority in both groups would prefer to be assessed by individual projects rather than group projects. This is certainly a strong warning that the implementation of PBGA has not been popular with university students. There seem to be many reasons why students would prefer to be assessed by individual projects. Clearly, fairness appears to be a main reason as some students explain that: "Assessing through an individual project can ensure the project is done by the student himself and it can assess the learning outcomes of the individual student more fairly" (S55, echoed by a number of other students including S51 & S56). "I truly believe one should get what he/she has paid for the assessments. So the learning outcome of the one is directly tested from exams or individual papers" (S48). As expected, some students tend to perform better with individual projects than group projects, as S44 puts it: "I usually do better in individual work because I can make my own decision more efficiently." S47 offers a different reason: "I enjoy doing group projects, but I would also like to be given chances to challenge myself, not being hindered or hindering the others". However, the following comment sounds more academic:

> *Individual project can truly reflect how much the students have digested and absorbed the concepts and are able to apply in the projects since they are working on the projects fully on their own without the help of others.* (S53)

But for students who prefer to be assessed by group projects, they also have their own reasons: "Sometimes it's nice to be at liberty to 'share' work the way we desire" (S43). Another student (S62) gives a very different, and very thought-provoking, reason for supporting group projects: "I prefer giving out ideas and let others execute, not everyone learns in class but some people may still be helpful."

Statement 6: If I can choose, I would prefer to be assessed through a traditional test than a group project for an important assessment, such as an end-of-term exam.

Again, as confirmed by Tables 9.1 and 9.2, the situation of this aspect does not look rosy: as many as 38 students (61% of the whole sample of the 62 students) are even more willing to be assessed by traditional tests when the grades are high-stakes. This tendency seems especially strong with the first-year group, in which 31 of the 42 students prefer to be assessed by traditional tests. Interestingly, the result of Mann-Whitney U Test has returned a statistically significant difference ($p < 0.05$) this time, due to the fact that many more first-year students than senior-year students prefer to be assessed by traditional tests.

Students have provided various reasons. One senior-year student (S48), who has indicated a strong agreement to this questionnaire statement, prefers tests because she does not like group projects. In a group project, "not everyone would like to work hard and there are always free-riders". A similar reason is provided by a fellow senior-year student (S53): "Traditional tests can ensure students will contribute to the assessment so that they will not rely on their group members." However, some students give a different set of reasons for supporting tests: "A test can accurately assess the students' understanding of the course" (S55); "Compared with a group project, an examination is better able to assess how much the students have learnt throughout the semester" (S56).

Doubtless there are students who prefer to be assessed by group projects. The reasons provided for supporting group projects vary greatly. One student (S43) chooses group projects because she hates tests: "I'm never good at them (tests)", she admits. "I will panic and become very nervous in exams. I don't think my ability can be assessed using one single test", said another student (S44). A more rational argument was given by S45: "I personally prefer projects over tests, because it is based over a longer period instead of performance during a single test." Student S50 dislikes tests because "There are more constraints in tests than in projects, such as time limit and stress." Another type of reason for rejecting tests is based on the nature of English major. "The subjects we study simply aren't designed for exams" (S62). Two other students echo this reason:

> *I don't think a traditional test is the only fair way to assess individual students. More importantly, not many courses in the language department*

are suitable for using traditional tests to assess students. Individual assignments might be better than tests. (S57) *I don't think some of the subjects in our programme can be assessed through a traditional test.* (S61)

Finally, a student who has chosen "disagree" for Statement 6, puts forth a more balanced view: "It depends on what the subject is. In my opinion, more academic subjects can make use of traditional tests while practical subjects can make use of group projects" S47).

Statement 7: My experience with project-based group assessment has generally been pleasant.

Statement 8: My experience with project-based group assessment has generally been unpleasant.

As reported in Tables 9.1 and 9.2, based on the responses to Statements 7 and 8 jointly, the picture of this dimension looks quite mixed even though the number of students who have had unpleasant experiences in PBGA (31 or 74% in first-year group; eight or 40% in the senior-year group) far exceeds the number who have had pleasant experiences (11 or 26% in first-year group; 14 or 70% in senior-year group) in PBGA. It is noteworthy that the first-year group appear more negative in perceiving their PBGA experiences. It should be admitted, however, that the numbers of responses to Statements 7 and 8 do not match perfectly, possibly due to the fact that a neutral point was not provided on the Likert Scale used in the questionnaire so that some students had to choose affirmative, or negative, answers for both Statements 7 and 8 as their experiences were mixed with both kinds of feelings. Separately, the results of Mann-Whitney U Test confirm that there exist statistically significant differences ($p < 0.05$) in the means of the students' responses. Nevertheless, it can be seen that, for Statement 7 concerning pleasant PBGA experiences, the senior-year group has a much higher group mean (2.95 out of 4.00) than the first-year group, which only scores a mean of 2.07; on the other hand, on Statement 8 focusing on unpleasant PBGA experiences, the first-year group returns a very high group mean (2.95) in comparison with the mean of 2.69 of the final-year group.

Reporting on pleasant experiences, a student (S43) reveals that her PBGA experience has been pleasant mainly because "I choose my team wisely most of the time." Another student (S44) has had similar experience: "I am generally lucky to have some good teammates who all are willing to work. We usually have even work allocation which everyone

understands and agrees with." Student S45 confides: "Team projects are fun. You also get to know your classmates better after working with them." A student feels happy about her PBGA because she received almost straight A grades for her group project work (S55). However, another student (S47) tells a more balanced story:

> *I could not say they* (PBGA experiences) *were perfectly pleasant as we argued sometimes. But I enjoyed the assessment because the group members I grouped with were always reliable, which I believed to be very lucky to meet them in my university life.*

A student (S49) shares her secrets of success:

> *Generally it depends on how well you know the ones you are working with. Most of the time these relationships are established gradually from Year 1 study. From time to time you will know who else would be a great partner for you and who else would be the last person you would like to see in your team. I am fortunate to have a group of people I can trust and cooperate in any projects without too many differences, since if there is too much disagreement and argument within a team, not only will it directly affect the outcome of the project but also waste everyone's time on handling these matters and harming the relationship between each other.*

Students reveal that there are some factors that are important in ensuring the successful completion of group projects, such as good communicative skills, ability to work as a team, willingness to listen to other teammates and accept others' suggestions (S51), and equal distribution of work (S53).

Speaking of unpleasant experiences, a major factor causing this feeling seems to be the presence of free-riders. Three students recalled their stories, which, although they differ in some ways, all have a focus on the "free-rider":

> *It was a subject about TESOL, where every team has to produce a modified textbook, lesson plan and scheme of work. There was a free-rider in the group which increased other teammates' workload.* (S50)
>
> *Sometimes, it is not easy to work with lazy group members. They usually did not contribute to the group work at all or did not show up during group meetings. At the end, the rest of us worked for his part as well, but the "free-riders" could still get the same grade as ours.* (S53)
>
> *For two or three times, there was a group mate who didn't contribute much to the projects but he presented well in the group presentation and got*

> *a good grade. Other teammates and I believed that it was not fair because he overran and presented most of the content prepared by us.* (S44)

Another source of unhappiness is related to the quality of group project work: "Some students did not have their work done on time. Some finished with a lot of mistakes" (S50). In addition, different personalities and work styles could also be sources of unpleasantness, as one student points out:

> *People with different personality and working style from me are always the sources of unpleasant group assessment experience. I can't force them to work hard or work in any preferred ways as they always have their own thoughts and things to be considered. Group projects always drive me crazy.* (S48)

Finally, an interesting case with S56 is indeed worth mentioning as she has had only one pleasant experience but a lot more unpleasant experiences:

> *I have had a pleasant experience once. The group members were responsible and we had clear division of duties. Everyone was able to follow the schedule and did the tasks well. We researched a lot of useful information and collected insightful real-life data. We also got some time to talk about the topic face-to-face so as to share our information. Finally, the project was coherent. We worked together to improve the whole project and we did quite well during the presentation. I found the whole project was really inspiring since we discussed the topic in-depth. (S56)*

However, on many other occasions, this student was so frustrated with group project work. It is alarming considering the total number of times the student has had to be involved in PBGA over the whole duration of her university study:

> *I have a lot of unpleasant experiences with group projects, which made me doubt the value of group work. The group members were completely irresponsible and claimed that they were too busy to have a short meeting. Nobody followed the working schedule. Though there was a clear division of duties, the members did not finish their work. When the deadline was coming soon, I could not contact them by person or even by phone. At last, I finished the whole PowerPoint on my own. When asking for a rehearsal before the presentation, nobody was willing to do so. Finally, for the sake of better grades, I even wrote the script for them since they did not know the content well.*

9.5 Discussion and conclusion

9.5.1 Addressing the research questions

The discussion section here aims to provide answers for the three research questions, which are first stated in Section 2.4.

1. *How do Hong Kong students generally perceive the adoption of project-based group assessment in English-as-a-second-language (ESL) classrooms?*
   Students' views of PBGA, regardless of which year of study they are in, are mixed. Since about 58% of the participating students hold negative views of PBGA, it might be appropriate to state that the majority of the participating students in the present study do not view PBGA favourably.
2. *Assuming students may change their perceptions in the course of study at university, how do senior-year language students' perceptions of PBGA compare with those of their freshmen counterparts?*
   The results of the analysis based on the response data to Statements 6, 7 and 8 indicate that the senior-year students view PBGA more favourably than their first-year counterparts. Based on this finding, I suggest that, depending on their experiences with PBGA at university, students may change their attitudes toward PBGA in a positive direction; they may dislike PBGA when they were in their first year of university study, but may gradually be more willing to accept this assessment method after a few times of experiencing positive encounters with PBGA. Therefore, whether or not teachers are able to handle the PBGA skilfully and appropriately will greatly influence students' attitudes toward PBGA.
3. *From the student perspective, what are the main aspects needing improvement for a better implementation of PBGA as an alternative assessment procedure?*
   Several issues are prominent in relation to addressing this research question. These issues all seem to be structural problems or inherent weaknesses of PBGA and many of them were also identified in previous research (Aggarwal & O'Brien, 2008; Bentley & Warwick, 2013; Burdett, 2003; Gibbs, 2009; Hall & Buzwell, 2012; Lima, Carvalho, Flores & Hattum-Janssen, 2007; Walker, 2001). These issues definitely need teachers' earnest attention if teachers want to see PBGA welcomed by their students.
   a. The present study confirms findings from previous research on PBGA that there exist free-riders in many project groups and this

problem seems difficult to solve unless the teacher takes some measures to combine group and individual assessments in PBGA.

b. As suggested in previous research, the busy schedules of some students may prevent students from finding a common time slot for group meetings, which has a negative impact on the successful progression and completion of group projects.
c. Different personalities of team members may be a negative force at work during a group project as this may make team members less accommodating when they do not see eye to eye with each other. Therefore, it is important to select teammates carefully so that teammates can see eye-to-eye with each other.
d. Relating to the above point, different views of teammates may also be an obstacle to a smooth and successful completion of a group project. While it is healthy to have arguments during a group project, willingness to accommodate different views seems to be key to the successful completion of a project. This again points to the importance of choosing teammates wisely. In that respect, perhaps senior students, knowing their classmates much better than their first-year counterparts, can do much better and gradually develop a more favourable view of PBGA.
e. A good factor on the students' part is their desire to get a good grade. This motivation can be an important driving force for keeping the project going even under difficult conditions. But students also see the wise choice of group members an important element for getting a good grade, as one student expresses: "For project-based assessment, especially group projects, the results very much depend on who you work with. If you are grouped with responsible students, you are lucky" (S41).
f. Finally, an extremely important issue is for the teacher to handle PBGA wisely. This particularly includes the skilful use of the marking scale for PBGA. Simply assigning a group grade may be easy for the teacher but in most cases will result in bad feelings among students when there are free-riders in the group. Previous research has also pointed out this problem. In some cases, even when there are no free-riders in the group, because of the uneven distribution of workload, some students will feel short-changed if the whole group receives the same grade. Therefore, a combination of group and individual assessments within a PBGA would in most cases be a better solution than to simply assign a group grade, especially to be fair to those group members who have made the greatest contributions to the project.

9.5.2 Preferences for assessment methods

As a former British colony, Hong Kong is heavily influenced by a strong British examination culture as well as a long history of examinations in China, where examinations are often of very high-stakes for selection and promotion purposes (Qian, 2008). In such an environment, individual assignments and assessments are naturally the norm throughout the local students' primary and secondary school years. Therefore, students in Hong Kong are generally more accustomed to tests and examinations than to alternative assessment methods. Even when alternative assessments are introduced into secondary-school classrooms in recent years, teachers still assert a dominant role in such assessments. For example, in group assessments, the groups are often prescriptive; teachers tend to decide who will work with whom in which group. Students do not really have the freedom to choose who they would like to work with. As a result, students tend to adopt a passive role in group assessment. This mentality may have influenced students' attitude towards PBGA when they first enter university, which is probably why the first-year group took a more negative view towards PBGA in the present study. In a way, it is surprising to see that an overwhelming majority (85%) of the participating students have indicated that they would prefer to be assessed through individual projects rather than group projects when the stakes are high, and as many as 61% of the participating students, including 31 (74%) first-year students, are even more willing to be assessed by traditional tests than by group projects when the grades are of high-stakes. This finding does not corroborate the finding from the research in the Swedish context (Hellström, Nilsson & Olsson, 2009). This new finding suggests that a great amount of work needs to be done to improve the image and mechanism of PBGA before the method can be readily and widely accepted by university students in Hong Kong. From their early years on, students need to be convinced that PBGA is an effective and fair way of assessing students so that they will be willing to accept the assessment procedure. Simply adopting PBGA in order to catch the trend of *Assessment for Learning* will not be very helpful if teachers are not aware of the inherent structural problems of PBGA that may negatively affect the implementation of the assessment procedure.

Acknowledgements

This study was financially supported by a research grant (G-YH65) from The Hong Kong Polytechnic University. The author would also like to

thank all the students who participated in this study and especially Enishi Lung for his assistance in collecting the data from the senior-year students.

References

Aggarwal, P., & O'Brien, C. L. (2008). Social loafing on group projects: Structural antecedents and effect on student satisfaction. *Journal of Marketing Education, 30* (3), 255–264.

Barfield, R. L. (2003). Students' perceptions of and satisfaction with group grades and the group experience in the college classroom. *Assessment & Evaluation in Higher Education, 28* (4), 355–370.

Bentley, Y. & Warwick, S. (2013). Students' experience and perceptions of group assignments. http://journals.heacademy.ac.uk/page/STEM_2013_Information; http://journals.heacademy.ac.uk/doi/book/10.11120/stem.hea.2013. University of Birmingham; *Proceedings of the HEA STEM Learning and Teaching Conference: Where Practice and Pedagogy Meet, 2013,* 80–84.

Bourner, J., Hughes, M., & Bourner, T. (2001). First-year undergraduate experiences of group project work. *Assessment & Evaluation in Higher Education, 26* (1), 19–39.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education.* London: Routledge.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32* (4), 653–675.

Burdett, J. (2003). Making groups work: University students' perceptions. *International Education Journal, 4* (3), 177–191.

Carless, D. (2005). Prospects for the implementation of assessment for learning. *Assessment in Education, 12* (1), 39–54.

Gibbs, G. (2009). *The assessment of group work: Lessons from the literature.* Centre for Excellence in Teaching and Learning in Higher Education, Brooks University, UK. Accessed August 30, 2014 from: http://www.brookes.ac.uk/services/ocsld/group_work/brookes_groupwork_gibbs_dec09.pdf

Hall, D., & Buzwell, S. (2012). The problem of free-riding in group projects: Looking beyond social loafing as reason for non-contribution. *Active Learning in Higher Education, 14* (1), 37–49.

Hellström, D., Nilsson, F., & Olsson, A. (2009). Group assessment challenges in project-based learning – Perceptions from students in higher engineering courses. *Proceedings of 2:a Utvecklingskonferensen för Sveriges Ingenjörsutbildningar*, Lund, December 2009.

Lima, R. M., Carvalho, D., Flores, M. A., & Hattum-Janssen, N. V. (2007). A case study on project led education in engineering: students' and teachers' perceptions. *European Journal of Engineering Education, 32* (3), 337–347.

Livingstone, D., & Lynch, K. (2000). Group project work and student-centred active learning: two different experiences. *Studies in Higher Education, 25* (3), 325–345.

Mello, J. A. (1993). Improving individual member accountability in small work settings. *Journal of Management Education*, 17 (2) 253–259.

Mills, P. (2003). Group project work with undergraduate veterinary science students. *Assessment & Evaluation in Higher Education, 28* (5), 527–538.

Pfaff, E., & Huddleston, P. (2003). Does it matter if I hate teamwork? What impacts student attitudes toward teamwork. *Journal of Marketing Education, 25* (1), 37–45.

Piezon, S. L., & Ferree, W. D. (2008). Perceptions of social loafing in online learning groups: A study of public university and US Naval War College students. *International Review of Research in Open and Distance Learning, 9* (2). Accessed August 30, 2014, from: http://www.irrodl.org/index.php/irrodl/article/view/484/1034.

Qian, D. D. (2008). English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing, 25* (1), 85–110.

Qian, D. D. (2010). Alternative assessment procedures for language classrooms: Perceptions of frontline school teachers. In: Kao, T., Lin, Y. (Eds.), *A new look at language teaching and testing: English as a subject and vehicle* (pp. 273–294). Taipei: Language Training and Testing Center.

Richards, J. C., & Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics*. London: Pearson Education.

Walker, A. (2001). British psychology students' perceptions of group-work and peer assessment. *Psychology Learning and Teaching, 1* (1), 28–36.

10
Chinese EFL Students' Response to an Assessment Policy Change

Qiuxian Chen and Lyn May

10.1 Introduction

The relationship between assessment and learning has been an enduring focus of research in educational contexts (Broadfoot, 2007). There is now a consensus that assessment and student learning are mutually and inextricably linked: it is clear that assessment shapes students' learning. This has been amply documented in research from a variety of educational contexts. Crooks (1988), for instance, in an extensive review, suggests that assessment plays a key role in influencing students' learning in multiple aspects, which range from the ability to retain and apply what has been learned to the development of students' self-efficacy as learners. Assessment was found to have considerable impact on how educational courses are perceived (Marton & Säljö, 1997), conditioning learning goals (Boud & Falchikov, 2007), shaping students' approaches to their learning (Ramsden, 2003), deciding on the quality of the learning outcomes (Biggs, 1999) and even the development of the students' future learning (Struyven, Dochy, Janssens, Schelfhout & Gielen, 2006). It is therefore unsurprising that assessment has been used in many educational contexts by policy makers as a "tool" (Hamilton, 2003) to effect changes in pedagogy and learning. This use has become more pronounced in the past decade, as the development of assessment theory has brought formative assessment and its potential to enhance learning into focus (Carless, 2011; Ross, 2008). The assessment initiative that the Department of Higher Education of the MoE (2007) proposed via the *College English Curriculum Requirements* (CECR) is one innovation of this type in the area of English as a Foreign Language (EFL).

While it is important to acknowledge that learning is an extremely complex process involving a range of participants and stakeholders

(Messick, 1996; Tang & Biggs, 1996), in the enactment of classroom assessment practices, it has been claimed that the pivotal role lies with students (Alderson & Banerjee, 2002; Bailey, 1996). Assessment in this sense becomes the "hidden curriculum" (Snyder, 1971), in response to which students "play the examination game" (Miller & Parlett, 1974). Studies further reveal that the way in which students view the nature, purpose and specific procedures of assessment directly relates to how they respond in their approaches to learning (Ramsden, 2003), and links to the quality of learning outcomes (Brown & Hirschfeld, 2008). Weaver (2006) concluded that it is students who decide if education change can possibly be truly realised. Therefore, this study takes up students' voices and their experience of the changed assessment policy and practice. The purpose is to identify, from the students' perspective, factors that afford or constrain the effectiveness of the formative assessment change initiated in the CECR, in terms of positioning them in a more active role in learning and assessment. The findings are intended to lead to a more nuanced understanding of the potential of formative assessment in the Chinese College English context.

10.2 Literature review

Formative assessment positions student engagement as the key to effective and improved learning (Broadfoot, 2007). This is because learners, according to constructivist theory, from which formative assessment is derived, are active meaning-constructors (Nuthall, 1997). Through participation in learning as well as assessment, learners develop self-confidence, identity and a capacity to monitor and manage their own learning (Broadfoot, 2007). This capacity, known as "meta-cognitive skills" or "meta-cognition" is regarded as crucial to the success of learners in the future, as it embodies learners' gains in terms of skills and the ability to "learn to learn" (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Broadfoot, 2007). Learner engagement or participation therefore is seen as a priority in formative assessment-related reforms.

In sharp contrast to the active role that formative assessment initiatives require students to play is the widely-held perception that Chinese students are relatively passive and reticent in the classroom (Dautermann, 2005; Ginsberg, 1992). They speak in class only when invited by the teacher (Carless, 2011), are reluctant to provide peer assessment and are actually sceptical of the value of peer feedback (Chen, May, Klenowski & Kettle, 2014; Hu, 2003).

The reasons for these manifestations of a less active approach to learning are largely cultural. To begin with, the Chinese culture of learning

has a tradition of positioning students as "listeners" or "recipients" and the teacher as an authority (Hu, 2002). In line with this understanding, the teacher is the sole judge and assessor, whereas others such as peers, students themselves and even parents are not perceived to have the credibility and authority to judge or assess students' work (Cortazzi & Jin, 1996; Hu, 2002). The traditional valuing of harmony in Chinese culture also tends to influence the Chinese students' behaviour in the classroom. Carson and Nelson (1996), for example, found that Chinese students showed reluctance in initiating comments and expressing disagreement in group discussions because of their concern for social harmony within the group. In addition, Hu (2003) foregrounded the role of students' prior learning experience and socioeconomic factors. Specifically, Hu (2003) found that the students from less developed regions demonstrated a notable reluctance to engage in group work, a lack of learner autonomy, lower adaptability and considerable resistance to a communicatively-orientated, learner-centred learning context and pedagogy when compared to those from economically more developed regions. These findings point to the complexity of implementing formative assessment, where the student is positioned as an active participant in learning and assessment, and thus could present serious challenges to the implementation of the CECR.

More recent studies, however, indicate that as multiple educational reforms have been introduced to the Chinese context, changes have occurred (Chan & Rao, 2009). Shi (2006), for example, found that Chinese middle-school students in Shanghai oriented to active participation in classroom interactions. Meyer (2003), too, revealed that although a preference for receiving feedback from the teacher remains, Chinese university students generally expressed positive opinions regarding feedback from peers. This change is echoed in Miao, Badger, and Zhen (2006), who found that although teacher feedback was adopted relatively easily and led to more improvement, both teacher and peer feedback were valued. Thus Miao et al. (2006) stated "even in cultures that are said to give great authority to the teacher, there is a role for peer feedback" (p. 179). These findings point to changes that could help to facilitate the realisation of the CECR policy intent on enactment in the classroom.

10.3 Context and background

10.3.1 CECR and its impact on teaching and learning

The CECR was formally issued and implemented by the Department of Higher Education of the MoE in 2007. As the name suggests, this document is designed to be a curriculum guideline for College English,

an English as a Foreign Language (EFL) programme designed especially for non-English major tertiary students in Chinese higher education institutions. This programme is compulsory in the Chinese universities. The potential influence of the CECR on teaching and learning English language is thus substantial.

Compared to previous guideline documents, the National College English *Teaching Syllabus* (Revision Team on College English Teaching Syllabus, 1986, 1999; Working Group/Revision Team on College English Teaching Syllabus, 1985), the CECR introduced several major innovations, which include "individualisation", and "autonomy" (Wang, 2010). In terms of assessment, the CECR states that: "College English should be assessed both formatively and summatively" (Department of Higher Education of the MoE, 2007 p. 6). The inclusion of formative assessment in this policy document initiated a profound change in the assessment of College English, which was previously comprised entirely of a summative achievement test at the end of each term and a national standardised testing system, the College English Test (CET) Bands Four and Six. The purpose of "enhancing English language learning" (Department of Higher Education of the MoE, 2007 p. 6) is explicitly stated in the CECR, which demonstrates the potentially transformative nature of this initiative.

This policy intent is particularly salient, given that "low efficiency" has been an issue of concern to College English educators for decades (Dai, 2001; Li, 1997). The official report that the Department of Higher Education of the MoE issued on two rounds of pilot implementation is quite encouraging. The largely positive findings include a more active involvement and an increase in morale on the part of both teachers and students, with the report concluding that the prospects for the implementation of this initiative in practice are promising (Department of Higher Education of the MoE, 2006). However, the negative washback on teaching and learning includes an increasingly heavy workload for teachers and the difficulty for teachers in catering to students' individual learning needs (Xue, 2006; Zhou & Qin, 2005). Moreover, the appropriateness of this initiative, which is borrowed from Anglophone contexts, is of concern (Carless, 2011; Chen, 2009; Chen, Kettle, Klenowski & May, 2013).

10.3.2 Assessment in the Chinese context

China has an examination tradition traceable back to the keju system in the Sui Dynasty (605 AD), when the monarchy used it to select civil officials. For over 2000 years the practice of examination and the use

of results for purposes such as increased social status and individual advancement have greatly influenced Chinese people's beliefs about education and schooling. Of the many influences, as Han and Yang (2001 p. 5) maintain, are "a stress on the key role that examination plays in education and an emphasis on the one-off result to the neglect of formative assessment". Thus the function of examinations are accepted in Chinese society as "a fair indicator of students' academic success" and hence "the goal of teaching and learning" in general (Cheng & Curtis, 2009a, p. 267). Today, as examinations continue to be used as a gatekeeper for selective purposes at virtually every layer of education, employment and promotion, the impact of examinations remains substantial (Cheng & Curtis, 2009b). The well-known saying that "tests are teachers' magic weapon; scores are students' lifeblood" [考考考，老师的法宝；分分分，学生的命根 kǎokǎokǎo, lǎoshīdefǎbǎo; fēnfēnfēn, xuéshēngdemìnggēn]" clearly illustrates the powerful influence of examinations in the Chinese context. This situation is described in Cheng and Curtis (2009a p. 267) as: "highly-valued, highly-selective, rather narrowly-defined and examination-driven".

To survive and excel within this "relentless and harsh" (Watkins & Biggs, 2001, p. 3) assessment climate, Chinese students have been trained or naturally oriented to the so-called "learning for tests" mode. As success in tests is "the key to succeed in study, work and life" (Cheng & Curtis, 2009a, p. 267), performance in tests, rather than substantive engagement in the process of learning, is the focus of most students. Chinese learners tend to take an instrumental attitude towards study and view high academic scores as goals for their efforts in study (Gao, Zhao & Cheng, 2007; Wang & Cheng, 2009); and they usually try to understand the learning materials in ways that they perceive will meet requirements, and adopt the most convenient approaches to maximise their grades (Tang & Biggs, 1996). That is, a surface approach for improvement and high-level cognitive skills is a striking feature in their behaviours (Ho, Peng & Chan, 2001). When they do foreground learning, it is often to serve the purpose of better performance in examinations. Qi's research (2005, 2007) demonstrated that the forces of examinations are so powerful in this context as to undermine the intended washback function of a national test to effect beneficial changes in teaching and learning. Chen (2009), on examining the current assessment realities in China, maintained that the dominance of tests and the resulting product-oriented learning approach are at odds with the principles of formative assessment and might therefore compromise the policy intent. A concern arising from this potential

disconnection is the extent to which the formative assessment initiative in the CECR could achieve its intended goals.

10.3.3 A culturally contextualised model

Emphasising the influence of localised settings and sociocultural contexts on the uptake of formative assessment practices, researchers have suggested a flexible approach to implementation. In their influential guidelines for formative assessment, Black et al. (2003) recommended that changes should be brought about incrementally. Pryor and Crossouard (2008) advocated that teachers and students take up appropriate variations of formative assessment in accordance with their situated contexts. Davison and Leung (2009) proposed a continuum of possibilities for formative assessment to be carried out in English Language Teaching (ELT) contexts. After closely examining the assessment and cultural characteristics of Confucian Heritage Culture (CHC) contexts, Carless (2011) suggests teachers in this context adopt a more extended form that suits the contextual realities rather than try to replicate the original form originating from Anglophone contexts.

Taking a sociocultural perspective, Chen et al. (2013, 2014) investigated the implementation of formative assessment in two Chinese universities at both institutional and classroom levels. They found the emergence of a culturally-adopted and contextually-situated form of formative assessment: "process assessment" (Chen et al. 2013). That is, varied proportions of College English assessment (10 per cent in one university and 60 per cent in another) were allocated to areas such as student participation in classroom activities, assignments and attendance. Teachers, according to their records, decide upon a grade, which is known as the "process grade" and used together with students' performances in a final-term exam for reporting purposes (Chen et al., 2013).

10.4 A case study approach

This study builds on the work of Chen et al. (2013, 2014) and aims to explore the potential impact that the CECR formative assessment initiative has on students and their approaches to EFL learning. This study focused on students for two reasons: first, students and their learning are the focus of the policy changes; second, it is students who play a vital and decisive part in the enactment and realisation of the policy. Particularly, it seeks to address the following research question:

How have Chinese university students responded to the changed assessment policy in terms of its positioning of them as more active participants in learning and assessment?

Aiming to explore the Chinese EFL students' experience of the implemented CECR assessment initiative, this study takes on an exploratory design and a case study approach. This approach allows for in-depth investigation into a research problem (Yin, 2003).

The case chosen for this study is a provincial university located in the northwest part of China. This university has a history of over 100 years. It enrols 3000 students each year, and is of medium size and medium ranking among the 2000 Chinese higher education institutions. In response to the CECR initiative, this university has allocated 10 per cent of College English assessment to the process of learning. The 10 per cent given to the process assessment encompasses three areas: attendance (3 per cent), assignments (3 per cent) and classroom participation (4 per cent). The remaining 90 per cent is based on performance in the original end-of-term achievement test and a speaking test, which is conducted individually.

The researcher approached a College English teacher, Lina, in this university, and with her permission, surveyed 100 students in two of Lina's classes. Altogether 93 surveys were completed, of which 91 were valid. The survey data was entered into SPSS16. The survey (Appendix 10.1) elicited details of students' backgrounds, and required students to respond to statements regarding the impact of the assessment policy change on their English language learning.

A primary analysis of the survey data shows a very high percentage of students considered themselves in terms of their involvement in classroom activities as "inactive" or "very inactive" (40.2 per cent), with 48.3 per cent of students describing themselves as "moderately active". Only 11.5 per cent of the surveyed students claimed that they participated actively in the classroom and none of the students considered themselves as "very active". While these findings appear to reflect the stereotypical perception of Chinese students' tendency to be relatively passive learners (Ho, Peng & Chan, 2001), it is important to note that as the survey entailed the self-reporting of levels of participation the students may not feel comfortable in categorizing themselves as "very active". Thus it was imperative to gain, through interviews, a more nuanced understanding of the perspectives of the students in order to explore the factors leading to their participation in class.

Thus informed, the researcher observed six continuous sessions of Lina's classes. Based on the classroom observations and in consultation with the teacher, the researcher selected four students from this class to interview, according to the degree of engagement they exhibited in classroom activities (with two students being categorised as active

participants and two as relatively inactive). In this sense, both the survey and the observation data were collected to help select the four informants.

The rationale for this selection was that student participation in classroom activities is currently one of the three major areas prescribed for process assessment. The four students were chosen according to their degree of participation in classroom activities during the six sessions of observation. The basis on which students were categorised is as follows: active students were observed to be attentive, engaged in class activities and consistently responded to the teacher's questioning. Moreover, they were always seated in the front rows, quite close to the teacher, which facilitated their communication with the teacher. The relatively inactive students, in contrast, remained silent, did not seem engaged in class activities and rarely responded when the teacher asked questions in class. They chose to sit at the back of the classroom, as far away as possible from the teacher. As the classroom has 120 seats and class size is 50, the students' decision to sit at the back made communication with the teacher difficult.

The researcher consulted the teacher, who confirmed the researcher's observation of these students. For the sake of anonymity, the four students were given the following pseudonyms: Ying (Active Student 1), Shan (Active Student 2), Lan (Inactive Student 1) and Ming (Inactive Student 2). The student demographic information including gender, age and the level of economic development of the area they are from is detailed in Table 10.1 below. The level of economic development was included to reflect the documented imbalance in the level of English education in China (Department of Higher Education of the MoE, 2007). This information is regarded as vital to the quality of English education the students received prior to their university study, and hence could be an important indicator of students' language levels, especially oral proficiency levels.

Table 10.1 Information on the student interviewees

Student	Participation profile	Gender	Age	Economic categorisation of the schooling area of students
Ying	Active	Female	20	Moderately developed area
Shan	Active	Male	20	Underdeveloped area
Lan	Relatively inactive	Female	19	Underdeveloped area
Ming	Relatively inactive	Male	20	Developed area

The four selected students were interviewed individually using an interview schedule (Appendix 10.2), which focused on the students' understanding of and response to the assessment policy change. This encompassed the influence of the changed policy on their approaches to learning. The interviews were conducted in Chinese, the first language of the interviewees, to ensure effective communication, and later transcribed and translated into English. Additional data sources were an interview with Lina and the researcher's notes taken while conducting the classroom observations.

10.5 Findings

10.5.1 Profiles of two active students and their participation

Both of the active students, Ying and Shan, sat in the first or second rows in the classrooms, quite near to Lina's desk. They were observed to respond to Lina's questions and also impressed the teacher as such. When asked about her participation in class, Ying agreed that she was an "active" learner, and commented:

> *I actively respond to the teacher's questions. … It is a kind of habit for me. Plus, I like taking the front seat in the classroom, which makes it convenient to communicate with teachers with both language and eye contact.*

Shan, however, did not perceive himself to be an "active" student. Instead, he attributed his response to the teacher's assessment practices: "the teacher usually ticks off [in her record sheet] if students speak out in class. So I sometimes try to respond when she asks questions". Shan also admitted that he was sometimes unable to respond because of difficulties in listening comprehension and expressing his ideas clearly: "it is embarrassing to admit that sometimes I can't understand what the teacher has said at all and when I actually want to participate I find myself struggling to organise what I want to say". For Shan, the overwhelming motivator was extrinsic: "I need to pass and graduate."

It appears the assessment change has impacted on Shan, but not Ying. The students' level of English proficiency is a factor in this. Ying, for example, prides herself on being a high-achieving student in English: "To be frank, my English has stood out from the very beginning [of my EFL learning] and remains so". Shan, on the other hand, sees English as his weakest subject: "English has been pulling my overall grades and ranking in the league table down all these years". He was especially

worried about his perceived weakness in speaking and listening, which he related to his earlier English learning experience:

> *As the speaking and listening test results are not included in the National Matriculation English Test (NMET), our teacher told us to spare our efforts [to focus on aspects of language included in the test].*

Through Shan's experience, we can see the washback of the absence of listening and speaking from the NMET. The narrowed curriculum that Shan, who was from an ordinary school in a small county of the local province, was subjected to was very different from that experienced by Ying, who came from the capital city of a neighbouring province and had been to a well-respected high school there. The differing attitude of the two students to their current learning context lies, at least in part, in their prior EFL learning experiences.

The orientations that these two students had to the role of assessment in their EFL learning could be due to another reason. Ying understood assessment as aiming "to improve English language learning", and believed that learning was paramount: "I think if you learn well, it doesn't matter whatever and however it is tested". Shan, in contrast, perceived the imperative to "pass rather than fail" as his goal, thus exhibiting a testing orientation.

When asked about how he would respond if greater weighting were given to process assessment, Shan replied: "I would definitely put more effort into speaking and listening". In contrast, Ying did not believe that this would change her current orientation to learning: "Not much [change from before], I suppose. I'd participate a lot as always".

To sum up, the incorporation of process assessment has impacted on Ying and Shan, both observed to be active participants, to differing degrees. While the impact on Shan's participation was substantial, Ying appears to have been minimally impacted. This difference can be attributed both to their proficiency levels and the extent to which they adhere to a learning or testing orientation. While Ying seemed to genuinely enjoy both the interactions with her teacher and peers and furthering her EFL learning, Shan's participation appeared to constitute a challenge to him, and was driven by the need to fulfil the criteria in order to gain the marks awarded for "participation".

10.5.2 Profiles of two relatively inactive students and their participation

Lan and Ming were observed to be relatively inactive in class. Ming, in all the sessions observed, took a seat in the back row of the classroom, as

far away as possible from the teacher. He seldom looked up, appearing to be preoccupied by reading from a book on his desk. When asked about this in the interview, Ming explained that he was looking through a dictionary, which is his preferred method of learning English. Lan sat in the far left corner of the classroom and was never observed to speak in class. In one session in which the teacher organised a group discussion activity, she was observed as "not participating" and "sat alone in the far corner through to the end of the session" (Observation notes). When asked about this during the interview, Lan explained that this particular day was as "an exception" and she attributed her lack of participation to her "poor oral English". She acknowledged that she did not often respond to the teacher in class, but maintained: "I'm not so inactive".

Lan had attended one of the best high schools in the local province, and felt that the school had provided adequate opportunities for English listening and speaking practice. Yet being a low-achieving student in English and knowing that English carried little weight in the NMET, she "did not spend much time on it". She recognised that her personality also played a part in her behaviour in class, describing herself as "A little introverted … it is difficult for me to talk with unfamiliar people in English".

Interestingly, Lan did take part in some English language learning activities outside of class: "Last Friday, I went to the English Corner and I talked with some girls from my school … I think it is not so bad to talk with them in English". She also mentioned her other experience of using English in the dorm: "We [Lan and her roommates] like watching English movies, and we use actors' lines from the movies when we feel like speaking in English or cracking a joke with each other." That is, non-participation is not actually the norm for Lan. She did involve herself, but only in activities in which she felt comfortable. This finding conflicts with Lan's behaviour in class during the observed sessions, and thus provides us with a deeper understanding of the complex factors underpinning an individual student's participation in class activities.

Ming, from a developed city in a coastal province in China, described himself a high-achieving student in English. He actively pursued his EFL learning by making foreign friends and chatting with them online, which has effectively "cultivated [his] language sense … and [enabled him to] learn beautiful pronunciation." Therefore it seems that a low level of listening and speaking proficiency, the reasons that have hindered Shan's and Lan's participation, should not apply to him. Ming attributed his non-participation in class to a lack of interest in the topics that the teacher put up for discussion and a dislike of the way that the English classes were taught. From his perspective,

"since I am not interested [in the activities in the EL class], I certainly will not take the trouble to do it". Instead, he showed a preference for practising English in more authentic situations: "Actually, I often go to places where foreigners are around and try to talk with them". This was confirmed by the teacher, Lina, who mentioned observing Ming speaking enthusiastically to foreign teachers in the English Corner. Ming related his approach to learning English to the achievement of his own goals:

> *I learn English because I want to communicate freely with people from other countries … I want to go abroad to have a look around the world after I graduate from university … Tests or exams are never my goal.*

Ming's ultimate desire to really use the language and travel help to explain his behaviour in class, as the activities did not seem to be able to meet his language learning needs or help him to achieve his individual goals.

Regardless of the different reasons for their participation in class, both Lan and Ming predicted they would behave differently in class if more weighting was given to process assessment. The excerpt below from Lan's interview is representative:

> *If more weight in the overall assessment is given to classroom performance, I will try better to participate, for the sake of grades. As we are told from our early years on, grades are what matter, at least it is so in China. It is no exception for me, and I do not see any sign of change.*

In summary, personality factors and the disconnection between individual learning orientations/goals and language teaching pedagogy influenced Lan and Ming's participation in class. However, both students felt that they would participate more if a higher percentage were assigned to process assessment, confirming the substantial influence of assessment policy on learning in this context.

10.6 Discussion

It is clear that a range of sociocultural factors influenced the extent of student participation and engagement in their English language class activities. The economic development of the region where students came from had impacted on some students' English proficiency, and oral English proficiency in particular. For example, Ying and Ming,

who are from relatively more developed regions and highly-regarded high schools, are both high-achieving students and do not experience problems communicating in English. In contrast, Shan and Lan, from disadvantaged areas, have only limited English proficiency. This in a sense confirms Hu's (2003) finding of the impact of imbalanced development of economy on English education in China. Nonetheless, this factor does not really explain the students' different behaviours in class or their different responses to the assessment policy change, since high-achieving students can be either active (Ying) or relatively inactive (Ming), and low-achieving students can force themselves to be active (Shan) or opt out of participating (Lan).

Individual learning style is the key factor that differentiates the two high-achieving students' (Ying and Ming) classroom behaviours. Ying, who values learning from interactions with the teacher, is always an active participant in class, whereas Ming prefers to learn from a dictionary or from what he regards as more authentic situations, and thus he opted out of participating in class. Ying, a low-achieving student, also opted out of participating in class, and instead chose to learn English in the manner and with the people she felt comfortable. Individual learning style, accordingly, appears to be an important factor influencing students' participation in class. These three students are similar in that, from the sentiments expressed in the interviews, they do not appear to be responsive to the current assessment policy change. Thus it is important for educators and policy makers to understand the pedagogical implications of students' preferred learning styles if students are expected to participate and engage more substantively in learning and assessment.

The feature that distinguishes Shan and makes him the only student responsive to the assessment policy change is his focus on test results, or a testing orientation. Driven by the desire to pass and the worry about his low English proficiency, Shan sought a strategic way out: since what is explicitly valued in the process assessment grade is participation, but not necessarily the quality of that participation, he forced himself to participate at a superficial level. In other words, he has adopted the most convenient approach to maximise his grades as the students in Tang and Biggs' study (1996) did, and has played "the examination games" that Miller and Parlett (1974) had documented. In this sense, Shan is the one who has been most obviously affected by the inclusion of participation in the assessment mix.

The percentage allocated to participation within the process assessment grade, as the analysis reveals, is a factor that has the potential to

bring about changes in classroom participation for most students, given that three out of the four students interviewed (Shan, Lan and Ming) indicated that they would more actively engage in the classroom if a higher weighting was allocated to this. This finding delivers a message for policy-makers at the institutional level: a higher weighting for process assessment is needed if this policy is to effect a change in students' approach to language learning. The present practice of awarding only ten per cent of the total grade to process assessment, whether intentionally or not, sends a message to students on the extent to which their participation is truly valued.

The findings from this study to some extent support the "reticent" profile of Chinese learners (Carless, 2011; Cortazzi & Jin, 1996; Dautermann, 2005; Ginsberg, 1992; Hu, 2003), with 40.2 per cent of the surveyed students describing themselves as "inactive" in class. However, it is important to acknowledge that students such as Ying and the 11.5 per cent of surveyed students who categorised themselves as "active" represent a profound change in the Chinese culture of learning, which Chan and Rao (2009), Shi (2006) and Meyer (2003) have documented.

10.7 Conclusion

From the findings, it can be said that the Chinese students interviewed in this study are, overall, not totally responsive to the assessment policy change as implemented in this university. Of the four students interviewed, only one was influenced by the policy change to involve himself more in classroom activities. However, three of the students claimed that they would participate more if a greater weighting were given to this, reflecting the washback potential of the policy as a lever for change (Hamilton, 2003).

The factors influencing students' responses are complex and interconnected. These factors include students' English proficiency level, their previous English language learning experiences, their individual learning style and the extent to which they could be described as primarily learning or testing-oriented.

From an institutional perspective, the findings indicate that if policy-makers at the institutional level aim to effect a bigger change in students' EFL learning approaches and further learning outcomes, a larger weighting to participation in process assessment is needed. Nonetheless, it is more important for the policy-makers at the institutional level to be aware that it is rather simplistic or wrong, in a sense, to equate the practice of allocating a certain percentage of the total

score to process assessment with the formative assessment initiative (Chen et al., 2013).

Teachers would benefit from professional development designed to enhance their EL pedagogy and assessment literacy, so that they can cater for different learning styles and create more authentic and engaging learning and assessment tasks. Otherwise, students will continue to opt out of participating, rather than actively engaging in the classroom community of learning.

While the policy-makers' as well as teachers' understanding of formative assessment affects to a large extent the effect of the assessment policy change, it is more important for teachers to encourage students to participate actively in classroom activities and help them develop learner autonomy. This is what formative assessment is meant to achieve (Assessment Reform Group, 2002) and the CECR intended for. Of course, if participation is to be included in the assessment mix, it is also important to clearly define "participation" and have meaningful criteria through which to assess it.

Students, too, need to understand the rationale and mechanics of the assessment change: particularly salient is that they are being positioned to take a different role in the new assessment regime. They need to be supported in order to meet these changing expectations. A top-down policy, without being fully understood by the key stakeholders involved, will experience difficulty in realising its original intent.

Acknowledgements

Data used in this chapter are from those that the first author collected for her doctoral studies. Thanks go to the sponsor, Shanxi Scholarship Council (Project No.: 2012–019) and others who participated or helped in data collection.

Appendix 10.1 Questionnaire for students' views and responses

Section 1. About you (Tick please)

1. I am:
   - ☐ female
   - ☐ male
2. I am from a:
   - ☐ developed region
   - ☐ moderately developed region
   - ☐ underdeveloped or rural region

3. I would classify my involvement and participation in classroom activities as:
   ☐ very active ☐ active ☐ moderately active ☐ inactive ☐ very inactive

Section 2. Your views and responses to College English assessment

Please give your immediate response to every answer.
1- strongly disagree (SD), 2 -disagree (D), 3 -neutral(N), 4 – agree (A), 5 -strongly agree (SA)

1. Formative assessment is a way to facilitate English learning.	1 2 3 4 5
2. Formative assessment is an alternative way to gain grades.	1 2 3 4 5
3. My English has improved as a result of the incorporation of formative assessment.	1 2 3 4 5
4. Involvement in assessment is not helpful for my English learning.	1 2 3 4 5
5. I adjust my approaches to learning depending on the mode of assessment.	1 2 3 4 5
6. How English learning is assessed does not affect my approach to learning.	1 2 3 4 5

Appendix 10.2 Student interview schedule

Section 1. About you

– Where are you from? Is that a developed region or …?
– Can you tell me about the high school you went to?

Section 2. About your views and responses to College English assessment

1. Please describe your understanding of the ways your College English learning is assessed. Are there any differences from the assessment you have experienced before?
2. How would you describe your involvement in College English classroom activities?
3. Does assessment usually influence your approach to College English learning? If yes, how?
4. How do you respond to the process assessment practice in this course?

5. Does the greater emphasis on the process of learning affect your approach to College English learning? If yes, how?
6. What are the purposes of College English assessment from your point of view?

References

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, *35*(2), 79–113.

Assessment Reform Group. (2002). *Assessment for learning: 10 principles*. Retrieved January 10, 2015 from http://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing, *Language Testing*, *13*(3), 257–279.

Biggs, J. B. (1999). *Teaching for quality learning at university.* Buckingham: The Open University Press.

Black, P., C. Harrison, C. Lee, B. Marshall, & Wiliam, D. eds. (2003). *Assessment for learning: Putting it into practice*. Maidenhead: Open University Press.

Boud, D., & Falchikov, N. (2007). Assessment for the longer term. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 3–25). London and New York: Routledge.

Broadfoot, P. (2007). *An introduction to assessment*. New York: Continuum.

Brown, G. T. L., & Hirschfeld, G. H. F. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education*, *15*(1), 3–17.

Carless, D. (2011). *From testing to productive student learning: Implementing formative assessment in Confucian heritage settings*. New York: Routledge.

Carson, J. G., & Nelson, G. L. (1996). Chinese students' perceptions of ESL peer response group interaction. *Journal of Second Language Writing*, *5*(1), 1–19.

Chan, C. F. S., & Rao, N. (2009). *Revisiting the Chinese learner: Changing contexts, changing education*. Hong Kong: Springer.

Cheng, L., & Curtis, A. (2009a). The impact of English language assessment and the Chinese learner in China and beyond. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 268–273). New York & London: Routledge.

Cheng, L., & Curtis, A. (2009b). The realities of English language assessment and the Chinese learner in China and beyond. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 3–12). New York & London: Routledge.

Chen, Q. (2009). The potential barriers to College English assessment policy change in China: A sociocultural perspective. In B. Garrick, S. Poed, & J. Skinner (Eds.), *Educational planet shapers: Researching, hypothesising, dreaming the future* (pp. 115–126). Brisbane: Post Press.

Chen, Q., Kettle, M. Klenowski, V. & May, L. (2013). Interpretations of formative assessment in the teaching of English at two Chinese universities: A sociocultural perspective. *Assessment & Evaluation in Higher Education, 38*(7), 831–846.

Chen, Q., May, L., Klenowski, V. & Kettle, M. (2014). The enactment of formative assessment in English language classrooms in two Chinese universities: Teacher and student responses. *Assessment in Education, 21*(3), 271–285.

Cortazzi, M., & Jin, L. (1996). Cultures of learning: Language classroom. In H. Coleman (Eds.) *Society and the language classroom* (pp. 169–206). Cambridge and New York: Cambridge University Press.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(1), 438–481.

Dai, W.[戴炜栋] (2001). On further improving English language learning in China: Suggestions for consideration [外语教学的'费时低效'现象-思考与对策]. *Foreign Languages and their Education* [外语与外语教学], *7*, 1–6.

Dautermann, J. (2005). Teaching business and practical writing in China: Confronting assumptions and practices at home and abroad. *Technical Communication Quarterly, 14*(2), 141–159.

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, *43*(3), 393–415.

Department of Higher Education of the MoE. (2006). Survey report on the progressive situation of College English Reform Program [关于大学英语改革进展情况的调查报告]. Retrieved January 10, 2015 from http://wenku.baidu.com/link?url=V-TNWO0C4I8xsjPrgIGSE_9v-0XgUFMYyaZ_RHrWbUTO4c9v6qc7MZ_aKGqVBpO2LW8LHXnzGp0P50SLGRticTOdfPSvtF6C9wbjWT02jPa.

Department of Higher Education of the MoE. (2007). *College English curriculum requirements* [大学英语课程要求]. Shanghai: Shanghai Foreign Language Education Press[上海外语教育出版社].

Gao, Y., Zhao, Y. & Cheng, Y. (2007) Relationship between English learning motivation types and self-identity changes among Chinese students. *TESOL Quarterly*, *41*(1), 133–155.

Ginsberg, E. (1992). Not just a matter of English. HERDSA News, *14*(1), 6–8.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education*, *27*(2), 25–68.

Han, M., & Yang, X. (2001). Educational assessment in China: Lessons from history and future prospects. *Assessment in Education*, *8*(1), 5–10.

Ho, D. Y., Peng, S., & Chan, F. S. (2001). Authority and learning in Confucian-heritage education: A relational methodological analysis. In C. Chiu., F. Salili & Y. Hong (Eds.), *Multiple competencies and self-regulated learning implications for multicultural education* (pp. 29–48). Greenwich: Information Age Publishing.

Hu, G. (2002). Potential cultural resistance to pedagogical imports: The case of communicative language teaching in China. *Language, Culture and Curriculum*, *15*(2), 93–105.

Hu, G. (2003). English language teaching in China: Regional differences and contributing factors. *Journal of Multilingual and Multicultural Development*, *24*(4), 290–318.

Li, L. [李岚清]. (1997) Primer Li Lanqing's talk on foreign language education conference [李岚清副总理在外语教学座谈会上的讲话]. *Jiangsu Foreign Language teaching and Research* [江苏外语教学研究], *9*(2), 53–53.

Marton, F. & SÄLJÖ, R. (1997). Approaches to learning. In F. Marton, D. Hounsell & N. Entwistle (Eds.), *The experience of learning: Implications for teaching and studying in higher education* (pp. 39–58). Edinburgh: Scottish Academic Press.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(2), 241–256.

Meyer, J. E. (2003). PRC students and group work: Their actions and reactions. In L. G. Ling, L. Ho, J. E. Meyer, C. Varaprasad & C. Young (Eds.). *Teaching English to students from China* (pp. 73–93). Singapore: Singapore University Press.

Miao, Y., Badger, R., & Zhen, Y. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing*, *15*(3) 179–200.

Miller, C. M. I., & Parlett, M. (1974). *Up to the mark: A study of the examination game*. Guildford: Society for Research into Higher Education.

Nuthall, G. (1997). Understanding student thinking and learning in classroom. In B. J. Biddle, T. Good & I. Goodson (Eds.), *International handbook of teachers and teaching* (Vol. 3, pp. 681–768). Dordrecht: Kluwer Academic Publishers.

Pryor, J., and B. Crossouard. 2008. A socio-cultural theorisation of formative assessment. *Oxford Review of Education*, *34*(1), 1–20.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(1), 140–173.

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education*, *14*(1), 51–74.

Ramsden, P. (2003). *Learning to teach in higher education*. London and New York: Routledge.

Revision Team on College English Teaching Syllabus. (1986). *National College English teaching syllabus for Arts and Science students*. Beijing: Higher Education Press [大学英语教学大纲（文科）北京:高等教育出版社].

Revision Team on College English Teaching Syllabus. (1999). *National College English teaching syllabus for liberal arts students* [大学英语教学大纲（文科）]. Shanghai: Foreign Language Education Press[上海:上海外语教育出版社].

Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing*, *25*(1), 5–13.

Shi, L. (2006). The successors to Confucianism or a new generation? A questionnaire study on Chinese students' culture of learning English. *Language, Culture and Curriculum*, *19*(1), 122–147.

Snyder, B. R. (1971). *The hidden curriculum*. Cambridge: MIT Press.

Struyven, K., Dochy, F., Janssens, S., Schelfhout, W., & Gielen, S. (2006). The overall effects of end-of-course assessment on student performance: A comparison between multiple choice testing, peer-assessment and portfolio assessment. *Studies in Educational Evaluation*, *32*(3), 202–222.

Tang, C., & Biggs, J. (1996). How Hong Kong students cope with assessment. In D. A. Watkins & J. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences* (pp. 159–182). Hong Kong & Melbourne: The Central Printing Press.

Wang, S. [王守仁](2010). On deepening the reform in college English teaching in China [全面、准确贯彻《大学英语课程教学要求》深化大学英语教学改革]. *Foreign Languages in China[中国外语]*, *7*(2), 4–7.

Wang, X. & Cheng, L. (2009) Chinese EFL students' perceptions of the classroom assessment environment and their goal orientations. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 202–218). New York & London: Routledge.

Watkins D. A. & Biggs J. B. (2001). *Teaching the Chinese learner: Psychological and pedagogical perspectives*. Hong Kong/ Melbourne: CERC & ACER.

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education*, *31*(3), 379–394.

Working Group/Revision Team on College English Teaching Syllabus (1985). *National College English Teaching Syllabus (For College and University Students of Science and Technology)* [大学英语教学大纲（理工科）]. Shanghai: Foreign Language Education Press[上海:上海外语教育出版社].

Xue, M. [薛默] (2006). A study of formative assessment in College English teaching and learning [大学英语教学中的形成性评价研究]. Unpublished master's thesis, Jilin University[吉林大学], Jilin, China.

Yin, R.K. (2003). *Case study research: Design and methods*, 3rd ed. London, New Delhi: Sage.

Zhou, P., & Qin, X. [周娉娣 & 秦秀白] (2005). The application of formative assessment in multimedia computer assisted language learning. [形成性评估在大学英语网络教学中的应用]*Computer-assisted language learning in China* [外语电化教学] 5, 6–10.

11

Students' Voices: What Factors Influence Their English Learning and Test Performance?

Ying Zheng

11.1 Introduction

Various factors come into play in affecting Chinese university students' English language learning and their test performance. There is no shortage of such studies that have examined the association between individual learner characteristics and language learning behaviours or outcomes. What is needed are empirical investigations of those associations in various localized second/foreign language learning contexts to provide a focused and targeted understanding of what indeed are the profound influencing factors.

This study collected students' voices from a qualitative study and aimed to explore what students considered as important factors that influence their English learning and test performance on College English Test Band 4 (CET-4). The major interest of this investigation focused on learners' motivation in learning English, including both internal forces and external influences. The study also explored possible changes in learners' motivation and their awareness of the importance of English in the current globalizing world.

11.2 Relevant concepts and studies

The affective factor that has attracted the most interest is learner motivation. However, motivation is multi-faceted concept that different researchers interpret differently. In this study, motivation is "the process whereby goal-directed activity is instigated and sustained" (Pintrich & Schunk, 2002, p. 5). Motivation has at least two fundamental features: directionality and intensity, that is, a desire to obtain certain goals and the amount of effort spent on attaining these goals. In second/foreign

language learning, motivation is usually operationalized as a combination of desires to learn a certain language, degrees of motivational intensity, and attitudes toward learning that language (Gardner, 2006).

The integrative aspect of orientation in learning a second/foreign language is associated with a positive disposition toward the second language (L2) group and the desire to interact with the target language community. Integrativeness is defined as "a genuine interest in learning the L2 in order to come closer psychologically to the other language community" (Gardner, 2001, p. 12). Instrumental motivation is usually defined as "potential pragmatic gains of L2 proficiency" (Dörnyei, 1994, p. 274), such as getting a promotion or a better-paying job.

Global awareness, a construct developed specifically for this study, reflects English learners' awareness of the importance of English as an international language in their particular learning contexts and their perception of that importance in their life. This concept is built upon the idea of situating integrative motivation in the current globalizing world (Csizer & Dörnyei, 2005).

Various researchers have conducted their research in attempts to identify and understand the factors that influence students' language learning. A few relevant empirical studies are reviewed here to set this particular study in its relevant background. For example, researchers have examined integrative motivation for learning a foreign language in the globalizing world in different contexts, namely, Kormos and Csizer (2008) and Csizer and Kormos (2008) in Hungary; Gan, Humphreys, and Hamp-Lyons (2004) in China; Hernandez (2008) in the USA; and Lamb (2004) in Indonesia.

Specifically, Kormos and Csizer (2008) examined motivations for learning English as a foreign language in three distinct learner populations in Hungary: secondary school pupils, university students, and adult language learners. The main factors affecting students' L2 motivation were language learning attitudes and the ideal L2 self, which provides empirical support for the main construct of the theory of the L2 motivational self system. The results demonstrated that models of motivated behaviour varied across the three investigated learner groups: for university students, as well as for adult language learners, "international posture" was an important predictive variable, instead of interest in English-language cultural products among secondary school pupils.

Furthermore, Csizer and Kormos (2008) examined the role of intercultural contact in the motivation of Hungarian learners. They used motivated learning behaviours as the outcome measures. According to Dörnyei (2005), motivated learning behaviour, one of the most

important antecedents of achievement in language learning, is defined as "effort expended to achieve a goal, a desire to learn the language, and satisfaction with the task of learning" (p. 6). Csizer and Kormos' (2008) results showed that these behaviours were determined not only by language-related attitudes, but also by the views of students about the perceived importance of contact with foreigners. The results of the study revealed that the perceived importance of contact was not related to students' direct contact experiences with target language speakers but was influenced by the students' milieu, that is, the social influence of the learners' immediate environment (parents' support and friends' attitudes toward L2 learning) and indirect contact with foreign media usage. Among the contact variables, it was only contact through media products that had an important position in the model examined, whereas direct contact with L2 speakers played an insignificant role in affecting motivated behaviour and attitude. Csizer and Kormos (2008) pointed out that this finding highlighted that, in a foreign language setting like Hungary, indirect contact by means of exposure to English-language media products, such as television, magazines, and the Internet, might take over the place of direct contact and might exert significantly more influence on attitudes to target language speakers and their culture than direct spoken contact.

Different levels of success in language learning among students may be explained by a complex and dynamic interplay of two aspects: the cognitive aspect and the affective aspect of language learning (Gan, Humphreys & Hamp-Lyons, 2004). The latter refers to emotion self-management in dealing with the positive and negative side effects that may be associated with language learning. A larger number of learning or practice activities used, and a more sophisticated use of strategies by the successful students, as compared to those employed by the unsuccessful students, might be related to the former's overall English learning goal, which was characterized by their apparent emphasis on a practical command of English. The majority of successful students in the study by Gan et al. seemed to be motivated both externally and internally.

Little (2007) pointed out that the ability to take charge of one's own learning characterizes learner autonomy. Independent or autonomous language learning has mainly been associated with Western educational settings; it is sometimes perceived to be more problematic in an Asian context, such as in Indonesia (Lamb, 2004). Hernandez (2008) identified integrative motivation as a significant predictor of scores on a simulated oral proficiency interview as well as of students' final exam grades.

He also found that integrative motivation was a significant predictor of students' desire to enrol in additional coursework after completing a four-semester foreign language requirement. It also had an important role in students' intention to major in the language. A negative relationship was found between the need to fulfil the language requirement and intent to continue with further studies.

To recapitulate, the importance of looking into the effects of motivation in the current globalizing world lends support to a closer investigation of motivation in different contexts. The purpose of this study was to investigate to what extent and in what ways interviews with students served to contribute to a comprehensive and nuanced understanding of the relationships, that is, what the perceived associations are between some of learners' individual characteristics and their English learning and testing experience. To be specific, what motivated them to learn English, what are the motivational changes they experienced, and whether the globalizing society they are in exerts any influence over their perception of and outcomes in English tests such as CET-4.

11.3 Methodology

Interviewing was chosen because it is presumed to be an appropriate and effective approach to inquiry about specific social processes or individual perspectives through direct contact with those involved in natural contexts (Locke, Spirduso, & Silverman, 2000). Participants were interviewed in an attempt to probe into their perceptions of the factors that were considered influential in their English learning, in particular in relation to their CET-4 performance.

Elicitation interviewing techniques were used to "better understand, describe, capture, and model tacit knowledge" of the interviewees (Johnson & Wellers, 2002). These techniques included developing a semi-structured interview protocol to maintain the consistency of the interview process. The interview protocol included three sections. I started with a lead-in section when I introduced myself and the study to the participants, preparing a setting for a friendly chat that would make them feel comfortable. The second section contained interview questions related to motivation and global awareness (see Appendix 11.1 for details). In the third section, two questions were asked, one regarding participants' suggestions for the Chinese English teaching and testing system, and the other about their future plans for learning English.

11.3.1 Participants

Participants of this study were chosen from a survey project with 830 second-year students from 32 classes in a south-eastern university in China in 2008. The participants involved in this part of the study were 12 participants who agreed to be interviewed. These interviewees were representative of the differences in gender (six males and six females), major in university (six from the Arts programmes and six from the Science programmes), and their self-reported English competence (four from each level: high, medium, and low). The participants were asked to rate their English competence based on their previous CET-4 mock test performances.

To facilitate interview data presentation and allow confidentiality of the interviewees, pseudonyms were given to these 12 interviewees. The pseudonyms were given to reflect their genders and their university programmes; within each group, the names are presented in a sequence of their self-reported proficiency levels from high to low. The six students from the Arts programmes were named with an initial A: Alan, Alex, and Adam are three male students from the Arts programmes who had different self-reported English proficiency levels; Alice, Anna, and Amy are three female students from the Arts programmes who had different self-reported English proficiency levels. Similarly, the six students from the Science programmes were named with an initial S: Simon, Scott, and Steve are three male students from the Science programmes who had different self-reported English proficiency levels; Susan, Sara, and Sally are three female students from the Science programmes who had different self-reported English proficiency levels. In addition, in data reporting, interviewees with a high self-reported proficiency level were assigned a number of 01, those who had a medium proficiency level were assigned a number of 02, and those who had a low proficiency level were assigned a number of 03. Table 11.1 presents the overall interviewee profile.

Table 11.1 Interviewee profile

Programmes of study	**Gender**	**Self-reported proficiency level**		
		High (01)	**Medium (02)**	**Low (03)**
Arts programmes	Male	Alan	Alex	Adam
	Female	Alice	Anna	Amy
Science programmes	Male	Simon	Scott	Steve
	Female	Susan	Sara	Sally

11.3.2 Interview data collection

The interviews were carried out in a cafeteria on campus where a separate room was reserved to make sure the interviews were conducted in a quiet setting. The interviewees were given a Letter of Information and a Consent Form. After the interviewees signed the Consent Form and gave permission to audio-tape the conversation, the interviews started, which lasted around 45 minutes for each interviewee. The interviewees were given the choice of conducting the interview in either English or Chinese or a combination of both. Nine of the interviewees chose to conduct the whole interview in Chinese, while three of them (Simon, Alice, and Alex) started with English and switched to Chinese as the interviews went on.

11.3.3 Interview analysis procedures

Several steps were followed in analyzing the interview data. First, to organize and prepare the data for analysis, I transcribed the interviews verbatim in Chinese and translated them into English. I then scanned through each interview, looking for potentially interesting or relevant material for analysis. I began with open coding to help build ideas inductively and remain more attentive to what the interviewees had to say, rather than pre-imposing my ideas. By reading over the transcripts and the highlighted parts, I started to make a list of the key topics, and then clustered together similar topics and made them into a summary table to be used as a preliminary organizing scheme (see Appendix 11.2). For example, students' motivations were clustered into two major categories: their motivation in English learning and their motivation in English testing. Each category was further broken down into several subcategories followed by interviewees' quotes to represent those subcategories.

After the list was made, I went back to the data, re-read the transcripts, and wrote codes next to the appropriate segments of the text according to the preliminary organizing scheme. Some of the codes were developed based on the literature that reflects certain theoretical perspectives (e.g., instrumental orientations), and some of the codes were developed solely on the basis of the emerging information collected from the participants (e.g., motivational change, globalization and Chinese culture). Interviewees' quotes that could best represent certain categories were included in the summary table. With the expansion of the summary table, I grouped some categories together and broke down some other categories. For example, influences from society, teachers, parents, and

peers were grouped together as external influences; reasons to learn English were broken down into mark orientation, further education orientation, and job orientation.

After each interview was coded and analysed, I adopted constant comparative methods (Merriam, 1998) to juxtapose responses from other interviewees. I performed selective coding by scanning data and previous codes when comparing students with different group memberships. I looked selectively for cases that illustrated themes, and made comparisons and contrasts on the students' reported affective factors and their views of the influences on their English proficiency.

11.4 Results

Among the 830 students who participated in the survey study, 73.9 percent of them were male students and slightly over a quarter were female students; 21.6 percent were from the Arts programmes, and the rest were from the Science programmes. Table 11.2 below presents a comparison of the test performance between different groups, i.e., male students vs. female students; students from the Arts programmes vs. students from the Science programmes. Overall, female students' average scores were higher than their male counterparts in all language skills including the total score. In contrast, Standard Deviations (SD) of female students were consistently smaller than male students, indicating score range of the former group was narrower than the latter group. However, the average score differences and SD differences between the students from Arts programmes and Science programmes were close and mixed with either group consistently higher or lower than the other group.

Table 11.2 Means and standard deviations of test scores

Section		**Overall**	**Male**	**Female**	**Arts**	**Science**
Listening	Mean	164.19	161.37	172.15	163.89	164.31
	SD	23.65	23.82	21.26	25.27	23.35
Reading	Mean	166.01	165.03	168.78	164.71	166.34
	SD	23.33	23.65	22.19	24.24	23.14
Writing	Mean	89.22	87.74	93.38	89.95	89.02
	SD	12.41	12.59	10.87	13.87	11.98
Cloze	Mean	46.86	46.08	49.07	47.28	46.77
	SD	7.97	8.10	7.17	8.12	7.93
Total score	Mean	466.27	460.22	483.38	465.84	466.44
	SD	51.05	52.44	42.58	55.56	49.89

The next section reports on the research findings of the interview data to further explore which factors are considered influential in students' English learning and test performance. The findings which emerged from the data are presented in the following sequence: motivation, external influences (which include influences from society, teachers, peers, parents as well as from CET-4), motivational changes, group differences, and the students' perceptions of global awareness. The results section ends with a summary of participants' voices related to how, in their opinion, the English teaching and testing practices could be improved. Direct interviewee quotes are included to support their opinions, followed by information in brackets including three elements: their pseudonyms, proficiency levels, and specific line numbers from the transcription.

11.4.1 Motivation

Students' perceptions of motivation are framed in two perspectives: English learning and English testing. With regards to motivation in English learning, students unanimously deemed that "English is very important" (Alice_01_19), and that English ability is "one important competition ability" for the country and for their personal development (Sally_03_03). Sally noted that being able to learn English well brought her joy (Sally_03_02). Another interviewee, Sara, also mentioned that being able to travel abroad was one of the reasons for her to learn English well (Sara_02_08).

Three interrelated instrumental orientations in learning English were identified: mark orientation, job orientation, and further study orientation. Students acknowledged that English is "a major course in the university," as it constituted the heaviest course credit of all (Sally_03_02). If the student was competing for any scholarship or honour, marks on English tests were among the important evaluation criteria (Simon_01_05). In addition, high English marks were required for graduate studies (Simon_01_05), because all graduate schools asked for good marks in English as one of the prerequisites for admission. Students, especially students from the Arts programmes, stated that future employers value job applicants' English abilities (e.g., Alice_01_19; Alex_02_24; Alan_01_31; Anna_02_28). With the current economic situation in China, finding a desirable job needed "an extra card" in the students' hands (Sally_03_02). In many cases, according to Sally, this extra card meant a certificate of the CET-4 and/or the CET-6 (Sally_03_02).

The concept of integrating with English-speaking people seemed to be a very remote idea. Amy explained that it was difficult to integrate

because different ethnic and cultural backgrounds kept a distance between the two worlds – the Eastern world and the Western world (Amy_03_17). Only two of the 12 interviewees had had previous direct encounters with English-speaking people. One student, Adam, had a marathon-running buddy from Sweden, but the inability to communicate smoothly with his Swedish friend in English made him feel very frustrated (Adam_03_34). Another student, Simon, had two English-speaking high school teachers about whom he had mixed feelings.

The other ten interviewees commented that they had not had direct contact with English-speaking people so far. Their impression of English-speaking people was indirectly obtained from mass media, including movies, TV, and Internet news. The concept of English-speaking people was more like a singular concept to them, and it was represented mostly by Americans. From their indirect experiences, they stated that English-speaking people were generally friendly and honest, probably because of their religious beliefs (Anna_02_29). In spite of the positive impression, students maintained that it was an "other" culture, and they would not consider integrating themselves into it, at least for now (Steve_03_40).

One interviewee, Anna, stated that being able to use English helped her "open a wider window," enabling her to get information from more resources and understand more. She thought that if a person could not speak the language of a certain country, his/her understanding of that culture would be indirect and limited. English was the bridge that led her to a wider world. She mentioned that her long-term goal was to make friends from all over the world and to have a deeper understanding of the issues happening in the world and people from different backgrounds (Anna_02_27).

11.4.2 External influences

Aside from being self-motivated for either instrumental or integrative reasons, the students were motivated by external influences from sources including society, teachers, parents, peers, as well as English tests, in this case, CET-4. The students realized that all these external factors were correlated to some extent but in various degrees (Scott_02_12; Simon_01_05).

11.4.2.1 Society

University students' interests in learning English were influenced by current social and political affairs. In 2008, there were a few important events in China that caught the world's attention. For example,

different opinions of the media were voiced about China's Tibet issue and the coverage of the Olympics. Students' interests in learning English were greatly triggered by these social/political events. They wanted to read the overseas news reports on these events (e.g., news from the CNN and the BBC) to be able to understand the differences between domestic and overseas reports in order to make their own judgments (Alex_02_23). However, some of the reports in English were beyond their level of English. Therefore, they felt more than ever that they, as current university students, should be equipped with good English abilities, so that their judgment would not be limited because of language barriers (Sara_02_09). Some students noted that websites like CNN and the BBC were good channels for keeping them up-to-date, especially when they had read reports of the same events on Chinese websites. They were able to apply their own critical thinking in evaluating these events by trying to understand voices from different sides (Alex_02_23), and not "follow like sheep" (Alan_01_34). "The Eastern and Western cultural clash" was sometimes reflected in these news reports, and to understand these clashes was a good reason for them to keep on learning English (Adam_03_36).

In addition, there were more and more international companies in China that offered higher salaries and better opportunities than the domestic ones (Simon_01_05). Simon, who majored in computer engineering, commented that all programming codes were written in English and the best books in this area were written in English. Therefore, learning English was a must rather than an extra asset for him (Simon_01_05). He regarded it as his responsibility, as a university student, to communicate with other international professionals in his field using the *lingua franca*: English. Sara, sharing Simon's opinion, said that, even though one could resort to translators, the information would be weakened or sometimes distorted if the translation did not go well. She pointed out that translated documents usually were delayed, and that if university students could truly grasp English, they would be able to access the most up-to-date information in the field without delay by reading documents in English (Sara_02_10). Alan also explained that if he had high proficiency in English, he would translate books on such advanced technology into Chinese (Alan_01_34), contributing to Chinese society in this way.

11.4.2.2 Teachers

Although these students recognized that the fundamental force driving them to learn English well was within themselves, rather than in

teachers, textbooks, or tests (Sally_03_02), the influence of teachers was important for them. "Whether I am motivated to learn English or not is partly dependent on the kind of teacher I get" (Simon_01_44). A student with a hearing disability, Sara, commented that her interest in learning English was largely dependent on the sympathy shown by the teacher to her situation. If the teacher was sympathetic about her situation, she would get certain accommodations in the class, and she would feel she was well attended to in a class of over 40 students. This feeling of being accommodated made her attached to the teacher and to the subject he/she was teaching – English (Sara_02_09).

Since, in this university, students had the option of choosing their own English teacher each semester, some students had had four different teachers at the time of the interview, one for each semester. Whether the teaching style suited the students' learning style was sometimes critical in motivating their learning (Simon_01_05). The factors that were deemed important included the teacher's personality, the information his/her class conveyed, and whether or not there was an emphasis on oral English in the class.

Compared with their English teachers in high school, most of the students agreed that their university teachers provided them with more flexibility in class activities, such as group discussion, oral presentation, and role-play activities. This classroom flexibility enabled them to gather information from different sources, including the Internet, English magazines, and movies, which, in turn, increased their engagement with English learning (Sara_02_09). On the other hand, if students felt uninterested in either the teacher or the content of his/her course, they would sometimes be absent from the class or use the class time to catch up with their sleeping (Steve_03_40). Whether the teacher was strict with the students or not also played a role in influencing their motivation to learn English. On the one hand, they preferred teachers who were friendly, outgoing, and easy to get along with. On the other hand, they realized that if the teacher was too lenient with classroom discipline, they would have less self-control and spend less time on learning English. If the teacher was strict enough, they felt that the pressure from the teacher could push them to work harder (Sara_02_10).

The teachers' professional qualifications also impacted on students' motivation in learning. In cases where the students believed that their teacher was very knowledgeable and qualified, they would be more willing to attend the class and participate in classroom activities (Anna_02_27). One of the interviewees had a teacher who had a Master's degree from an English-speaking country; according to this

student, this teacher was more popular than other teachers who had a Master's degree or sometimes even a Doctoral degree from domestic universities (Anna_02_27).

11.4.2.3 Peers

Peer pressure was an important reason for low proficiency students to learn English and pass the CET-4. One student stated: "I am a competitive person, if my classmate passed the test; I want to pass the test too" (Adam_03_34). Another student acknowledged that "Everybody else is studying for the test, I also have to study and pass the test. Otherwise, I will be at a disadvantage in the job market" (Steve_03_41). In contrast, students from the high proficiency level regarded peer pressure as peer support. One student mentioned that he had a classmate who learned English and prepared for English tests with him; they prepared for the CET-4 together during the weekends. This student said that, without his study buddy, he might not be able to control himself that well and spend that much time on English. He maintained that peer support was a small environment that students created to encourage themselves to learn English (Simon_01_45).

11.4.2.4 Parents

Parental influence was relatively weak compared to other external influences. According to the interviewees, since most of them lived on campus and parents were mostly concerned with their general wellbeing, they usually would not be very worried about one specific university course or test. Only two of the 12 interviewees commented on their parents' influence. One student reported that her motivation to learn well was more her responsibility not to disappoint her parents rather than her responsibility to the country or society (Amy_03_18). This interviewee, from the low English proficiency group, maintained that her responsibility to the country or society seemed very remote to her, and the way she saw university students' contribution to society was to do well in university. Another interviewee, Alex, who was also from the low proficiency group, reported pressure from his parents. He said his father was a high school English teacher who was concerned about his English development and constantly reminded him to pass the CET-4 as soon as he could (Alex_03_50).

11.4.2.5 CET-4

CET-4 is considered as one of the vital external influences. According to the interviewees, CET-4 was like a hurdle race, and English learning

was like a running race without a finish line (Alan_01_32). As university students, they had already passed the first hurdle: the university entrance examination. They were then faced with CET-4 as the next hurdle. The students felt the positive side of the CET-4 hurdle, in that it gave them a focus and a direction in learning English (Anna_02_30). In addition, an interviewee commented that tests provide test scores that could be used to screen students for different purposes, e.g., ranking, scholarships (Adam_03_35). The negative aspect of English tests as hurdles was that the teaching and learning of English was made all about jumping the hurdles, that is, the existence of the test turned the objective of teaching and learning English into more or less only about passing the test, no matter whether CET-4 certificate was attached to an undergraduate degree or not (Alex_02_24). Alice echoed that tests like the CET-4 were part of a rat race that this society fostered. She commented that society rewarded results, but not processes, arrivals but not journeys (Amy_03_17).

The interviewees considered it a big drawback that the CET-4 did not include a compulsory spoken part. The test, therefore, could not test students' communicative competence, which they felt should be the focus of language teaching and testing (Adam_03_50). The CET-Spoken test was open only to candidates whose CET-4 test score was over 560 out of 710. Since a spoken component was not included in the regular CET-4 and was not open to everybody, it was less emphasized in the classroom and less practised by the students. However, as one student pointed out, reading and writing skills could be developed by oneself, even with no or few interactions with others. In contrast, speaking skills usually needed practice with others, preferably with those who had higher speaking abilities (Simon_01_05). Some interviewees attributed the phenomenon of "dumb English", meaning the majority of English learners in China are less able to communicate confidently in speaking, to the lower priority placed on the speaking component in testing and teaching (Susan_01_41).

The students reported that the CET-4 was among the major reasons to keep them learning English and spending more time in learning English. Although this test was no longer a requirement for their undergraduate degrees, its importance had not diminished (Simon_01_06). Instead, this kind of detachment from their degrees provided the students with some flexibility; that is, they could graduate with a university degree without passing any English test (Amy_03_16). This possibility was especially beneficial to students who did well in their majors, but not in English. Meanwhile, the presence of the test itself was a pressure to

the majority of the students if they wanted to compete with their counterparts (Amy_03_16). Because nowadays universities are graduating more and more students each year, the competition in the job market in China is becoming fiercer. Tests, including CET-4, have played and will continue to play an important role in selecting candidates for various reasons (Adam_03_35).

The students also felt that, even though the CET-4 had some negative aspects, they did not see any better alternatives. One student argued that classroom assessment might be a choice, but it would be comparatively subjective because it would involve a teacher factor; it would also be difficult to compare students across different classrooms and universities, and it would be extremely difficult to carry out on a nationwide scale (Alex_02_24).

11.4.3 Motivational changes

Interviewees were asked if they had experienced motivational changes in their learning of English. All of them agreed that certain motivational changes occurred from high school to university. Before entering university, they were more-or-less driven by the school curriculum and the upcoming university entrance examination to learn English. At that time, English classes were offered almost every day, and the way they learned English was mostly dominated by the content of the university entrance examination. Achieving a high score on that examination was the major impetus (Alex_02_23).

The students recognized that "there is a disconnection between the university entrance examination and the College English Test" (Simon_01_43). For example, in the university entrance examination, listening and speaking are not tested. Therefore, the students spent less time on listening and speaking while they were in high school. In contrast, in university, maybe because listening was included in the English tests, in preparing for both the CET-4 and the English examinations at the end of each semester, both the students and their teachers spent more time on listening, but, in many cases, still not enough on speaking. Students noted that: "English in high school is mainly for the university entrance exam, while English in university is mainly for our future work and graduate schools" (Simon_01_43). Although they realized that tests could guide teaching and learning, the students agreed they considered the drive from the language tests as positive because the tests gave them direction in learning English; otherwise, they would get lost as to what to learn and what to focus on (Alex_02_23).

11.4.4 Group differences

Group differences were identified throughout the interviews. In addition to some of the differences that were described or hinted at in the previous sections, some other salient group differences are summarized here. Students were asked to give an estimated proportion of their study-time spent on English and on their major subjects. Generally speaking, the female students reported spending more time each week learning English than the male students, the result of which was reflected in the CET-4 score differences between the two gender groups (see Table 11.1). The female students reported that their allotted time in learning English ranged roughly from 30–40 percent of their total available time. The male students, on the other hand, focused more of their study time on their subject areas, claiming to spend only about 10–30 percent of their time on English learning. The male students reported that they usually spent less time learning English in their dorms, because there was not an "atmosphere" that emphasized English learning among male students outside the classroom (Steve_03_42).

Moreover, students with high English proficiency usually had clear objectives for learning English, and they reported possessing a certain degree of flexibility in resorting to different resources that could help them improve their English. For example, they would practise English with students who majored in English in university, and they would attend some English activities and student societies to help them learn English (Alice_01_21). They viewed it as important to maintain a variety of resources to keep up their motivational intensity, their interest in English, and their efforts spent in learning English. On the contrary, students with low English proficiency commented that they had no choice but to pass the test, because everybody else was doing so. They tended to use limited learning resources, focusing mainly on textbooks and anything else that was available (Steve_03_43).

11.4.5 Global awareness: "English is the Bridge"

In this study, the concept of global awareness refers to the awareness, possessed by EFL students, of the important role English plays in the globalizing world. This global awareness concept is reflected by the unanimous understanding of the importance of English in current Chinese society. In the context of globalization, English was deemed as an international language (Susan_01_39) and it was "the *lingua franca* in the world" (Amy_03_16).

Generally speaking, it was not considered a bad thing that every university student was learning English. Simon commented that, even though the Chinese economy was developing very fast, "we are still behind some developed countries in many aspects, especially in terms of advanced technology" (Simon_01_05). It was considered realistic and beneficial for university students to learn English well and use it as a communication tool to be able to understand advanced technology. Simon also noted that English was the bridge in linking Chinese traditional culture and advanced Western technology, especially after China entered the World Trade Organization in 2001 (Simon_01_06).

Students sensed that a balance was needed between maintaining Chinese language and culture and encouraging every university student to learn English well (Scott_02_12). Under the current global circumstances, it was believed to be more important to introduce advanced technologies and good values from outside the country than to simply preserve Chinese language and culture (Simon_01_07). When talking about a balance, the students were open-minded and considered it more crucial to extract the positive aspects from outside than to passively maintain the status quo.

While recognizing the importance of internationalization, the interviewees commented that, because Chinese was their mother tongue, its influence was deep rooted and would not easily fade away. Since everyone who was born and grew up in China had been immersed in Chinese language and culture, its status was solid (Scott_02_12). Scott made an analogy that it was just like many foreigners learning to use chopsticks to eat dumplings; it was the culture exchange that mattered, not the act itself (Scott_02_13). Adam, however, voiced a contrary opinion, arguing that: "it went too far." He asserted that English was important but that it was not necessary for everybody to learn it (Adam_03_34).

Aside from recognizing the importance of English in their future study and career, one interesting thought on why every university student needed to learn English was expressed: "Since they asked us to learn English, there must be some sort of reason behind it" (Alex_02_26). "They" here referred to the university authorities, society, and the government. This student commented that he was more concerned with his personal development than with thinking about the "big picture" – the amalgamation of Chinese culture and world culture. He felt that contributing to the amalgamation of cultures should be one of his responsibilities as a 21st century university student, but it was too far in the distance and not realistic in his current daily life.

11.4.6 Suggestions for English language teaching

Interviewees were asked about their suggestions for English teaching and testing in China. In the area of English teaching, first of all, students stated that two major aspects were crucially important in their English learning at the classroom level in university: vocabulary and spoken English. Since grasping more vocabulary can be solitary work, they suggested that more opportunities for oral communication should be introduced and maintained. For example, qualified teachers, especially those who had received training from overseas and English native speakers, were needed to provide quality practice and guidance. Moreover, the students advocated for more group discussions and oral presentations to be offered in smaller classes. In their experiences, a class normally consisted of 40–50 students, and they suggested that 20–25 was a desirable size for an English class (Alice_01_19).

Second, suggestions were also made to increase the allotment of time for English classroom teaching. The students recommended that English be taught at least three times a week instead of twice a week, so that they could have more practice inside the classroom, which was perhaps the only English-speaking environment they had. Third, in terms of the teaching focus, students suggested that more emphasis should be placed on the two productive skills: writing and speaking. They argued that these two skills were good indicators of one's actual English communicative levels.

Different opinions were voiced as to the necessity of offering English courses to every university student. Some students pointed out that, for some of them, there would be no need to have English proficiency in their future jobs, so students should have the option of not taking English as a compulsory course (Alex_02_40). In terms of cost and benefit, these students felt that the investment in making every university student learn English was larger than its possible benefits – producing university graduates equipped with adequate English abilities to be applied in their future work (Scott_02_12). While some students felt they should have more choices instead of being blindly guided, they all agreed that, even though English might not be needed in their future jobs, certificates of English tests like the CET-4 and/or the CET-6 might be helpful for them in securing a good job.

11.4.7 Suggestions for English language testing

In terms of English testing, suggestions were raised in four aspects. First, it was recommended that more variety of test formats should be

introduced. For example, some test formats from other internationally recognized English tests (e.g., TOEFL and IELTS) could be borrowed (1) to raise the difficulty level of the test; and (2) to be able to test integrated language skills in interactive ways (Anna_02_29). According to Anna, integrated skills referred to a combination of four language skills, for example, testing reading and writing simultaneously; interactive ways referred to testing activities that resembled daily language uses.

Second, the students agreed that the spoken English test should be included in the regular CET-4, and that it should be open to every test-taker, not just test takers who achieved certain scores on the written paper (Alan_01_32). Only by integrating the speaking component into the test and making it open to every test-taker could the CET-4 have a positive washback effect on the teaching of the language; that is, creating a positive cycle in teaching, learning, and testing.

Third, most students needed to use radios to do the listening comprehension part. However, in some cases, problems such as tuning to the right channel or noises from other students in the classroom could trigger a certain level of anxiety. It was suggested that better equipment should be used to facilitate the listening comprehension test (Sara_02_10). Fourth, almost all interviewees commented that they were rushed to complete the test because of time pressure. Therefore, it was suggested that the timing of each section of the test should be optimized based on the average speed of previous test-takers.

11.5 Discussion and conclusion

The results of this study demonstrated different levels of factors that influenced Chinese university students' motivation to learn English and CET-4 test performance. At the personal level, promoting the importance of English in further education and less emphasis on marks might be an incentive to learn. At the classroom level, it is obvious that the teacher factor is important. Students' views on the course and on the teacher constitute vital elements that influence their motivation. Relevance in L2 teaching is important because students will not be motivated to learn if they see no relevance in their English class to the potential use of the language as it may be applied in their current and future life.

The results also showed that this group of students were more motivated by instrumental orientations than by integrative orientations,

and the idea of integrating to an English-speaking community was considered not realistic and not applicable to them. In this particular Chinese university context, in which direct contact with L2 speakers was limited, the L2 community was still well-known to the learners through indirect contact with it, through their exposure to a range of L2 cultural products and artefacts. The effects of the milieu and such indirect contacts (media and the Internet) are important in developing language proficiency in EFL contexts (Csizer & Kormos, 2008).

Participants expressed their unanimous realization of the importance of English in their current study, further study, and future career. At the same time, they acknowledged the balance they thought Chinese university students should strive to keep: developing English competency and maintaining Chinese culture. Previous research has tended to focus on the role of integrativeness in learning English, either integrating into the English-speaking community as suggested by Gardner's socio-educational model (Gardner, 2006), or integrating into the global community as suggested by some other researchers (e.g., Lamb, 2004; Yashima, 2009).

With the current rise of China and Chinese culture, G. Cheng (2008) argued that along with economic globalization comes political and cultural globalization. He suggested that Chinese cultural strategy should switch from a passive "defensive" type to one that is active and "enterprising", from stressing the protection of China's traditional national culture to stressing participation in world culture, and from "taking" to "giving". These ideas were shared by the interviewees in this study who argued that English is a bridge to connect Chinese traditional culture and Western technology and that there should be a balance between maintaining Chinese language and culture and encouraging university students to learn English well. All of the interviewees perceived the significance of English as an important international language in their learning contexts as well as its potential value in many aspects of their lives. They were aware of their own identity, and they did not think their interest in learning English, improving their English skills, and passing English tests (e.g., the CET-4) would mean sacrificing their Chinese identity, language, and home culture.

In terms of group comparison, the relatively higher test scores among female students may be accounted for by the differences in learner motivation. The results showed that females tended to spend more time on English learning than males, and their interest in learning English was higher. Ryan (2009) suggested that the higher performance

and stronger motivation to learn English among female students was because of the common perception of foreign languages as feminine subjects. Kissau (2006a, 2006b) also argued that traditional societal perceptions of language learning ability among male and female students account significantly for the gender difference in language development. In addition, the significant difference in integrativeness found in this study is congruent with Mori and Gobel's (2006) findings, which suggested females placed a significantly higher value on integrativeness in motivation. Similarly, Dörnyei and Clément (2001) found that females scored significantly higher than males on direct contact with L2 speakers, integrativeness, and cultural interest.

Furthermore, the results of this study showed that the value of English is generally conceded by Chinese university students. However, what differentiates students with respect to English achievement is their willingness to commit the time and effort to attain it. High proficiency and low proficiency students differed in how they resorted to different learning resources. A successful language learner is less of a passive recipient of knowledge or input, but usually actively seeks out resources in her or his local context of learning.

The results also showed that the high proficiency students regarded it as their responsibility, as contemporary university students, to learn English well so that they could help to communicate the most up-to-date information in their area of study between English and Chinese speakers. Also, high proficiency students had a clearer vision of their goal in English learning, and they were able to make use of more resources, including learning materials and opportunities. Furthermore, this group of students was usually more strategic in preparing for and taking English tests; for example, they tended to spend more time familiarizing themselves with writing test formats. Learner autonomy characterizes good language learners in learning strategy and learner self-regulation (Benson, 2000; Dörnyei, 2001; Wenden, 1991). Independent interactions with learning materials, educational technology, and learning strategies, as well as learner control over the planning and evaluation of learning and the curriculum are approaches that are associated with the development of learner autonomy (Benson, 2000).

Whether students come from the Arts programmes or Science programmes seemed not to make much difference in their English learning as well as their CET-4 performance. From the interview results, university major differences were found in the interviewees' perception of the importance of English in their future jobs. Even though students from

the Science programmes acknowledged the importance of job orientation as one of the personal level instrumental orientations, students from the Arts programmes seemed to be more emphatic in this respect. Not many previous studies have examined the academic discipline difference in influencing English as second/foreign language learning. Only Andreou, Andreou, and Vlachos (2004) conducted a study with 452 Greek undergraduate students. In their findings, academic discipline difference was a variable that significantly affected students' performance on verbal fluency tasks in English as a second language. The authors explained that different workloads and different departmental cultures might cause this difference. More research needs to be carried out in this respect to further understand the impact of discipline differences in English learning.

Relevance in L2 testing is equally important because, as a large-scale test, CET-4 serves as a benchmark around which English language teaching and learning in China revolves (Zheng & Cheng, 2008). Gronlund (2002) suggested that factors that lower the validity of assessment results, such as inadequate allowed time and poorly controlled conditions, should be decreased. As evident from this study's interview findings, both of these two administrative problems were found in this research context. First, timing analysis of test items, if not yet done, should be conducted to better adjust to the speed of the majority of test-takers. Accuracy will definitely be compromised if test-takers are concerned with their speed in completing the test. Second, testing facilities, in this case, radios used to complete listening comprehension, should be tested to ensure their stable quality to eliminate construct irrelevance that may be introduced because of the unstable quality of the radios used.

Finally, some students from this study stated that they did not currently see any better alternative to replace CET-4 in China, either due to its history or due to its prevalence. L. Cheng (2008) commented that with many university faculty members expressing concern over the reliability and validity of in-house proficiency tests, notwithstanding the negative washback, there is apparently little momentum to replace the CET. Research has suggested that testing should be an efficient agent for pedagogical change (Qi, 2007). General concerns about the use of high-stakes testing usually includes, but is not limited to: (1) the defining of the purposes for which tests were developed (Solorzano, 2008), (2) the alignment of the tests to the curriculum taught in individual classrooms, and (3) the use of the testing instrument to inform high-stakes decisions.

Appendix 11.1 Interview guide

Section I Lead-in conversation

Section II Questions related to motivation (and global awareness)

1. How do you like your English course?
2. How do you like your current English teacher?
   a. How many English teachers you've had so far in university
   b. Are the teachers helpful in preparing you for the test?
3. How do you like the CET-4?
4. How would you describe your desire to learn English?
5. Please tell me why you learn English, and what keeps you learning English (e.g., self- motivated, societal, parental, or peer pressure). Please give examples to illustrate your point.
6. In the current Chinese society, what do you think of the situation of Chinese university students learning English?
7. What do you think of the status of CET in China? What are the positive and negative aspects related to this situation?
8. Are there any motivational changes in your experience in learning English? If so, how and why? What are the reasons behind the change? (e.g. people, important events, etc.)
9. What opportunity do you see will be opened up if you speak good English/perform well in CET-4?
10. Could you please give me reasons for your choices in the section of global awareness (refers to the questionnaire)?

Section III Additional Questions

1. If it were up to you, what suggestions or changes you would make to the status of English learning and testing among tertiary Chinese students?
2. What are your future plans regarding learning English after the CET-4, what do you think will motivate you to continue learning English?

Appendix 11.2 Summary of key data from the interviewees

Themes	English Language Learning	English Language Testing
Motivation	Importance of English • English ability is one kind of competition abilities (SALLY_03_01) Learning English for instrumental reasons Further-study orientation • Further study may require English (SIMON_01_05) Future job orientation • Future employees may value English abilities (SALLY_03_02) Mark orientation • English is a major course which has the highest credits (SALLY_03_02) Learning English for integrative reasons • The desire to integrate with native speakers is remote (AMY_03_17) Interest/Joy • Learning English well can make me happy (SALLY_03_02) External influences • External influences (society, parents, peers, and teachers) were correlated (SCOTT_02_12) Motivational changes • Motivational changes occurred from high school to university (STEVE_03_19)	A hurdle race • CET is a hurdle race (ALAN_01_32) General perception of the CET-4 • CET should not be the goal of college English teaching and learning (SALLY_03_02) Negative side of the CET-4 • CET doesn't adequately test communicative skills (ADAM_03_50) Positive side of the CET-4 • CET gave directionality in learning English (ALEX_02_23) CET changes and its influences • Detachment with degrees provides students with flexibility (AMY_03_16) • Although no longer a requisite for graduation, its importance had not diminished (Simon_01_06). • Currently no better alternatives can replace CET (ALEX_02_24) Test design/time and its impact • Speed compromises accuracy (ALICE_01_19) • Vocabulary is a crucial component in getting good scores (SUSAN_01_42)

(*continued*)

Appendix 11.2 Continued

Themes	English Language Learning	English Language Testing
Global Awareness	Globalization and its relation to English • English is the bridge (SIMON_01_06) • Globalization requires English as an international language (SUSAN_01_39) Globalization and Chinese culture English learning will not influence Chinese culture (SCOTT_02_40)	
Suggestions for Teaching and Testing	More qualified teachers • Qualified teachers, especially those who were trained overseas and English native speakers, are needed (ANNA_02_27) Increase class time • Increase class to at least three times a week instead of twice a week (SARA_02_44) Have smaller class size • 20–25 students each class was a desirable size for English class (ALICE_01_19	Alignment with other standardized English tests • Borrow test formats from TOEFL or IELTS (ANNA_02_29) Inclusion of spoken English test in the regular CET • CET-SET should be open to every test-taker (ALAN_01_32) Improve testing facility • Better equipment to facilitate the listening test (SARA_42_10)

References

Andreou, E., Andreou, G., & Vlachos, F. (2004). Studying orientations and performance on verbal tasks in a second language. *Learning and Individual Differences, 15*(1), 23–33.

Benson, P. (2000). *Teaching and researching autonomy in language learning*. London: Longman.

Cheng, G. (2008). Chinese culture: Self-awareness and self-confidence. *Social Sciences in China, 29*(4), 195–204.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15–37.

Csizer, K., & Dörnyei, Z. (2005). The internal structure of language learning motivation and its relationship with language choice and learning effort. *The Modern Language Journal, 89*(1), 19–36.

Csizer, K., & Kormos, J. (2008). Modelling the role of inter-cultural contact in the motivation of learning English as a foreign language. *Applied Linguistics, 30*(2), 166–185.

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal, 78*(4), 273–284.

Dörnyei, Z. (2001). *Teaching and researching motivation*. Harlow, UK: Pearson Education.

Dörnyei, Z. (2005). *The psychology of the language learners: Individual differences in second language acquisition*. Mahwah, NJ: Erlbaum.

Dörnyei, Z., & Clément, R. (2001). Motivational characteristics of learning different target languages: Results of a nationwide survey. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 391–424). Honolulu: University of Hawaii.

Gan, Z., Humphreys, G., & Hamp-Lyons, L. (2004). Understanding successful and unsuccessful EFL students in Chinese universities. *The Modern Language Journal, 88*(2), 229–244.

Gardner, R. C. (2001). *Language learning motivation: The student, the teacher, and the research*. Paper presented on March 23–24, 2001 at the Texas Foreign Language Education Conference, University of Texas at Austin.

Gardner, R. C. (2006). The socio-educational model of second language acquisition: A research paradigm. *EUROSLA Yearbook, 6*, 237–260.

Gronlund, N. E. (2002). *Assessment of student achievement* (7th ed.). Boston: Allyn & Bacon.

Hernandez, T. A. (2008). Integrative motivation as a predictor of achievement in the foreign language classroom. *Applied Language Learning, 18*(1), 1–15.

Johnson, J. C., & Wellers, S. C. (2002). Elicitation techniques for interviewing. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research: Context and method* (pp. 491–514). Thousand Oaks, CA: Sage.

Kissau, S. (2006a). Gender differences in motivation to learn French. *The Canadian Modern Language Review, 62*(3), 401–422.

Kissau, S. (2006b). Gender differences in second language motivation: An investigation of micro- and macro-level influences. *Canadian Journal of Applied Linguistics, 9*(1), 73–96.

Kormos, J., & Csizer, K. (2008). Age-related differences in the motivation of learning English as a foreign language: Attitudes, selves, and motivational learning behaviour. *Language Learning, 58*(2), 327–355.

Lamb, M. (2004). "It depends on the students themselves": Independent language learning at an Indonesian state school. *Language, Culture and Curriculum, 17*(3), 229–246.

Little, D. (2007). Language learner autonomy: Some fundamental considerations revisited. *Innovation in Language Learning and Teaching, 1*(1), 14–29.

Locke, L. F., Spirduso, W. W., & Silverman, S. J. (2000). *Proposals that work: A guide for planning dissertations and grant proposals* (4th ed.).Thousand Oaks, CA: Sage.

Merriam, S. (1998). *Qualitative research and case study applications in education* (2nd ed.). San Francisco: Jossey Bass.

Mori, S., & Gobel, P. (2006). Motivation and gender in the Japanese EFL classroom. *System, 34*(2), 194–210.

Pintrich, P. R., & Schunk, D. (2002). *Motivation in education: Theory, research and applications* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education: Principles, Policy & Practice, 14*(1), 51–74.

Ryan, S. (2009). Self and identity in L2 motivation in Japan: The ideal L2 self and Japanese learners of English. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, Language Identity and the L2 self* (pp. 120–143). Bristol, UK: Multilingual Matters.

Solorzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research, 78*(2), 260–329.

Wenden, A. (1991). *Learner strategies for learner autonomy: Planning and implementing learner training for language learners*. New York: Prentice Hall.

Yashima, T. (2009). International posture and the ideal L2 self in the Japanese EFL context. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 144–163). Bristol, UK: Multilingual Matters.

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

12

Standard English or Chinese English? Native and Non-Native English Teachers' Perceptions

Ying Zhang

12.1 Introduction

This study focuses on the comparison of native and non-native English speaking university teachers' attitudes on varieties of English in English education in China. The impetus for this study comes, firstly, from the ongoing debate about the role of native speaker (NS) norms in English language teaching and testing and in communication more generally, and, secondly, from the tendency amongst non-native users of English to distrust local norms, viewing them as insufficient or at least inappropriate for communicative purposes.

Traditionally, a native English speaker (NES) is defined as a speaker of Standard English (Davies, 1999) and, as such, has long served as the norm for language teaching and testing. The assumption here is that native speaker competence is the ideal that ESL/EFL learners are aiming for. This ideal underpins much of the ESL/EFL teaching/testing industry. Teaching materials are usually developed based on NS conversations and texts. Educational initiatives drawing on the assumed expertise of NES teachers are prevalent in many non-English medium countries (Cook, 1999). In the field of language testing, NSs are often used in test pilot studies, on the assumption that if they cannot do the tasks or score poorly, the test must be flawed or unreasonable in its demands. In addition, NS raters are commonly employed to mark a test, and NS standards are invoked in drawing up assessment criteria (Davies, 2004; Lazaraton, 2005; Lowenberg, 2002; Seidlhofer, 2001). Furthermore, Standard English (SE) remains a point of reference in language testing. According to Elder and Harding (2008), the default model for language testing is still often Standard English (SE) as used in 'inner circle'

countries and as codified in English grammars, dictionaries and the like (p. 34.3), on the grounds that 'it allows for greater certainty about what is best assessed', and the appeal of SE lies in its apparent neutrality, in the sense that it is the variety most likely to be equally familiar to all test taker groups (p. 34.5).

However, in the wake of the rapid spread of English as a *lingua franca* (ELF) in world communication, the status of the NS norm has been widely challenged. These challenges take a number of forms, covering issues of definition (e.g., Davies, 1991, 2003), ownership (Beneke, 1991; Crystal, 1997; Gnutzmann, 2000; Graddol, 1997, 1999; Seidlhofer, 2001), identity (Brutt-Griffler & Samimy, 2001; Lazaraton, 2005; Phillipson, 1992; Rampton, 1990) and language use (Bruthiaux, 2003; Davies, Hamp-Lyons & Kemp, 2003; Jenkins, 1998, 2002; Lowenberg, 2002) (See Zhang & Elder, 2011 for detailed review).

The other side of the debate on NS norms focuses on issues of practicality as well as on attitudes and ideology. Quirk (1990) clearly states that the use of non-native varieties of English as pedagogically acceptable models is unacceptable since these varieties are not adequately described (also see Elder & Davies, 2006). As for language attitudes, it is undeniable that NS norms are still regarded as an ideal by most ESL/EFL learners. NS norms are widely believed to be superior and it is not uncommon for users to stigmatize local varieties of English, especially their own (Davies, 1996, 2003; Elder & Harding, 2008; Kirkpatrick & Xu, 2002; Lukmani, 2002; Medgyes, 1999; Quirk, 1990; Seidlhofer, 2001; Timmis, 2002). Empirical studies provided evidences that 'standard' varieties of English generally receives more favourable evaluation from NNSs than the 'non-standard' varieties of English (Xu, Wang & Case, 2010; Zhang & Hu, 2008). Practically speaking, therefore, regardless of one's ideological stance on this debate, it is not easy to simply abandon the NS ideal, especially in the fields of English teaching and testing, which are by their nature normative in operating in the relevant target language use. Furthermore, it is not altogether clear whether the differences between Standard (native speaker) English and other varieties (to the extent that they have been described) are great enough to warrant entirely different tests (McKay, 2002). It is also unclear whether local English users, who are often employed as assessors of locally-developed tests of English, do indeed judge candidates differently from their NS counterparts, and in turn English language teaching to ESL/EFL learners in universities.

12.2 Literature review

12.2.1 Influence of NS norms in English education in China

NS norms, expressed as Standard English, are upheld in English education at tertiary institutions in China, both in the employment of NS teachers in universities, and in the national curriculum and teaching materials of university English education.

Firstly, NS teachers are the ones most sought after in EFL/ESL countries (Jenkins, 2003), including China. It is acknowledged that NNES teachers have an advantage in teaching NNES students in many aspects such as teaching methodology, and conversely, NES teachers show limitations in teaching NNES students in that they do not understand local teaching and learning contexts and often lack teaching experience (e.g., Kirkpatrick, 2002). Nonetheless, NS teachers are still accepted as experts and authorities (Widdowson, 1994), and preferred by students (Liao, 2008; C. Tang, 1997). The local norm in China, which is referred to as Chinese English, or sometimes China English, is not welcomed by key stakeholders, including parents, teachers, business leaders, schools, students, examination authorities, and the government itself (Li, 2007, p. 14). As a result, institutions offering English language programmes often promote themselves as employing NES teachers, and advertisements for teaching positions often require that applicants are native speakers (Clark & Paran, 2007; Liu, 1999; McKay, 2002; Medgyes, 1994; Todd & Pojanapunya, 2009). In addition, few non-Chinese NNES teachers from different first language (L1) backgrounds other than Chinese teach in China. This indicates that NES teachers are appreciated, and students do not have the opportunity to be aware of different varieties of English other than Standard English (Li, 2007).

Secondly, the English teaching curriculum in China has traditionally been dominated by NS-based pedagogical models (Li, 2007, p. 11). This is shown in the implicit acknowledgement of NS norms as stated in the goals of the national College English teaching syllabus. For example, Bai (2008), in a review of the *College English Curriculum Requirements* (Ministry of Education, 2004), argues that NES ability is an implied goal, rather than English as a *lingua franca*. This NES-based goal is indicated in the emphasis on English-speaking countries' cultures and language. For instance, the basic requirement of oral English ability as described in the *College English Curriculum Requirements* is:

> Students should be able to communicate in English in the course of learning, to conduct discussions on a given theme, and to talk about everyday topics *with people from English-speaking countries*. They should be able to give, after some preparation, short talks on familiar topics with clear articulation and basically correct pronunciation and intonation. They are expected to be able to use basic conversational strategies in dialogue (p. 9).

Thirdly, the material used in textbooks for non-English major university students is influenced by NES countries and their cultures. In order for textbooks to be authoritative, they are compiled by a government-appointed panel of experts according to the curriculum set by the government, and are universally used by universities and colleges throughout the country (Wang, 1999, p. 47). The textbook writers always acknowledge the value of 'authentic texts', which are defined as texts which use the 'real' language created by native speakers of that language in pursuit of communicative outcomes (Little, Devitt & Singleton, 1989). Thus, writers of major College English textbooks all claim that only authentic texts are included in their textbooks (e.g. Cai & Tang, 2008; Dong, 1992; Zhai, 1986). Authentic texts as defined by these textbook writers are materials which originate from authentic sources, such as newspapers, journals, magazines, encyclopedias, EFL/ESL textbooks, novels and short story books. These materials are usually written by English native speakers and published in NES countries, particularly in the United Kingdom and the United States (Dong, 1992; Feng & Byram, 2002), and/or are adapted according to pedagogical needs (Cai & Tang, 2008). These textbooks mainly adopt the target-culture-only approach (Cortazzi & Jin, 1999; B. Zhang & Ma, 2004, p. 61). Material in these textbooks is mainly about the language and culture of native English-speaking countries, such as the US and the UK, while neglecting information about other non-native speaking countries including China, where English may be used as a *lingua franca* (Bai, 2008; Feng & Byram, 2002; J. Tang, Xing & Yang, 2005).

The controversy surrounding NS norms as described above is highly relevant to the field of language testing, raising the question of whether such norms do still serve as a benchmark for testing purposes, in a world where the mother tongue of speakers of English is, more often than not, a language other than English. This question is of particular importance with a local test like the College English Test-Spoken English Test (CET-SET), where performance is assessed by NNES raters from China.

The CET-SET is a nationwide spoken English test designed to assess the oral communicative ability of Chinese university and college students based on the *National College English Teaching Syllabus* (National College English Testing Committee, 1999, p. 14). The teaching syllabus is an official document which guides English teaching at university level. College English is a compulsory course for university students in China whose majors are not English. It is a spoken sub-test of the paper-and-pencil test of the College English Test (CET). The CET-SET is a local English test in the sense that the test content is about local life, the candidates and interviewers are non-native speakers of English, and the rating is carried out by averaging the scores given by two non-native speaking raters. The rating criteria used to assess performance on the CET-SET are: *Accuracy and Range*, *Size and Discourse Management*, and *Flexibility and Appropriacy*. The sub-score of the first criterion (i.e., *Accuracy and Range*) carries a heavier weighting (the first sub-score is calculated as 'raw score × 1.2') and the third criterion (i.e., *Flexibility and Appropriacy*) is weighted least (the third sub-score is calculated as 'raw score × 0.8') (see Y. Zhang & Elder, 2009, for details of conversion between weighted scores and band grades, and band grade description).

The CET-SET rating scale implies a NES norm, as does the national teaching syllabus discussed above. For instance, the remarks relating to linguistic accuracy make reference to 'errors', while descriptors of pronunciation at the top levels 4 and 5 pay attention to 'residual accent'. Because the resources which teachers and students look on as authorities of correctness and appropriateness are reference works such as dictionaries, grammar books, and guides to usage, all of which are based on native-speaking models (Tsui & Bunton, 2000), any linguistic features which deviate from the standard are treated as errors in need of correction.

An interesting question is whether the non-native raters employed to rate CET-SET performance are indeed operating according to these norms, or whether they are accepting a more local variety of English, comprehensible and acceptable to users of English in the Chinese context. This kind of exploration has both theoretical and practical implications. It will provide insight into claims made by advocates of World Englishes and English as *lingua franca* about the distinctiveness of non-native varieties of English and will have practical implications for the recruiting of raters, indicating whether native and non-native speakers of English can be used interchangeably as judges of oral proficiency.

The issue was investigated in the two parallel studies to this paper, using the CET-SET as a research instrument, and comparing the NNES

and NES teacher rater behaviours when they gave marks according to their interpretation of the spoken English ability (Zhang & Elder, 2011) and according to the test rating scale (Zhang & Elder, 2014). The two studies reach similar findings in that, although the outcomes of the rating process are broadly similar, the NES and NNES teacher raters have somewhat different approaches to rating and may arrive at their judgements via somewhat different pathways and show different degrees of tolerance of breakdowns in relation to particular features of speech. Linguistic competence features such as accuracy were more powerful in determining the NNES teacher rater group's ratings, while NES teacher raters paid attention more widely to other aspects of oral communication features and showed greater leniency on pragmatic competence features such as appropriacy. Chinese NNES raters are not operating according to a different code. If anything, they may be more oriented to Standard English than the NES group.

12.2.2 Studies on attitudes to English in Chinese universities

The importance of teachers' beliefs has been well documented in second language acquisition research. Teachers' beliefs play a role in teachers' decisions, judgments and behaviours (Shavelson & Stern, 1981), and teachers' attitudes/beliefs have a considerable impact on the ways in which varieties are selected for teaching purposes.

Four surveys undertaken by Kirkpatrick and Xu (2002), Hu (2004, 2005), Jin (2005), He and Li (2009) and He and Miller (2011) investigate both students' and teachers' views on English and Chinese English in universities in the context of China. These investigations show that, during the seven years from 2002 (Kirkpatrick and Xu's study) to 2009 (He and Li's study), the NS norm is still the goal of students and teachers, although there is a clear trend of their being increasingly aware of the varieties of English in the world.

Kirkpatrick and Xu's (2002) analysis of the outcome of a questionnaire survey administered to university students in China shows that, in the context of China, the NS norm is deeply embedded and difficult to change. Although the students being surveyed were open to the idea of the existence of several varieties of English and felt that Standard English was not the sole preserve of native speakers, they were unimpressed by the notion of a Chinese variety of English, and very few appeared happy to sound Chinese when they spoke English. They felt that the variety 'China English' was not yet socially acceptable, and they preferred to learn American English as a Standard English.

Hu's (2004) questionnaire survey to over 1200 Chinese students and Hu's (2005) investigation into 586 Chinese English teachers' attitudes yielded similar results, in that American English was the most favoured variety to be learnt by students followed by British English. However, Hu's (2005) survey indicates that teachers more commonly accepted Chinese English than students did, on the grounds that, according to their teaching experience, they could see that their students could never hope to attain American native speaker competence. The teachers also referred to Chinese English as a kind of Standard English, and they had no wish to disguise their Chinese identity, whereas in Kirkpatrick and Xu's (2002) study, 60.8 percent of students strongly objected to being recognized as Chinese when speaking in English.

J. Jin (2005) investigated Chinese students' views on Chinese English and NS/NNS teachers using data from questionnaires, group discussion and interviews. Moreover, the author introduced Hu's (2004) paper and a lecture about World Englishes (WEs) to students as an intervention, and used two sets of questionnaires with the same topics but different wordings to compare pre-lecture and post-lecture outputs. It was found that the students reacted positively to the information presented and became more positive towards English as an International Language (EIL) after the lecture. For instance, they largely came to disagree with NS norms (increasing from 17.7 percent to 76.5 percent) and the number of respondents seeking to get rid of a Chinese accent decreased from 82.3 percent to 29.4 percent. Although the students still felt reluctant to learn Chinese English, they became more accepting of the idea that Chinese English should stand alongside British English and American English. In addition, it was also found that students' reactions to World Englishes (WEs) seemed to have slightly influenced their preference for NS/NNS teachers, and the number of students preferring Chinese teachers of English doubled after the lecture.

He and Li (2009) investigated college teachers' and students' perceptions of the ideal pedagogic model of college English in China, which they called 'China English', as opposed to a native-speaker-based standard. The participants were 795 non-English major students and 189 teachers of English from four universities in China. The findings indicate that the preferred teaching model of college English in Chinese classrooms is a standard variety of English, supplemented with salient, well-codified and properly implemented features of 'China English'. Most findings of this study are similar to those of Kirkpatrick and Xu (2002), in that (1) most participants did not want to be identified as Chinese while speaking English; (2) most participants

believe that 'the non-native speakers can also speak Standard English' and that 'there are many (kinds of) Standard English'. Furthermore, there is a trend (3) whereby college students (and most probably their English teachers too) in China are becoming increasingly aware of and tolerant toward 'China English'; and (4) the majority of student participants and teachers in China prefer an exonormative, NS-based model of English as the teaching model. In addition, the main goal for Chinese people in learning English remains unchanged from the way Kirkpatrick and Xu (2002, p. 277) set it out seven years earlier, not so much for 'intraethnic communication' as for communication with NNSs from other L1 backgrounds. This is a trend which is becoming more and more evident as NSs of English are outnumbered by NNSs of English by an ever-widening margin, notably in Asia (Crystal, 2003; Dalby, 2001; Graddol, 1997, 2006; Jenkins, 2003). However, slightly more respondents in Kirkpatrick and Xu's (2002) study believed that the purpose for Chinese to learn English is for communication with NSs rather than with NNSs, while the opposite trend is true in He and Li's (2009) study. The main difference in findings between these two latter studies is that, in Kirkpatrick and Xu's study, the possibility of a variety of English in China in the future is acknowledged by a minority (46 percent) of respondents, whereas close to two-thirds of the participants (60.5 percent) in He and Li's study regard such a development as possible. This difference shows an increasing awareness of 'China English' as being a legitimate alternative to a NS-based pedagogic model of English.

He and Miller (2011), while investigating whether it was the NNES or NES that Chinese non-English majors prefer as their English teacher, provided a questionnaire survey to 984 college students and their teachers at four universities in different parts of China asking about their views on varieties of English and Standard English. The results of the questionnaire survey suggested that Chinese teachers and learners of college English alike realized that although English is the global language there are more models to imitate than only British or American standards. Participants have high expectations of the level of English speaking they wish to achieve and the majority of them (82%) preferred to sound like a NS. It was also concluded that the participants believe that college English should be taught by both NNES and NES teachers in China.

These investigations show that in the context of Chinese universities, there is a clear trend that the university students and teachers are increasingly aware of different varieties of English in the world, and

their reactions to Chinese English and different varieties of English have been influenced after they are educated about the WEs concepts. However, NS norm and Standard English is still the goal of students and teachers.

12.3 Research methods

Differently from previous studies, this study looks at both NES and NNES teachers' attitudes towards varieties of English in China and addresses the research question: What are NNES and NES teachers' attitudes to Standard English and other varieties?

As a parallel part of an investigation of the impact of teachers' L1 background on their evaluation of university students' spoken performance in a nation-wide university English language test in China (see Zhang & Elder, 2011, 2014), this study analyses their responses to a questionnaire survey and considers the possible links between these responses and their teaching and judgments of students' English ability.

This paper focuses only on the views of English language teaching and testing to non-English majors at universities in China. Firstly, English majors in China are expected to graduate with near-native proficiency in English and are a minority group of learners within the Chinese educational system. Secondly, since non-English majors constitute the majority of the potential English-speaking and English-using population in China, we believe their teachers' attitudes on English should be investigated (He & Miller, 2011).

12.3.1 Participants

40 teachers participated in the study, all being university English teachers with some experience of teaching English to students from China. 20 of these teachers were China residents as NNES and 20 were NES. While it is acknowledged that defining the native speaker is a complex and controversial matter (Davies, 2003), in this study we have settled for a simple operational definition of a NES as someone who is born and educated in an English-speaking country, or educated in English from an early age and has completed at least their secondary schooling and tertiary education in an English-speaking country.

The 20 NNES teachers recruited for the study, seven males and 13 females, were from five universities in China. The NNES teachers ranged in age from their 20s to their 50s, with English-teaching experience ranging from five to 32 years. They had all learned English at school and had obtained either a bachelor or master degree in English in all

but one case. They could therefore be assumed to be highly proficient in English.

Of the 20 NES teachers, ten males and ten females, ten were currently teaching ESL or language-related subjects in two universities in Australia, and another ten were employed as teachers of EFL in four universities in China, originating from the USA, New Zealand or Canada. While the age range of the NES group was similar to that of the NNES teachers, their median age was higher with most in their 50s rather than in their 30s as was true for the NNES group. While the English/EFL-teaching experience of the NES teachers was similar to that of the NNES group, varying from two to 35 years' duration, they were somewhat more educated in that six NES teachers had obtained a PhD qualification, usually in a language-related discipline.

It should be noted that both the NES and NNES groups varied widely in their prior experience of rating, with some NESs having worked as assessors for high-stakes tests like International English Language Testing System (IELTS) (for between two and 12 years with a mean of six years of prior rating experience) and some NNESs having experience of rating local tests including the CET-SET (for between two and six years).

It should be acknowledged that although the two groups of NNES and NES teachers were broadly similar in a number of respects, the sample is very small and unlikely to be representative of the wider NNES and NES populations. In addition, their similar affiliation to a university/city may also have a potential impact on the study's outcome.

12.3.2 Instrument

All participants answered the questionnaire about their attitudes to English varieties and their views about the norm appropriate for English language teaching and assessment. The participants surveyed did not actually mark CET-SET for this study (although some had marking experience). This questionnaire consists of eight items which were developed based on relevant questionnaire surveys in the literature (e.g., Kirkpatrick & Xu, 2002; He & Li, 2009; Hu, 2004, 2005; J. Jin, 2005). Items 1, 2, 3, 4, and 7 are statements requiring an indication of agreement or disagreement and an accompanying explanation. Items 5 and 6 require a choice from the options or an open-ended response if the provided options are not satisfactory. Respondents were allowed to choose more than one of the provided options. Item 8 required either a 'Yes' or 'No' response, and investigated teachers' self-perception of norms they applied when rating. Before administering the questionnaire, the researcher provided a comprehensive introduction to CET-SET

to all teacher participants (unless they were the trained CET-SET raters) to familiarize them with the CET-SET assessment, so that they were able to refer to the test when responding to the questionnaire items, especially Item 8. All teacher participants watched the official CET-SET rater training video which includes the introduction of CET-SET format, content, rating criteria, and sample test video-recording with justification of scores awarded to each test-taker in the video.

Table 12.1 displays each questionnaire item and its purpose.

12.3.3 Data analysis

The answers to each item of the questionnaire were grouped according to agreement or disagreement. If the participant agreed with the statement, the answer was categorized as 'YES'; if the participant did not agree with the statement, the answer was categorized as 'NO'. Other categories were also created, such as 'Not sure' and 'Do not care' when they are not minding if the issue in the item happens. The frequency of mentions of each kind of response and their explanations were compared between the NNES and NES teachers.

12.4 Findings and discussions

12.4.1 Findings

Table 12.2 summarizes the questionnaire findings.

As Table 12.2 shows, while similarities were found in frequencies of YES and NO responses to Questions 1 and 7, there are noticeable differences between native and non-native teacher groups in other respects.

Both the NES and NNES teachers believed that Standard English (SE) was 'owned' not only by native speakers of English (Item 1). Comments by members of both groups expressed the view that SE could be mastered fully through hard work. They also questioned the definition of SE in respect of whether a speaker is a NES or NNES. For instance, one NNES teacher commented that:

> Firstly, native English speakers do not necessarily speak Standard English, which is closely related to the degree of a person's education; secondly, although our native language is Chinese, many of us can speak perfect Standard English, which is closely related to a person's language environment and background. (NNES 8)

Furthermore, the NNES and NES teachers were similar in the way they defined native speakers of English (Item 7). Members of both teacher

Table 12.1 Questionnaire

Item	Purpose	Sample responses
1. Only NES can speak standard English	To investigate participant's views on the ownership of Standard English.	Don't agree. Many learners/users of English can speak Standard English because of the needs they have to use English, the number of years of learning, etc. (NES 10*)
2. There are many Standard Englishes	To investigate participant's understanding about the varieties of English. The answer to this question may reflect the participant's degree of tolerance to English spoken by university students.	Yes, and not only British and American: my own 'NZ standard' is a little 'different' from 'Australian standard' which is identifiably different from British 'R.P'; 'British standard' is actually varied (I'm sure you've listened to modern BBC announcers) and American also varies. There are Canadian, South African, Indian, etc. ... (NES 1)
3. One day Chinese English will become a language in its own right	To examine views on Chinese English and understandings about the NS norms, which may underpin participant's teaching and judgments.	I don't agree. Chinese English is a problem which needs to be addressed in current English education. (NNES 12)
4. When I speak English, I want people to recognize where I come from	To examine views on varieties of English and their perceived status.	I don't agree. I think the highest level of English learning is to be able to speak English in the same way as native English speakers do. So I hope that I don't have an accent. (NNES 12)
5. Target variety/varieties of English which Chinese university students should learn is ……	To further probe the participant's view on the status of the NS norm and Standard English (SE), especially in the Chinese context.	American English; British English; Any standard varieties of English; English as it is spoken in Asia; English as it is spoken in China; Other: (All above options were ticked including Other). Comment: Learners of English are best served if they are made aware of the range of social, regional and functional varieties (dialects, sociolects and registers) that exist in the performance of this language around the world, and have opportunities to adopt and adapt varieties of English in line with their own communicative needs and priorities. (NES 15)

6. The main reason why Chinese people study English is ……	To explore participant's answer to Item 5.	Other: As English is one of the *lingua franca*, learning English is for communication with all who speak English in the world; however, many people just stick to conventional ways of learning, and learn English in the same way as they learn other subjects. (NNES 17)
7. Are you a native speaker of English? How would you define a native speaker of English? Please explain.	To investigate the participant's definition of a native speaker.	No. The person who was born in an English speaking country and whose first language is English is a native speaker of English. (NNES 2)
8. Will you apply NES norms while rating the local English test such as College English Test-Spoken English Test (CET-SET)?	To ask directly about whether the participants apply a native speaker norm while rating the local English test such as CET-SET test.	YES or NO answer.

**Note*: The 20 NNES participants are identified by being named as NNES 1 to NNES 20; the 20 NES participants are identified as NES 1 to NES 20.

Table 12.2 Teachers' attitudes to English (frequency of comments)

Item	Attitude	NNES	NES
1. Only NESs can speak Standard English (SE)	NO	18	20
	YES	2	0
2. There are many Standard Englishes	YES	12	17
	NO	7	1
3. One day Chinese English will become a language in its own right	NO	17	10
	YES	3	8
	NOT SURE	0	2
4. When I speak English, I want people to recognize where I come from.	NO	9	2
	YES	1	4
	DO NOT CARE	4	14
	NOT SURE	2	0
5. Target variety/varieties of English which Chinese university students should learn is	American/British English	13	2
	Any standard variety of English	5	15
	English spoken in China	0	7
	English as it is spoken in Asia	0	4
	Other	1	8
6. The main reason why Chinese people study English is	To communicate with NES in the world	14	7
	To communicate with other NNES in the region	1	3
	Other	5	11
7. Are you a native speaker of English? How would you define a native speaker of English? Please explain.	YES	0	20
	NO	20	0
8. Will you apply NES norms while rating the local English test such as College English Test-Spoken English Test (CET-SET)?	YES	10	5
	NO	10	10

groups invoked the same criteria for native speaker-ness. In their comments they defined native English speakers variously in terms of (1) whether English was their first language; (2) whether English was the language they had learned at early childhood or since birth; (3) whether they had grown up and lived in an English-speaking country or society; (4) whether English was used for communication rather than simply for learning or study purposes; and (5) whether they were educated

in English at school. Two NES teachers also mentioned the criterion of being identified by other native English speakers as having native speaker competence.

Despite general agreement about the ownership of English, more NESs (n = 17) accepted the idea that there were many Standard Englishes than did the NNESs (n = 12) (Item 2). Comments by NNES teachers show that their YES answers are qualified by their understanding of the term 'standard'. A typical comment is '… The Standard English that I recognize is British English and American English' (NNES 3). Answers regarding the appropriate target varieties of English for Chinese learners (Item 5) confirm the non-native teachers' preference for American and British English. More NESs believed that any standard variety of English should be the target for Chinese learners (n = 15), whereas more NNES teachers believed that American/British English should be the target (n = 13). On the one hand, 15 out of 36 responses to this item provided by the NES teachers were 'any standard varieties of English should be learned in China', as Chinese people 'will meet people from all parts of the world'. Among the 'Other' answers, these NES teachers implied that any English that meets learners' needs should be the target. For instance, one teacher commented:

> Learners of English are best served if they are made aware of the range of social, regional and functional varieties (dialects, sociolects and registers) that exist in the performance of this language around the world, and have opportunities to adopt and adapt varieties of English in line with their own communicative needs and priorities (NES 15).

Another seven NES teachers believed that English as it was spoken in China was a valid target, as well as other varieties of English in Asia (four NESs), because 'this is realistic' (NES 1). Overall, only two of the 20 NES teachers indicated British and/or American English as the preferred target in China. On the other hand, while five NNES teachers regarded any standard variety as acceptable, the majority (13 out of 19) chose American English and/or British English as the target in China.

A difference between groups was also observable in their divergent opinions in relation to Chinese English (Item 3). As many as 17 NNES teachers disagreed that Chinese English would eventually become a language in its own right, while only ten NESs disagreed with this proposition. Various reasons were provided in support of these responses. Five of

the NNESs did not accept Chinese English because 'Chinese English will at most be a style or variety of English, not an independent language' (NNES 2), 'China is not internationalized enough' (NNES 6), 'Chinese English is a problem which needs to be addressed in English education' (NNES 12), 'In terms of population, not enough Chinese people speak English in China' (NNES 19). Those NESs who were opposed to the notion of Chinese English becoming a language in its own right offered reasons such as 'Great languages such as English and Chinese should not be mixed up and lose their unique features' (NES 9), and even 'This is conceivable, but unlikely until such time as China chooses to adopt English as a national standard language, as have, for example, India, Singapore, South Africa, New Zealand, Australia, Canada and the USA. However, recognition of and tolerance for the characteristics of English learned and used by Chinese in China is likely to grow in some areas' (NES 15). Interestingly, both NES teachers residing in China and NNES teachers chose to comment further on the development and encouragement of Chinese language and culture, for example, 'Chinese language is a powerful language itself and many foreigners are learning Chinese. In the future, Chinese language may be as important as, or even more important than English language' (NNES 14), hence, '… such great languages as English and Chinese shouldn't get mixed up and lose their unique features' (NES 9), and there is no market and no need to create Chinese English in China (NNESs 4, 8, 9). On the other hand, eight NES teachers agreed with the statement that Chinese English would become a language in its own right, while only three NNES teachers agreed with the statement. The NES teachers accepted Chinese English because they believed that in terms of 'pronunciation, usage and vocabulary features' (NES 1), Chinese English already existed, or was quite possible. They commented that 'teachers should present 'the standard', they should take a relaxed view of variation', and it was important for speakers to 'keep an accent identity' and to 'reflect pride in their home culture' (NES 2). However, although admitting that Chinese English was likely to become a language in its own right in the future, one NES teacher hoped not.

The NNES and NES teachers expressed different attitudes about being recognized by their accent (Item 4). Nine of the 16 NNES teachers who responded to the item did not want to show their Chinese nationality when they spoke English. They aimed at speaking native-like English, specifically British English and/or American English, and were proud to be recognized as being able to speak American English. On the contrary, most NES teachers (n = 14) stated that they did not care, as long as they

were able to achieve communication. Another four NESs did want people to recognize where they came from. They hoped to maintain their identity, because they were proud of their native culture. Only two NES teachers did not want it to be evident where they were from when they spoke English.

Most NNES teachers (n = 14) recognized communication with NESs in the world as the main reason for learning English in China (Item 6), while the NES teachers acknowledged a broader range of reasons. Only one NNES teacher regarded communicating with other NNES speakers as the main reason, the other five NNESs being more inclined to invoke reasons like passing examinations, getting a degree and looking for jobs. Three of the NNESs also mentioned the reason of communication with anyone who speaks English in the world (NNES2, NNES10) as English is the *lingua franca* (NNES17). In contrast, only seven NES teachers regarded communication with NESs as the reason for Chinese people to study English, six of them were NES teachers living in China. The NES teachers pointed out the same reasons proposed by the NNESs under the 'Other' category, such as communicating with all people in the world (n = 6), curriculum requirements (n = 2) and job hunting (n = 2). In addition, the NES teachers also mentioned immigration (n = 1) and the status of English as a global language of power (n = 2) as possible motivations for studying it.

Moreover, more NNESs (n = 10) admitted applying a NS norm to evaluate their students' English than did NESs (n = 5) (Item 8). This finding corresponds with the findings that NNES teachers showed preference for SE, or more specifically, American and/or British English. On the other hand, only one third of the NES teachers (five out of 15) admitted to applying NS norms, and they all lived in China at the time of data collection.

12.4.2 Discussion

In summary, the questionnaire survey found that, in spite of general agreement on the ownership of English and the definition of native speakers of English, the NES teachers appeared more open-minded about varieties of English than the NNES teachers.

This finding is evidenced, firstly, by the NES teachers' broader range of answers about varieties of Standard English (Item 2) and about the target English for Chinese people to learn, and secondly, by the fact that the NNES teachers prefer to master SE (Items 3, 4). This finding suggests that the NES teachers were more tolerant of a range of varieties of English and NNES speech, and hence would be more tolerant

in assessing their students' English performance. This inference corresponds with the findings in the parallel studies (Zhang & Elder, 2011, 2014) which further investigated what features of speech the NESs would be more tolerant with when assessing a locally-developed English language test such as CET-SET. For instance, the NES teachers, when trained to rate the CET-SET performance, were more lenient in evaluating the test takers' ability to deal with different situations and topics and use linguistic resources according to context – the pragmatic competence embodied in the rating criterion *Flexibility and Appropriacy*.

On the other hand, the NNES teachers expressed a greater preference for a NS norm and SE which was defined as American/British English. More NNES teachers did not want to be recognized as NNESs. This may relate to the fact that since NESs are usually considered to be 'foreign experts' and few NNESs, except those with high education, can become fluent English speakers, being able to speak English like a NES is a symbol of being well educated and intelligent (Lin, 2006). This may also relate to the rater training and practice experience, as it was the NNES teachers who had prior CET-SET training and rating experience who were more likely to identify themselves as having a British or American accent, the so-called SE accent in China. However, on the other hand, these NNES teachers were more reluctant to acknowledge that they applied the NS norms in rating perhaps because, on account of their training, they were wary of mentioning anything other than the official CET-SET rating criteria as the basis for their ratings.

The NNES teachers' preference for NS norms rather than the Chinese variety of English is similar to the finding of Timmis' (2002) study which revealed a similar desire to conform to the NS norm in survey responses from students and teachers from over 45 countries. It also corresponds with the results of previous studies conducted among Chinese teachers and students on their attitude towards China English (e.g., Kirkpatrick & Xu, 2002; He & Li, 2009; Hu, 2004, 2005; Jin, 2005). In addition, this finding accords with parallel research on teacher rater behaviour when assessing the CET-SET (Zhang & Elder, 2011, 2014), showing that while NES and NNES teachers generally agreed in rating Chinese students' oral proficiency in English with respect to overall scores, the latter group paid more attention to deviations from the SE norm; on the contrary, as indicated in this attitude survey, the NES teachers greater openness to English varieties might explain the reason the NES group tended to be more lenient to breakdowns in relation to particular features of speech, such as the communicative competence embedded in the criterion *Flexibility and Appropriacy*, and their tendency to refer to categories other

than those included in the CET-SET rating scale (while the NNES teachers were more likely to refer to the CET-SET rating scale and especially to the form-focused *Accuracy and Range* category (Y. Jin, 2000) which covers accuracy in pronunciation, stress/intonation, use of grammar and vocabulary and the range of vocabulary and grammatical structures). The NNES teachers' form-focus and SE-orientation is at odds with recent descriptions of WEs or EFL communication, where it has been argued that in real world communication getting the message across is what matters for NNES participants or EFL users rather than approximation to NS models (Jenkins, 2000). This finding indicates that, in spite of the WEs movement's attempt to legitimize non-standard varieties of English, the ideal NS norm is still the goal or inspiration for L2 learning and assessment, and it may be premature to abandon the NS norm, especially in an EFL context like China (e.g., Han, 2004).

In this study the NNES teachers' belief in the NS norm appears to have been influenced by their understanding of the purpose of learning English in China. For the NNES teachers the aim was to communicate with the NESs in the world, and most of them held a narrow definition of NESs, defining them as those located in America and/or Britain.

It was also seen in the findings that teachers based in China, whether native or non-native speakers of English, shared similar beliefs about English varieties. Firstly, both NNES and NES teachers in China, regardless of their language background, were reluctant to accord Chinese English the status of a language in its own right (in contrast with Australia-based NESs' recognition of Chinese English as an independent language), and at the same time insisted on the status of Chinese as a great language. This could be because the Australia-based NESs have not much knowledge about 'China English', and the teachers in China believe that there is no such a thing as 'China English'. In addition, they also agreed, by and large, in contrast to Australian-based NESs, that the purpose of learning English in China was to communicate with NES speakers in the world (rather than with NNESs in the region), and acknowledged adherence to NS norms when teaching and rating. Secondly, both the NNES teachers and the China-based NES teachers shared the view that American/British English or Standard English was the appropriate target of instruction in China, whereas NES teachers in Australia rejected NS norms as a suitable target.

The more purist attitudes of NNES and NES teachers living in China is perhaps understandable, given that they are not exposed on a daily basis to the varieties of English characteristic of a multicultural society such as Australia and are therefore more prone to look to an idealized

variety such as Standard English. Or it may be that exposure to Chinese varieties of English breeds contempt rather than tolerance. The more permissive attitudes of the Australian NES group may also be attributed to their socialization in an institution which has a strong tradition of acknowledging English as an international language and of according validity to various non-standard varieties. These different perceptions may support the inference in the parallel studies on NNES and NES teacher rater performance (Zhang & Elder, 2011, 2014) that different orientations to English proficiency may be due to the teachers' different social, cultural and educational experiences, deriving from the fact that English is (or is not) the first language used at home or in the wider society. Native speaker-ness, therefore, may be more a question of experience than biology.

However, teachers' persistence of NS norms in China may be challenged with the spread of English use in China. Recent studies show that university students in China are becoming increasingly bilingual, and constantly switch between different worlds using different languages and language varieties as they acquire English at school and outside of their formal curriculum through the Internet, music, computer games, movies and television series (Bolton, 2013). The varieties of English these students are exposed to no doubt press on research on the spread and use of English in China, and at the same time, challenge the teachers' attitude to Standard English. Hence, it is important to educate NNES teachers in China regarding WEs or ELF, so as to raise their awareness of the existence of a whole range of local varieties of English worldwide and to avoid the blind adoption of NS norm (Jin, 2005).

It is worth noting that past experience in rater training as well as rating the local English test like the CET-SET may lead to higher value and expectation of the NES-like English ability, as the response by the NNES teachers with CET-SET rating experience indicated. The NNES and NES teachers' different attitudes towards varieties of English need to be taken into consideration in terms of rater training. This is especially important when both NNES and NES are involved in the rating of the local English test, in view of the fact that NES raters may be tolerant to even non-standard varieties of English and may be more lenient in rating certain aspects of test performance. Furthermore, the raters, regardless of their L1 background, need to be familiar with the local varieties of English and the local context, to reach shared attitudes to English.

The differences in attitudes towards varieties of English between the NNES and NES teachers do not indicate that Chinese NNES teachers are operating according to a different code (same as the findings in the

parallel studies in Zhang & Elder, 2011, 2014). If anything, they may be more form-focused, and more oriented to Standard English than are the NES teachers. This finding indicates that the attitudes and concepts of NNES teachers, who are currently assessing the CET-SET performance, accord with the assessment criteria which embodies the test construct, and hence provides a validity evidence of the test.

Teachers' attitudes to varieties of English may have policy implications for the implementation of formative assessment of English learning in universities. The recent assessment and curriculum reform in Chinese universities requires the incorporation of formative assessment components into the summative assessment framework (Chen & Klenowski, 2009). The purpose of formative assessment is to aid learning and teachers play a critical role in the teaching and formative assessment. As teachers' assessment skills and practice are mainly gained from experience, teachers' attitudes will have impact on setting standards and criteria for assessment. As research indicates that students and teachers tend to believe that college English class should be taught by both NNES and NES teachers since students can benefit from the strengths of both types of teachers (He & Miller, 2011), different attitudes to English between the NNES and NES teachers may result in different teaching and learning outcome as well as the use of different criteria for the formative assessment. The NNES teachers may stick to SE norms and be form-focused, emphasizing linguistic competence perspectives such as accuracy, while the NES teachers may focus on message, and pay more attention to pragmatic features such as appropriacy. This may cause issues in the formative assessment of students' learning progress on what standard of English the teachers should follow, and have washback effect on teaching and learning. The findings of this study may indicate that NES teachers may need to fit into the local context while the NNES teachers need further training so that they will be aware of different varieties of English.

12.5 Conclusion

The rise of English as a world language challenges the concept of the native speaker by raising the question of which variety should constitute the standard. However, what matters in all cases is that 'the community should be confident in choosing its own solution' (Davies, 2003, p. 170). This inspires the research about English in China. What norm do students and teachers refer to there in English teaching, learning and testing? According to the attitudes in the survey responses by

the NES and NNES university teachers in this study, the NS norm is still what teachers pursue in China. The findings of the study shed light on the implications for language policies governing the status of English in China in terms of English teaching, testing and use. This allegiance to standard language norms may explain the preference of the NS-based pedagogic models of English in university in China and the privilege of the SE in the locally-designed tests such as the CET-SET, as well as NNES teachers' greater focus on linguistic accuracy compared to NES teachers, who were more tolerant towards English varieties and may operate broader constructs of communication in testing and teaching.

References

Bai, H. [柏会力]. (2008). 英语全球化背景下我国《大学英语课程教学要求》分析及教材研究 [An analysis of the College English Curriculum Requirements and textbooks in the globalization of English]. *中国高教研究 [China Higher Education Research], 5*, 92–93.

Beneke, J. (1991). Englisch als lingua franca oder als Medium interkultureller Kommunikation [English as lingua franca or as medium of intercultural communication]. In R. Grebing (Ed.), *Grenzenloses Sprachenlernen* (pp. 54–66). Berlin: Cornelsen.

Bolton, K. (2013). World Englishes, globalization, and language worlds. In N. L. Johannesson and G. Melchers (Eds), *Of butterflies and birds, of dialects and genres: Essays in honour of Philip Shaw* (pp. 227–252). Stockholm: Acta Universitatis Stockholmiensis.

Brutt-Griffler J., & Samimy, K. K. (2001). Transcending the nativeness paradigm. *World Englishes, 20*(1), 99–106.

Bruthiaux, P. (2003). Squaring the circles: Issues in modeling English worldwide. *International Journal of Applied Linguistics, 13*(2), 159–178.

Cai, J., & Tang, M. [蔡基刚 & 唐敏] (2008). 新一代大学英语教材的编写原则 [Designing principles of new generation college English textbooks]. 大学英语教学 [College English Teaching], 4, 85–90.

Chen, Q., & Klenowski, V. (2009). Assessment and curriculum reform in China: The college English test and tertiary English as a foreign language education. In: *Proceedings of the 2008 AARE International Education Conference*, 30 November–4 December 2008, Queensland University of Technology, Brisbane.

Clark, E., & Paran, A. (2007). The employability of non-native-speaker teachers of EFL: A UK survey. *System, 35*(4), 407–430.

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly, 33*(2), 185–209.

Cortazzi, M., & Jin, L. (1999). Cultural mirrors: Materials and methods in the EFL classroom. In E. Hinkel (Ed.), *Culture in second language teaching and learning* (pp. 196–220). Cambridge: Cambridge University Press.

Crystal, D. (1997). *English as a global language*. Cambridge: Cambridge University Press.

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press.
Dalby, D. (2001). The linguasphere: Kaleidoscope of the world's languages. *English Today, 17*(1), 22–26.
Davies, A. (1991). *The native speaker in applied linguistics*. Edinburgh: Edinburgh University Press.
Davies, A. (1996). Proficiency or the native speaker: What are we trying to achieve in ELT? In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 145–159). Oxford: Oxford University Press.
Davies, A. (1999). Standard English: Discordant voices. *World Englishes, 18*(2), 171–186.
Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon: Multilingual Matters.
Davies, A. (2004). The native speaker in applied linguistics. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 431–450). Malden, MA: Blackwell.
Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes, 22*(4), 571–584.
Dong, Y. [董亚芬] (1992). 大学英语教学的回顾与展望[College English teaching: What we've done and what we're looking forward to]. *外语界 [Foreign Language World], 47*(3), 23–26.
Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics, 26*, 282–304.
Elder, C., & Harding, L. (2008). Language testing and English as an international language: Constraints and contributions. *Australian Review of Applied Linguistics, 31*(3), 34.1–34.11.
Feng, A., & Byram, M. (2002). Authenticity in College English textbooks: An intercultural respective. *RELC Journal, 33*(2), 58–84.
Gnutzmann, C. (2000). Lingua franca. In M. Byram (Ed.), *The Routledge encyclopedia of language teaching and learning* (pp. 356–359). London: Routledge.
Graddol, D. (1997). *The future of English?: A guide to forecasting the popularity of the English language in the 21st century*. London: British Council.
Graddol, D. (1999). The decline of the native speaker. In D. Graddol & U. Meinhof (Eds.), *English in a changing world. AILA review 13* (pp. 57–68). Guildford: Biddles.
Graddol, D. (2006). *English next: Why global English may mean the end of 'English as a foreign language'*. London: British Council.
Han, Z. (2004). To be a native speaker means not to be a nonnative speaker. *Second Language Research, 20*(2), 166–187.
He, D., & Li, D. C. S. (2009). Language attitudes and linguistic features in the 'China English' debate. *World Englishes, 28*(1), 70–89.
He, D., & Miller, L. (2011). English teacher preference: The case of China's non-English-major students. *World Englishes*, *30*(3), 428–443.
Hu, X. (2004). Why China English should stand alongside British, American, and the other world Englishes. *English Today, 20*(2), 26–33.
Hu, X. (2005). China English, at home and in the world. *English Today, 21*(3), 27–38.
Jenkins, J. (1998). Which pronunciation norms and models for English as an International Language? *ELT Journal, 52*(2), 119–126.

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*(1), 89–103.
Jenkins, J. (2003). *World Englishes: A resource book for students*. New York: Routledge.
Jin, J. (2005). Which is better in China, a local or a native English-speaking teacher? *English Today, 21*(3), 39–46.
Jin, Y. [金艳] (2000). 大学英语四、六级考试口语考试对教学的反拨作用 [The washback effects of College English Test-Spoken English Test on teaching]. *外语界 [Foreign Language World], 80*(4), 56–61.
Kirkpatrick, A. (2002). ASEAN and Asian cultures and models: Implications for the ELT curriculum and for teacher selection. In A. Kirkpatrick (Ed.), *English in Asia: Communication, Identity, Power & Education* (pp. 213–24). Melbourne: Language Australia Ltd.
Kirkpatrick, A., & Xu, Z. (2002). Chinese pragmatic norms and 'China English'. *World Englishes, 21*(2), 269–279.
Lazaraton, A. (2005, September). *Non-native speakers as language assessors: Recent research and implications for assessment practice*. Paper presented at the 38th annual meeting of the British Association of Applied Linguistics, Bristol.
Liao, Y. [廖亦斌] (2008). 外籍教师对中学生英语学习作用的调查 [A survey on the importance of foreign language teachers to high school students]. *教育与管理 [Journal of Teaching and Management], 18*, 63–64.
Li, D. C. S. (2007). Researching and teaching China and Hong Kong English. *English Today, 23*(3–4), 11–17.
Lin, J. [林娟] (2006). *我国大学生眼中的英语 [The image of English in Chinese college students' eyes: Their attitudes towards English in China]*. 吉林大学硕士论文 Unpublished MA thesis, Jilin University, Jilin, China.
Liu, J. (1999). Nonnative-English-speaking professionals in TESOL. *TESOL Quarterly, 33*(1), 85–102.
Little, D., Devitt, S., & Singleton, D. (1989). *Learning foreign languages from authentic texts: Theory and practice*. Dublin: Authentik.
Lowenberg, P. H. (2002). Assessing English proficiency in the expanding circle. *World Englishes, 21*(3), 431–435.
Lukmani, Y. (2002). *English in India: Assessment issues*. Paper presented at the Hong Kong Seminar, Hong Kong.
McKay, S. (2002). *Teaching English as an International Language*. Oxford: Oxford University Press.
Medgyes, P. (1994). *The non-native teacher*. London: Macmillan.
Medgyes, P. (1999). Language training: A neglected area in teacher education. In G. Braine (Ed.), *Non-native educators in English language teaching* (pp. 177–191). Mahwah, NJ: Lawrence Erlbaum.
Ministry of Education. (2004). *大学英语课程教学要求（试行）* [*College English curriculum requirements (For trial implementation)]*. Beijing: Tsinghua University Press.
National College English Testing Committee. (1999). *大学英语口语考试大纲 [College English Test-Spoken English Test (CET-SET) syllabus]*. Shanghai: Shanghai Foreign Language Education Press.

Phillipson, R. (1992). *Linguistic imperialism*. Oxford: Oxford University Press.
Quirk, R. (1990). What is Standard English? In R. Quirk & G. Stein (Eds.), *English in use* (pp. 112–125). London: Longman.
Rampton, M. B. H. (1990). Displacing the 'native speaker': Expertise, affiliation, and inheritance. *ELT Journal, 44*(2), 97–101.
Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics, 11*(2), 133–158.
Shavelson, R.J., & Stern, P. (1981). Research on teachers' pedagogical judgements, decisions, and behaviour. *Review of Educational Research*, 51(4), 455–498.
Tang, C. (1997). The identity of the nonnative ESL teacher: On the power and status of non-native ESL teachers. *TESOL Quarterly, 31*(3), 577–580.
Tang, J., Xing, Q., & Yang, P. [唐建国, 邢倩&杨平] (2005). 大学英语教材中中国题材问题初探 [A preliminary research on Chinese theme in College English textbooks]. *大学英语（学术版） [College English (Academic Edition)]*, 302–304.
Timmis, I. (2002). Native-speaker norms and International English: A classroom view. *ELT Journal, 56*(3), 240–249.
Todd, W. R., & Pojanapunya, P. (2009). Implicit attitudes towards native and non-native speaker teachers. *System, 37*(1), 23–33.
Tsui, A. B. M., & Bunton, D. (2000). Discourse and attitudes of English teachers in Hong Kong. *World Englishes, 19*(3), 287–304.
Wang, Y. (1999). 'College English' in China. *English Today, 15*(1), 45–51.
Widdowson, H. G. (1994). The ownership of English. *TESOL Quarterly, 28*(2), 377–389.
Xu, W., Wang, Y. & Case, R. E. (2010). Chinese attitudes towards varieties of English: A pre-Olympic examination. *Language Awareness 19*(4), 249–260.
Zhai, X. [翟向俊] (1986). 大学英语精读课程 [Intensive reading course of College English]. *外语界 [Foreign Language World], 23*(4), 25–28.
Zhang, B., & Ma, L. [张蓓 & 马兰] (2004). 关于大学英语教材的文化内容的调查研究 [A survey on the cultural content of college English teaching materials]. *外语界 [Foreign Language World], 4*, 60–66.
Zhang, W., & Hu, G. (2008). Second language learners' attitudes towards English varieties. *Language Awareness, 17*(4), 342–347.
Zhang, Y., & Elder, C. (2009). Measuring the speaking proficiency of advanced EFL learners in China: The CET-SET solution. *Language Assessment Quarterly, 6*(4), 298 – 314.
Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31–50.
Zhang, Y., & Elder, C. (2014). Investigating native and non-native English speaking teacher raters' judgments of oral proficiency in the College English Test-Spoken English Test 9CET-SET), *Assessment in Education: Principles, Policy & Practice, 21*(3), 306–325.

13

The Power of General English Proficiency Test on Taiwanese Society and Its Tertiary English Education

Shwu-Wen Lin

13.1 Introduction

Reacting to the government's policies to increase Taiwanese university students' international competitiveness by raising their English proficiency, universities in Taiwan have set up a graduation requirement of English proficiency in recent years. This chapter reports on how the implementation of the English graduation requirement has affected the university students and the English curriculum. The requirement accepts scores from various English proficiency tests as proof of proficiency, instead of one particular test. Thus, the findings of this study have implications as to what determines the strongest washback that any language test can have in the context of multiple tests existing and competing for influence.

13.1.1 Research context

The implementation of the graduation requirement for English proficiency originated from the idea of establishing a common index of English proficiency for university students in order to promote global competitiveness. In 2004, the Ministry of Education sent an official document to all universities, encouraging them to include English proficiency as a criterion for student graduation (Zhang & Tu, 2007). Since then, universities began to require their students to pass an external test at a designated level of English language proficiency. However, the autonomy of the universities allows them to differ in not only their approaches to attending the government policies but also the entailed details of the actual requirements. Despite the potential differences, most graduation requirements are similar, in that they state which

English proficiency tests are acceptable and at what level. What needs to be noted is that most of the tests included in the requirement are proficiency tests that are not tied to any university English curriculum.

Among the tests clearly stated in the requirement, GEPT seems to be one of the most important (Shih, 2006, 2007, 2008, 2010), because of its popularity as one of the major English tests taken by the Taiwanese people (LTTC, 2013; Vongpumivitch, 2009). Thus, the GEPT could be the most influential English proficiency test for undergraduates in Taiwan. By exploring the impact of the requirement on the university students and the English curriculum this study aims to investigate whether the GEPT has brought about the greatest washback and whether there is washback of other English language tests in the universities. Through this investigation, this study hopes to provide insights on improving English teaching and learning in universities in addition to understanding test washback and impact further.

13.1.2 The General English Proficiency Test

The GEPT is an English proficiency test developed by the Language Training and Testing Centre (LTTC) in 1999, commissioned by the Ministry of Education with the original goal of promoting life-long learning in English. The GEPT official website states that GEPT test scores can now be used for a variety of purposes, including job searching, career advancement, university entrance and exit. It is a criterion-referenced test based on communicative approach with five levels: elementary, intermediate, high-intermediate, advanced and superior. There are two stages of the test, and test takers have to pass the first stage in order to be advanced to the second stage. The first stage consists of the listening and reading components while the second stage consists of the writing and speaking components (The only exception is the elementary level which includes writing component).

As stated in the general description of the high-intermediate level, test takers who pass this level have the proficiency equivalent to CEFR B2 (see https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/hi_intermediate.htm), which is described in Vongpumivitch (2009) as the level targeting non-English major undergraduates. A search of the requirement regulations in Taiwanese universities also shows that universities with a ranking of above average mostly set up the high-intermediate level of the GEPT as the standard in their requirement. However, some universities may accept a pass at GEPT intermediate level, which is equivalent to a high school graduate's English proficiency.

13.2 Literature review

The educational context of this study presents two interesting topics that have received little attention in washback studies so far. The requirement accepts scores of not just one test, rather, students can provide evidence of their English language proficiency from any of the English proficiency tests listed in the requirement. Most previous studies centre on the influences of one particular high-stakes test or assessment system, which is closely related to the curriculum (Alderson & Hamp-Lyons, 1996; Cheng, 2005; Green, 2007; Wall, 1996; Wall & Alderson, 1993; Watanabe, 1996). In the very few studies that have probed into the effects of more than one test (Shohamy, 1993; Shohamy, Donitsa-Schmidt, Ferman, 1996; Watanabe, 1996, 1997, 2001, 2004), the contents of those tests are still aligned with the prescribed curriculum. However, none of the tests stated in the graduation requirement in this research context are developed according to Taiwanese university English curriculum. This presents an interesting and rare opportunity to explore which test among the list of English proficiency tests has the strongest degree of washback and why.

The implementation of the English requirement for graduation is one of the recent developments in English curriculum in Taiwanese universities (Shih, 2007) and has received little formal research. Recent washback studies in Taiwan have investigated more in the context of high schools than universities (Chen, 2002; Wu and Chin, 2006). Wu and Chin (2006) explored the potential washback of the General English Proficiency Test (GEPT) but they centred on senior high school English curriculum. The more recent study by Shih (2006, 2007, 2009, 2010) investigated the GEPT washback on learning in the context of higher education. The context of his study was similar to the present study, also a case study of two universities, but his research investigated washback from the GEPT only. His study revealed that the GEPT only brought about limited a degree of washback on learning. Students spent no more than two months on GEPT preparation, unlike the year-long preparations they spent on their university entrance examination. Shih attributed the lack of strong washback to the fact that his participants were all English majors, a limitation of his study. How the GEPT may influence the majority of university students, the non-English majors, remains unexplored.

Tsai and Tsou (2009) probed into learners' viewpoints on the adoption of standardised English language proficiency tests as a tool to assess their English competence for graduation purposes. They collected questionnaires from 520 university students of different academic fields in a

technical university. Their findings suggested that instead of using tests as the only tool, there should be multiple measures in assessing university students' English proficiency for graduation purposes. Tsai and Tsou did not limit their participants to English majors. However, the questionnaire they designed contained only nine items, which is limited in depth in its exploration of learners' perceptions of the graduation requirement and the impact it might bring, or have brought to them.

The present study attempts to fill the gaps in the above studies by adopting a case study approach to investigating the impact of the graduation requirement for English proficiency in Taiwanese universities, with a particular focus on non-English majors.

13.3 Research methods

This study employed an ethnographic case study approach to research design, for the purpose of encapsulating the complexity and the depth of washback produced by different tests in the graduation requirement in different institutional contexts. Two universities were selected as cases. Case A was a relatively new university with a very low ranking. It did not require its non-English majors at the undergraduate level to provide evidence of English proficiency for graduation. Case B was a university with a hundred year history, which received much higher ranking than Case A. This university, on the other hand, implemented the requirement that non-English majors should pass the GEPT intermediate level or its equivalents upon graduation (See more details of the regulations for the graduation requirement in Appendix 13.1).

To explore whether the implementation of such requirements had washback in the two cases, this study collected data through observations in the real classroom setting, and interviews of both teachers and students.

13.3.1 Participants

Altogether there were seven teachers who took part in both the classroom observations and the interviews. In Case A, the participants were four teachers teaching 'English Integrated Skills Training' (the only EAP course for non-English majors in Case A). In Case B, they were two teachers who taught the general EAP course and one who taught the test-related remedial course, 'English Reading and Writing'. Eighteen students, who were non-English majors from a variety of departments, were interviewed (nine in Case A and nine in Case B).

Table 13.1 shows the pseudonyms of the participants involved in this study.

Table 13.1 Teacher and student participants in this study

	Teachers	Students
Case A	Adam, Alice, Amy, Anna	Aaron, Abel, Aiden, Alex, Archer, April, Alvin, Andrew, Anson
Case B	Becca, Ben, Betty	Bess, Bianca, Billy, Blair, Bonnie, Brad, Brenda, Bridget, Bryan

13.3.2 Data collection and analysis

This study took on an ethnographic approach to classroom observation due to the nature of the study and also the complexity the graduation requirement encompassed. The ethnographic classroom observation allows context-specific evidence and the more subtle and covert forms of washback to emerge in data collection. Classroom observation data collected comprised 17 lessons of the seven teacher participants. The 13 lessons of Case A were all of the same 'English Integrated Skills Training' course, captured in ten video and three audio recordings. All the four lessons of Case B were audio-recorded data, with three lessons of the EAP course and one lesson of the test-related remedial course, 'English Reading and Writing'. The following were collected and provided as supplementary to the observation data: field notes during the observation, during private talks with teacher participants and teaching materials used in the lessons.

Semi-structured interviews were conducted with the teachers after lesson observation. The interviews focused on eliciting their perceptions of the influences brought about by the graduation requirement, their attitudes towards the requirement and also their comments on their observed lessons. During the observations and the interviews, when issues related to other stakeholders such as parents or publishers arose, they were further probed in the interviews.

Instead of a questionnaire with a limited number of items, semi-structured interviews are conducted for more an in-depth investigation of non-English majors' perceptions towards the graduation requirement. In addition, the interviews with the 18 students probed further into which test accepted by the requirement was considered as most influential and the factors that motivated them to learn English. The student interviews were used to triangulate with the teacher interview data, for exploring the impact of the requirement on the students from both students and teachers perspectives.

There were two main sets of data for analysis. The first set consisted of the observation data and the teacher interviews, which were analysed to explore the influence of the graduation requirement on the non-English majors' EAP curriculum. Pertaining to the ethnographic approach to classroom observation, the first stage of the analysis started with a more grounded approach and went through a reiterative process of inductive coding. Such analysis allows not only the macro impact of the requirement, but also the micro impact of different tests on the curriculum to emerge from the raw data, including the lesson observations, field notes, teaching materials and teacher interview data. The first stage of coding and data analysis revealed an implicit evidence of GEPT washback mentioned by the teachers in Case A and thus resulted in further collection of test papers used in that university. After the initial data analysis, the observed lessons were then coded with Transana (Woods & Fassnacht, 2012), a piece of software for analysing qualitative video and audio data. The observed lessons were segmented according to related activities. Next, the segments were coded with keywords derived from the initial data analysis and literature review. Lastly, the recorded lessons were presented in the form of a timeline, in order to depict their chronological flows to provide a more structured, systematic arrangement of the data, supplementary to the main grounded analysis. The second set of data was the teacher and student interviews. Inductive coding of the data prioritized the participants' voice and provided the emic perspective of how teachers and students perceived the graduation requirement and its impact on the students.

13.4 Findings and discussions

13.4.1 Impact on the EAP curriculum

The analysis of the lessons, triangulated with the teacher interviews, private talks and field notes in both universities and the subsequent collection and analysis of the test papers in Case A revealed the following findings concerning the impact of the graduation requirement on the EAP curriculum.

Despite the fact that the graduation requirement did not specify any particular English proficiency test, GEPT seemed to have a more profound impact than other tests that were also acceptable by the requirement. There was evidence of GEPT washback on teaching and assessment materials, and there was significantly more evidence of the GEPT washback in Case B than in Case A. Case B, which had implemented the graduation requirement for a few years, included a remedial

course that was directly linked to the GEPT (See Section 3). The test affected the teacher, Becca, on the choice of teaching materials, course planning and assessment. Becca knew the course she taught was a remedial course and was suggested to be oriented towards GEPT preparation. She thus adopted the GEPT preparation textbook as the main teaching material and incorporated the GEPT tests as pre-test (a mock test) and post-test (a test delivered by the Language Training and Testing Center) for that class. Ben chose to use commercially-available, monthly-issued English learning magazines as the teaching materials. The magazines incorporated GEPT elements, i.e. GEPT practice items, topics and contents related to local culture (See also 4.3).

On the other hand, in Case A, which did not impose the graduation requirement on non-English majors, there was little explicit evidence of GEPT washback on its EAP curriculum for non-English majors. The only exception was the washback on testing materials. The local Taiwanese publisher, Tunghua Books, developed the mid-term and final test papers modelled after GEPT test item types (See Appendix 13.1) so as to promote the sales of the international EFL teaching materials they introduced. Thus, it seems that this facet of washback in Case A was not mediated by the teachers, but by the publisher instead (See also 4.3).

Another finding was that the GEPT washback was relatively intensive on some aspects. As distinct from several previous studies (Alderson and Wall, 1993; Cheng, 1997, 2005; Stecher, Chunm, and Barron, 2004), the present study found little evidence of curriculum narrowing, focusing only on skills tested, or changing teaching activities as intended by the introduction of a test. This was even true in the GEPT-related remedial course in Case B. Becca, who taught the course, incorporated a listening activity of a pop song and a speaking activity when she discussed the food pyramid with her students, even though her course was designated to focus on reading and writing. The analysis of the language focus and language skills targeted by the activities in the observed lessons revealed individual differences among the teachers of the same courses. For example, all the four teachers in Case A teaching the same course demonstrated a different focus on language skills: Alice on pronunciation, Adam on vocabulary, Anna on listening and conversation, and Amy on integrated skills. Such differences were also observed in Case B teachers, where pronunciation and read-aloud sessions could be seen in Ben's lessons but not in Betty's, even though they taught the same course. Teachers' individually different responses to English learning, test-taking or the graduation requirement were more highlighted than the influences of the GEPT or other tests on teaching. Similarly, there

was little evidence that teaching methods were significantly affected by any single test accepted by the graduation requirement.

Individual differences between the teachers were evident not only in their lessons, but also the interviews. This finding reflects the 'washback variability' (Green, 2006) among the teachers. The teachers' lessons reflected different degrees of washback. Their perceptions of the GEPT washback and the impact of the graduation requirement impact also varied. For example, Betty and Alice were supportive of the requirement to motivate students to learn. Betty said, 'I think that the graduation requirement can urge the students to re-evaluate their language proficiency, to see if they need to enhance their ability.' On the contrary, Amy believed that students with low English proficiency who were not already motivated in English learning would hardly change even with the requirement. Adam further pointed the loopholes of the requirement, by mentioning that remedial courses were 'just an easy way out'. The findings of this study also provided evidence on the roles that teacher factors can play in explaining the presence or absence of washback on teaching (Watanabe, 1996, 2004). The teacher factors that mediated or prevented washback from happening could be seen from the following findings. First, whether the teachers would urge students to take English proficiency tests and boost their English proficiency depended on teachers' perceptions of how effective the graduation requirement could be. Second, the teachers' willingness to comply with what the requirement demanded and their beliefs in what they should teach explained why there was washback in some teachers' classes but not in others. Third, the teachers' preference for one test over the others determined if there was washback of GEPT or other tests on teaching.

13.4.2 Impact on the students

The individual interviews with the 18 non-English majors (nine in each university) were analysed in relation to the influence of the GEPT or other English tests on their learning. The findings were also triangulated with related parts of the teacher interviews.

According to the students, the GEPT was considered as the most important test among all the tests accepted by the graduation requirement. Students aligned the requirement with passing the test, and the majority of the interviewees from Case B had either taken the test or had planned to do so. Students had insufficient knowledge of other English tests and what scores were set as benchmarks for graduation. Most of them also took it for granted that the graduation requirement was equivalent to a GEPT requirement. Such findings indicated a strong

presence of the GEPT influence. This also applied to the students in Case A. There were also references to TOEFL and TOEIC when the learners considered which English proficiency test to take. However, students from both Case A and B, such as Archer and Bess, mainly associated GEPT with the requirement. TOEFL and TOEIC were associated more with their own academic plans in the future.

The interview data of both teachers and learners further revealed differences between the perspectives of the two groups, and between the learners themselves. First, similar to the previous washback studies on learners, the findings of this study suggested that the learners viewed the impact of the requirement on their English learning differently from their teachers. From a negative perspective, the teachers were sceptical of its intended effect in promoting the students' English learning. The teachers were concerned with the possible adverse effects the requirement might bring, and warned about loopholes in the requirement, such as students avoiding taking any external English language tests. Nevertheless, what concerned the teachers was not necessarily what the students cared about. The learners were more concerned about the difficulty in meeting the benchmark set by the graduation requirement and the role of the requirement in their English learning during their university years.

The learner interview data shed further light on the individual differences in their perceptions of the graduation requirement and its impact on the learners. Their attitudes towards the implementation of the graduation requirement varied according to how they perceived the compulsory nature of the requirement and the entailed stakes. Three of the students acknowledged the need and the benefits for such regulations to compel them to learn more English (Aiden, Bess, Brenda); yet, others questioned the necessity of making it compulsory for their graduation (April, Billy). The perceived stakes of fulfilling the requirement could also affect the learner's attitudes towards the implementation of the requirement. Those who believed that their English was not good enough for them to reach the benchmark were reluctant to accept the implementation (Andrew, Bonnie), whereas those who considered it easy to fulfil the requirement did not take it too seriously (Bianca, Blair, Brad, and Bridget).

The learners' perceptions of the impact of the graduation requirement also revealed some individual differences. There were students who perceived little impact but there were also students who associated stress and anxiety with the compulsory requirement in order to obtain their degrees. Those students who considered themselves to be little

influenced, such as Brad and Bianca, were more concerned with their opportunities to learn more English than with being required to provide official proof of their English proficiency. For them, their motivation to learn English would not be influenced much by the implementation of the graduation requirement as they had their own learning goals, and the requirement would only make them work harder for this particular high-stakes purpose. The rest of the students, on the other hand, associated stress and anxiety with the graduation requirement. However, they were still quite different in terms of how they chose to face the implementation. Some viewed the pressure accompanying the requirement as a positive force on their English learning and thus, welcomed the implementation (Abel, Anson, Aiden, Bess). There were those who did not support the implementation as they disliked the 'side effects' the requirement might bring (Andrew, Billy, Bonnie). Others were not supportive either, not so much because they had negative feelings towards the implementation, but because they feared that they would not be able to meet the requirement, because of their poor proficiency and their past failures in test-taking (Aaron, Alex, Bryan).

13.4.3 Social impact of GEPT

The social impact of the GEPT was evidenced in this study through different types of stakeholders, beyond or within the university system. External stakeholders included the parents, the publishers of monthly-issued English learning magazines for all citizens, and also the publishers that represented the international EFL teaching material. Stakeholders within the university system, like non-English majors, might also take the GEPT for purposes other than fulfilling the graduation requirement.

The most explicit evidence of the social impact of the GEPT is within the community of local publishers who develop English learning magazines for lifelong learning. An important feature that stems from the commercial nature of those monthly-issued magazines is to cater for the current needs of the potential buyers in order to promote sales. With the popularity of the GEPT in society, the magazines that were once not test-oriented have been changed into materials that can prepare readers for the GEPT. The inclusion of the GEPT-related contents and practice items that explicitly refer to the test in the issues is thus a reflection of how big the influence the test is on Taiwanese society. In this study, Ben used the magazines as the teaching materials in his lessons, but denied having the intention to prepare his students for the test. He further explained that he chose to use the magazines because the assigned textbook for his course was not interesting enough while the magazines

provided articles of different and up-to-date topics every month. It was clearly evidenced in his lessons that there was no sign of test preparation, practice of mock test items or any reference to the test. Thus, since teachers use the magazines with no specific purpose for test preparation, it seems that the washback of the GEPT test on the teaching materials is a product of the social impact of the test on the magazines.

Similarly, local publishers who represent international EFL materials designed for institutional use can also come under the strong social impact of the GEPT. The publishers here are different from those mentioned above, in that they import international EFL materials and promote the sales of those materials in local educational settings. The findings of this study revealed that the evidence for the washback of GEPT was not in the teaching material itself, but in the mid-term and final tests developed by the local publisher. The item types in these tests were modelled on GEPT item types (See Appendix 13.2 for a comparison of the test instructions for the 'question and statement response' section. See also an item of the listening test developed by the local publisher and the equivalent in the practice GEPT test provided by the official website: http://www.lttc.ntu.edu.tw/geptpracticee.htm).

The comparison shows that there is little difference in the instructions between the two tests, and the test item in the test paper used in Case A is exactly the same as that in the GEPT. The teachers revealed that the test papers were developed by the local publisher based on the contents of a non-GEPT-related, international teaching material. Thus, the local publisher's attempt to model a small part of the test after the GEPT, while preserving the internationally-recognised contents, suggested how the local publishers adapted the global materials to meet the locals' needs. The local publisher's action was clearly evidence of the social impact of the GEPT, and via the test the local publisher designed, the GEPT has exerted indirect impact on the testing of the universities.

The social impact of the GEPT can also be realised via the students' parents, who can influence the learners on the selection of which English proficiency test to take. The student interviews revealed that it was the parental influence that made some students prioritise the GEPT over other English tests, as illustrated by Bridget's case. Her decision to take the GEPT instead of other tests was because her father asked her to do so. Although the graduation requirement did not favour the results of one test over another, the parents' eagerness for their children to pass the GEPT was probably attributable to the higher value that the parents attach to the GEPT over other English proficiency tests available. Thus, parental influence in the context of this study can be considered as a

manifestation of the social impact of the GEPT on parents, which in turn shaped the impact of the test on the learners.

Learners themselves may also directly feel the social impact of the GEPT. For example, Alvin claimed that he would choose to take the GEPT instead of other English proficiency tests, mainly because he considered a certificate of English proficiency test as essential for job searching. His assumption of the GEPT as being a test that will be accepted or requested by future employers illustrates the strong impact of the GEPT on the society generally.

The above findings suggested that the GEPT probably has much stronger impact on the stakeholders from general society, like parents or local publishers, than the university teachers and students in the context of the requirement of English language proficiency. The government's policies to increase university students' international competitiveness, and the accompanying measures the universities have taken to boost their students' English proficiency, reinforced, albeit indirectly, the impact of the GEPT on teaching and learning within the universities.

13.4.4 Reinforcement by graduation requirement

The findings suggested that the implementation of the English graduation requirement reinforces the social impact of the GEPT in two ways. Firstly, for the majority of the students, the need to provide a proof of English proficiency for graduation is an imperative of the need to pass the GEPT. In other words, the students are compelled to select the GEPT over many other English proficiency tests because of the perceived high social status of the GEPT. This may be the reason why the teachers speculate that their students are most likely to take the GEPT. The direct alignment of passing the GEPT with fulfilling the graduation requirement by both teachers and students is the evidence of the strong impact of the GEPT, being reinforced in the university system.

Secondly, through the hands of the curriculum designers, who are usually the English department in universities, the English curriculum may be affected by the GEPT. The graduation requirement introduced some intended curriculum changes like preparation or remedial courses. The social impact of the GEPT is reinforced in English teaching and learning when the curriculum designers shape the direction of these courses to focus particularly on the GEPT but not other tests. An example is the remedial course in Case B (See also 4.1). Despite having a generic course title (English Reading and Writing) that did not suggest a link to any test, teachers were instructed by the English department to

incorporate GEPT contents and related teaching materials. The original course aim was to provide assistance for students who failed to meet the graduation requirement. However, the aim was narrowed down to assist them particularly to pass the intermediate level of the GEPT. In this case, the washback of the GEPT manifested in those courses is mediated by the curriculum designers who are influenced by the power of the GEPT in the society.

13.5 Conclusion

This study offers some explanations for how one particular test can exert the strongest influence when test takers may choose from a number of tests. The conditions or purposes (e.g. admission, promotion, placement or graduation) of a test, as Madaus (1990) pointed out, can determine whether a test is of high stakes or not. However, this cannot be fully applied in determining the stakes that any test receives in the context of this study. When students are given the liberty to choose any test stated in the graduation requirement, each test can be considered as high stakes. However, the findings suggest that in addition to test use, the importance of a test in society is another factor to consider. What makes the GEPT stand out from the internationally recognised English proficiency tests such as TOEFL or TOEIC is the wide recognition of GEPT among Taiwanese citizens. This may be the reason why Taiwanese parents prefer their children to take the GEPT, the test they know most about. In addition, unlike TOEFL, TOEIC, or IELTS, which are usually used for a certain purpose like further studies abroad, the GEPT test scores can be used in a wider range of areas in the Taiwanese society such as university admission, job application or governmental scholarship application. The GEPT is thus perceived by the majority of stakeholders as very important in the society, showing its strong social impact. The implementation of the graduation requirement, which uses English test results for high stakes purposes, further reinforces the already strong social impact of the GEPT in the university system. The advantage of test scores for multiple uses and the status of the test in society are similar to Gates' (1995) ideas of test 'utility' and 'monopoly' that can determine the extent of washback intensity. Nevertheless, both Gates' ideas refer to the social context in which a test is used and how important a test is in society. Therefore, in the high stakes context where stakeholders are given multiple choices, the most eminent washback would be from the test the stakeholders perceived as the most important in society. Since language testing 'is and always has

been a social practice' (McNamara & Roever, 2006), such social agendas embedded in a language test should be taken into consideration when designing and implementing an educational policy involving the test concerned.

Appendix 13.1 Regulations for promoting students' English proficiency

提升學生英語能力實施辦法

第三條　本校學士班學生，除依第五條得免修者，應於三年級起修習2學期之英語能力補強課程 … 其他學系之英語能力補強課程為「英語會話聽力」及「英文閱讀寫作」。

Regulation no.3:
All students who are receiving the Bachelor's degree, besides those who can have exemption according to Regulation no.5, are required to take two semesters of English remedial courses in their third year … English remedial courses for non-English departments are 'English Conversation and Listening' and 'English Reading and Writing'. (Note: Contents of the regulation not related to non-English majors are omitted.)

第四條 英語能力補強課程為必修，零學分，及格分數為60分，課程未通過者，必須重修，全部通過者始得畢業。

Regulation no.4:
The English remedial courses are required courses with no credits. The fail and pass grade is 60. Those who failed the courses have to re-take the courses and only those pass all the courses can be graduated.

第五條　本校學生符合下列條件之一，並於三年級加退選時間結束以前繳交相關證明文件，得免修英語能力補強課程：
2.其他學系學生：
(一)全民英檢（GEPT）中級初試（含）以上。
(二)托福（紙筆測驗）457（含）以上。
(三)托福（電腦測驗）137（含）以上。
(四)托福（網路測驗）47 分（含）以上。
(五)多益測驗（TOEIC）550（含）以上。
(六)國際英語測驗（IELTS）4 級（含）以上。
(七)其他經教務處及相關科系認可之語文檢定或測驗標準。

Regulation no.5:
Students in this university can receive exemption from English remedial courses by reaching one of the standards listed below and by providing documents of proof before the end of the add and drop period (i.e. when students decide which course to take) in their third year.

2. Non-English majors:
(1) GEPT Intermediate Level 1st Part and above
(2) TOEFL (paper-version) 457 and above
(3) TOEFL CBT 137 and above
(4) TOEFL iBT 47 and above
(5) TOEIC 550 and above
(6) IELTS 4 and above
(7) Other proficiency tests or standards approved by the Office of the Academic Affairs and related departments.

Appendix 13.2 Comparison between University A test item and the GEPT elementary level

Listening	University A test paper	GEPT elementary level
Question and statement response: test instruction	每題請聽錄音機播出一英語問句或直述句之後，從試題冊上A、B、C三個回答或回應中，找出一個最適合的作答。每題只播出一遍。(For each item, please listen to a question or a statement from the audio recorder. Choose the most appropriate answer from the three answers or responses in a, b, and c. Each item is played only once.)	請聽錄音機播出一個問句或直述句後，從下面A、B、C三個回答或回應中，找出一個最適合的作答。每題只播出一遍。(Please listen to a question or a statement from the audio recorder. Choose the most appropriate answer from the three answers or responses stated in a, b, and c. Each item is played only once.)
Sample test item	1. (Audio: How often do you clean the house?) A. Yes. My house is very clean. B. Twice a week. C. We usually clean the house on Sunday.	1. (Audio: Who's that tall handsome man?) A. He's studying. B. He's my cousin. C. He's not very happy.

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*(2), 115–129.
Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education*, *11*(1), 38–54.
Cheng, L. (2005). *Changing language teaching through language testing: A washback study*, Cambridge University Press.

Cheng, L., Watanabe, Y. & Curtis, A. (2004). *Washback in language testing: Research contexts and methods*, Mahwah, NJ: Lawrence Erlbaum.

Chen, L. (2002). *Taiwanese junior high school English teachers' perceptions of the washback effect of the Basic Competence Test in English*. Unpublished PhD thesis, College of Education, The Ohio State University, Ohio, United States.

Council of Europe (2001). *Common European Framework of Reference for Languages: learning, teaching and assessment*. Cambridge: Cambridge University Press.

Gates, S. (1995). Exploiting washback from standardized tests. In J. Brown & S. Yamashita, eds. *Language testing in Japan*. (pp. 101–106). Tokyo: Japan Association for Language Teaching.

Green, A. (2006). Washback to the learner: Learner and teacher perspectives on IELTS preparation course expectations and outcomes. *Assessing Writing*, *11*(2), 113–134.

GEPT Research Highlights, 2013. [online] Available at: https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT/files/GEPT_Research_Highlights.pdf [Accessed 30 March 2014].

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner, ed. *Critical issues in curriculum* (pp. 83–121). Chicago, Illinois: Chicago University Press.

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

Shih, C.M. (2006). *Perceptions of the general English proficiency test and its washback: A case study at two Taiwan technological institutes*. Unpublished PhD thesis, Department of Curriculum, Teaching and Learning, Ontario Institute for Studies in Education, University of Toronto, Toronto, Canada.

Shih, C.M. (2007). A new washback model of students' learning. *Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes*, *64*(1), 135–161.

Shih, C.M. (2008). The General English Proficiency Test. *Language Assessment Quarterly*, *5*(1), 63–76.

Shih, C.M. (2010). The washback of the General English Proficiency Test on university policies: A Taiwan case study. *Language Assessment Quarterly*, *7*(3), 234–254.

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

Stecher, B., Chun, T. & Barron, S. (2004). The effects of assessment-driven reform on the teaching of writing in Washington state. In L. Cheng, Y. Watanabe & A. Curtis, eds. *Washback in Language Testing: Research context and methods* (pp. 53–72). Mahwah, NJ: Lawrence Erlbaum.

Tsai, Y. & Tsou, C.H. (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.

Vongpumivitch, V. (2010). The General English Proficiency Test. In L. Cheng & A. Curtis, eds. *English language assessment and the Chinese learner* (pp. 158–172). New York: Routledge.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*, *13*(3), 334–354.

Wall, D. & Alderson, J.C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, *10*(1), 41–69.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, *13*(3), 318–333.

Watanabe, Y. (2001). Does the university entrance examination motivate learners? A case study of learner interviews. In A. Murakami, ed., *Trans-Equator Exchanges: A Collection of Academic Papers in Honour of Professor David Ingram* (pp. 100–110). Akita, Japan: Akita University.

Watanabe, Y. (2004a). Methodology in washback studies. In L. Cheng, Y. Watanabe & A. Curtis, eds. *Washback in language testing: Research context and methods* (pp. 19–38). Mahwah, NJ: Lawrence Erlbaum.

Watanabe, Y. (2004b). Teacher factors mediating washback. In L. Cheng, Y. Watanabe & A. Curtis, eds. *Washback in language testing: Research context and methods* (pp. 129–146). Mahwah, NJ: Lawrence Erlbaum.

Woods, D. and Fassnacht, C. (2012). Transana v2.50. http://www.transana.org. Madison, WI: The Board of Regents of the University of Wisconsin System.

Wu, R., & Chin, J. (2006). An impact study of the Intermediate-Level GEPT. Proceedings of the Ninth International Conference on English Language Testing in Asia, Taipei, 41–65.

Zhang, R. & Tu, Y. (2007). 國內技專校院學生英語能力畢業門檻現況與省思 [The English graduation requirement for students in domestic universities and colleges of technological and vocational education, the present state and review]. Cross-Strait Technological and Vocational Education Conference. Taichung: Chaoyuang University of Technology.

14

Twenty Years of Cambridge English Examinations in China: Investigating Impact from the Test-takers' Perspectives

Xiangdong Gu and Nick Saville

14.1 Introduction

In the wake of the political reforms which opened up China to the outside world in 1978, and with China's entry into the World Trade Organization in 2001, the Chinese economy has become increasingly integrated into the international community. In keeping up with globalisation more generally around the world, English has become widely accepted in China as a utilitarian tool for international mobility, study purposes and career development. This has been reflected in the way that English language teaching and learning has been introduced into compulsory education in China. In 2001, the Chinese government established a national policy whereby children start learning a foreign language (mainly English) from Grade Three in primary school (at age nine), instead of from Grade Seven in junior middle school as before.
In line with these social, political and educational changes, English language education in China has also become internationalised. Along with the restoration of the National Matriculation Test (commonly referred to as Gaokao, in which English is a compulsory subject, with equal status to Chinese and Mathematics since 1983) in 1977 after a ten-year suspension, and the implementation of the National College English Test for students of non-English majors (CET) in 1987, an increasing number of internationally recognised English tests have been introduced into China.

In particular, the focus of this chapter is on the Cambridge English examinations that were introduced into China in the early 1990s: the Cambridge English: Business Certificates, originally known as the Business English Certificates and abbreviated to BEC when they were

introduced in 1993; and the Cambridge English: Key for Schools and Preliminary for Schools examinations which have been available in China since 2009 and are also known as the Key English Test (KET) and the Preliminary English Test (PET) for Schools.

The BEC examinations were among the first international certificates to be introduced, having been specifically developed for use in China in partnership with the National Educational Examinations Authority (NEEA). These examinations now comprise a suite of internationally-recognised certificates that provide a progressive way to develop and improve English ability for use in workplace contexts, including the Preliminary, Vantage and Higher certificates. They are typically taken by young adult learners preparing for a career in business, or seeking to advance their present career where English skills are necessary in their work context. They are tests of English at Levels B1, B2 and C1 respectively in the Common European Framework of Reference for Languages (CEFR) and are qualifications which show that candidates can use English confidently for communication in international business-related contexts. (see www.cambridgeenglish.org/exams/business-certificates/)

More recently, the general English examinations targeting the CEFR A2 and B1 levels for school-age children – Key and Preliminary for Schools were introduced into China in 2009. These examinations are designed for learners aged between 11 and 14 years (Papp and Nicholson 2011) and are qualifications that show that pupils can deal with everyday written and spoken English at basic and lower intermediate levels respectively. (see www.cambridgeenglish.org/exams/general-english-and-for-schools/)

Since 1993, several-million Chinese learners have taken Cambridge English examinations, in part due to the growing demand for English language services in keeping with the expansion of the Chinese economy and its international profile. For example, the annual test-taking population for the Business Certificates increased from less than 10,000 in 1993/4 to over 100,000 in 2012.

To ensure that Cambridge English examinations are fit for their intended purposes, Cambridge English has adopted the concept of "impact by design" as a fundamental principle of good practice, building on the organization's four maxims for achieving and monitoring impact. These are summarised as: Plan, Support, Communicate, and Monitor/Evaluate (Milanovic and Saville, 1996). The principles of good practice emphasize the importance of anticipating the impact that Cambridge English examinations may have on a wide range of stakeholders, and on the test takers in particular, and on adjusting

the examination system appropriately as a result of evidence collected under operational conditions. This is important because the perceptions and behaviours of the stakeholders may affect the validity of the examinations in important ways. Moreover, it may be especially relevant when the examinations are introduced into new contexts where English language assessments carry high stakes and where test-taking constitutes a significant part of the language learning experience – as in China, for example.

Against this background, therefore, this chapter explores the notion of impact in the Chinese context and considers the potential effects and consequences Cambridge English examinations may exert on English language learning in Chinese educational settings, in particular, on the test-takers' perceptions towards the tests, their motivations for taking the tests, and their test-preparation processes which may influence their learning/test-taking behaviours. These perceptions may also influence how learners engage in their language studies and ultimately affect their progress in reaching the level of English proficiency they need for study or career purposes. In addition, it is important to monitor and evaluate the uses of the Cambridge English examinations in Chinese contexts more broadly to understand how the results are interpreted by other stakeholders, such as parents and employers.

To illustrate these issues and the methodological approach being adopted, two empirical studies into the impact of Key and Preliminary for Schools and BEC Vantage and Higher in China are reported. These studies were collaborative projects conducted by two teams led respectively by the authors of this chapter – one from the Research and Validation Group in Cambridge, and the other from Chongqing University in China.

In line with the impact model of Cambridge English, a mixed method approach was used, combining data from structured questionnaires and interviews. The implications for English language learning and assessment in China are considered with reference to test validation procedures, stakeholder communications, and long-term intercultural relationships in the field of language education which seek to integrate local, national and global perspectives.

14.2 Literature review

Broadly speaking, the Cambridge approach adopts Wall's (1997: 291) concept of impact, which is defined as "any of the effects that a test may have on individuals, policies or practices, within the classroom,

the school, the educational system or society as a whole". Impact is therefore a superordinate term encompassing the well-known concept of washback which traditionally has been defined as the influence of testing on teaching and learning (Alderson and Hamp-Lyons, 1996; Alderson and Wall, 1993; Bachman and Palmer, 1996; Davies, Brown, Elder, Hill, Lumley, and McNamara, 1999; Hamp-Lyons, 1997; Hughes, 1989; Messick, 1996; Shohamy, 2001; Shohamy, 1992; Wall, 1997).

What has emerged in recent years is a consensus that impact is a complex phenomenon with wider scope and influence than washback (Saville, 2010) and with implications for the ways in which assessment systems are developed, validated and revised.

For example, as Hawkey (2006: 10) points out:

> … whether impact is intended or unintended, it would seem to be a legitimate and crucial focus of research, both micro and macro, to "review and change" tests and programmes in the light of findings on, among other aspects of programmes or tests, "how the stakeholders use the exams and what they think about them".

Moreover, influenced by Messick's (1989, 1996) unifying concept of validity, many researchers view impact as one of the indispensable aspects of test validation, including but not limited to the impact aspect in Bachman and Palmer's (1996) framework of "test usefulness", the social consequences in Kunnan's (2004) "test fairness" framework; the consequential aspects of validity in Weir's (2005) "socio-cognitive framework" for test validation; and the warrants for intended consequences in Bachman's (2005) "assessment utilization argument".

Studies on washback effects of tests can, therefore, be subsumed under the category of impact studies. Since Alderson and Wall's "washback hypotheses" (1993), an increasing number of such empirical studies within various contexts and different types of tests have been carried out. These can be classified broadly into three types: key factors within the tests or testing systems; the stakeholders; and the contexts in which the tests are developed and used (Liu and Gu 2013).

The first set of factors is concerned with *tests themselves* and includes the test construct and test content; test formats; the use, function or purpose of a test, and especially the stakes attached to it (e.g. Alderson and Hamp-Lyons, 1996; Alderson and Wall, 1993; Andrews et al, 2002; Cheng, 2005; Green, 2007; Gu, 2007; Hawkey, 2009; Qi, 2004, 2007; Shih, 2007; Shohamy et al, 1996; Tang, 2005; Wall, 2005).

The second set of factors is concerned with stakeholder *groups* and *individuals* in educational settings, especially the teachers and learners, including their knowledge, beliefs, attitudes, and so on (e.g. Alderson and Hamp-Lyons, 1996; Andrews, 1995; Andrews et al, 2002; Green, 2007; Gu, 2007; Hawkey, 2006; Shih, 2007, 2009; Tang, 2005; Wall, 2005; Wall and Alderson, 1993; Watanabe, 1996, 2004).

The third set of factors relates to both the *micro-* and *macro-contexts*, as noted by Hawkey (op cit). These contexts range from multiple classroom settings at a local level (micro contexts), to the prevailing sociopolitical milieu in a society (macro contexts) (e.g. Hawkey, 2006; Hayes and Read, 2004; Green, 2007; Gu 2007; Saville, 2009; Tsagari, 2009; Wall, 2005; Watanabe, 2004).

The complexity of test impact phenomena has not only been demonstrated by the various factors mediating the *process of its realization,* but also by the *variability* it displays. Watanabe (2004), for example, has conceptualized this in relation to five dimensions: "specificity", "intensity", "length", "intentionality" and "value".

Studies on the working mechanisms of test impact/washback have also generated a number of "washback models", for example: Hughes' (1993, cited in Bailey 1996) trichotomy; Bailey's (1996) basic model of washback; Burrows' (2004) curriculum innovation model; Shih's (2007) washback model of students' learning; Greens' (2007) model of washback, incorporating intensity and direction; Saif's (2006) conceptual framework for washback; and Qi's (2004) basic model for the consequential aspect of validity. Liu and Gu (2013) suggest that these models originated from the basic question into the existence of washback (Does washback exist? – best embodied in Alderson and Wall's washback hypotheses), evolved into the question of how it works (the models of Bailey, Burrows and Shih), and more recently into how to achieve intended impact and washback effects (models proposed by Green, Qi and Saif). This later point accords with the concept of "impact by design" proposed by Saville (2009).

However, despite the considerable progress in the last two decades, many aspects of impact are still under-researched and remain to be addressed in a wide range of educational contexts, particularly the impact of international tests on test-takers in local contexts. In this chapter we report two small-scale studies conducted with two different groups of English learners in China which have received little attention so far:

1. The impact of international language tests on younger Chinese learners in compulsory education;

2. The impact of international language tests on Chinese learners seeking career development in work-related contexts.

In these two studies, the learners/test-takers we targeted ranged from school-age children starting at age nine at the CEFR A1 to B1 levels, to university undergraduates at CEFR B2/C1 levels. In both cases, the studies are briefly summarized and only a small part of the findings are discussed to illustrate the nature of the research and the insights obtained in carrying out this kind of impact-related research. Our focus is mainly limited to finding out more about the test-takers' perceptions, motivations and preparation processes.

14.3 Study 1: the impact of Key and Preliminary for Schools

The macro-context of Study 1 is in part set by the process of globalization, whereby English has become a world language, and the starting age for learning English is becoming increasingly younger worldwide (Graddol, 2013). This is the case in many parts of China where children as young as five are attending English classes out of school. Correspondingly, more and more young learners are being assessed using national and international English language proficiency tests. This part of the chapter reports on the impact of the Key and Preliminary for Schools on young Chinese learners in Beijing, through a combined methodology of questionnaires and semi-structured interviews.

14.3.1 Research methods: questionnaires and interviews

Two separate questionnaires were designed for learners of each of the exams but the questions were broadly the same. The main sources of information for the questionnaire design were:

1. The Handbooks for Teachers for each exam (Cambridge ESOL, 2011a and 2011b);
2. Consultation with test developers and validation researchers of these exams (personal communications);
3. Cambridge's Impact Study pamphlet and proposal templates (Cambridge ESOL, 2011c);
4. Previous Cambridge English washback and impact studies (e.g. Hawkey, 2006; Green, 2007; Saville, 2009, 2010);
5. The China team's washback and impact studies (e.g. Gu, 2007, 2011) and their local knowledge of the participants.

Altogether the questionnaires had four parts and 40 items. Part One sought participants' demographic information (nine items). Part Two covered their perceptions of the exams (20 items). Part Three investigated their EFL learning processes (five items). Part Four concerned their experiences in the relevant test preparation courses (15 items).

The questionnaires have five item types: blank-fill, multiple-choice with one answer only, multiple-choice with more than one answer acceptable, five-point Likert scale question type and open-ended questions. The questionnaires have both Chinese and English versions of the same content to facilitate the data collection and communication between the two teams. A two-month design process was iterative, with more than a dozen revised versions. The validation of the questionnaires was done mainly through the expert judgment of the Cambridge team and informal interviews, trials and pilots by the China team with the targeted participant groups in China. The semi-structured interviews were designed following the same iterative design and validation processes as the questionnaires. Interviews were conducted for two main purposes: to triangulate the questionnaire data and to explore in-depth information not revealed through the questionnaires.

14.3.2 Participants and data collection

The Key and Preliminary for Schools examinations were administered at test centres in Beijing in December 2011. The China team took this opportunity to administer the questionnaires and conduct interviews immediately after the exams. Altogether 592 test-takers were surveyed and among them 20 received individual interviews lasting two to five minutes. Both the questionnaire surveys and interviews were conducted in Chinese.

14.3.3 Findings and discussion

The findings are presented and discussed from the following perspectives: 1) the test-taker characteristics, 2) their perceptions of the exams, 3) their motivations for and anxieties about taking the exams, 4) their preparation for taking the exams, and 5) their latest exam results.

14.3.3.1 Test-taker characteristics

Test-taker characteristics are considered an important component in test validation (Weir, 2005) and are a fundamental starting point in impact-related studies beginning with demographic data. In the Key for Schools survey, girls account for 55% and boys for 46% of the total

number of test-takers surveyed. In the Preliminary for Schools survey, there was a bigger difference between the percentage of girls (62%) and boys (38%). Most of the children in the study had started learning English at about five years old and over 95% of them were studying at primary school (ages from nine to twelve) at the time of the investigation. Interestingly, the average age difference between the two groups was only one year: Key for Schools was ten and Preliminary for Schools was 11. In both cases this was younger than the targeted age groups for the two exams in the global cohort, that is, 11–14 years old (Papp and Nicholson, 2011). This was thought to be a factor that could impact on test performance because their cognitive ability might not be at the level required to deal with the questions effectively (see Gu and Saville, 2012 for a discussion of this issue).

In the Key for Schools group, 53% were in Grade Five and in the Preliminary for Schools group 67% were in Grade Six. This suggested that most children who had passed KET for Schools took PET for Schools just one year later (even though they may not have reached the required B1 proficiency level).

Interview data confirmed that the Key for Schools test-takers were more divergent in their language proficiency, whereas the Preliminary for Schools test-takers, having already passed Key for Schools, were more similar in their language proficiency.

For both groups, the majority of them came from Non-Key schools or Key schools at the district or county level that generally have lower level educational infrastructure and/or more unqualified teachers. During the interviews some reported that, if they wanted to enter better junior middle schools, they thought it was better for them to have obtained internationally-recognized certificates like the Cambridge one before entering those schools (again an aspect explored by Gu and Saville, 2012).

14.3.3.2 Perceptions of the exams

As the learners were younger than the targeted groups in the rest of the world cohort, some of them felt that they were not very familiar with the exams, even after taking them. Interestingly, despite the issue of unfamiliarity, both groups held positive perceptions of the exams, particularly the Preliminary for Schools test-takers (see Table 14.1). In a similar way, the Preliminary for Schools children, who are generally one year older than their Key for Schools counterparts, held more positive perceptions towards learning English generally, and had better feelings about the language than the younger test-takers. There were bigger individual differences among the Key for Schools children in their

Table 14.1 Perceptions of Key for Schools and Preliminary for Schools (%)

	KETfs	PETfs
The time allotment for answering each part of the exam is just as needed	64	91
The number of items in each part of the exam is appropriate	64	87
The difficulty of each part of the exam is reasonable	65	78
The exam can accurately assess my English language proficiency	66	89

perceptions and feelings (also confirmed by the parent data for the two groups reported in Gu and Saville, 2012).

14.3.3.3 Motivations for and anxieties about taking the exams

Each group had a range of motivations for taking the exams, but shared similar overall patterns. The main motivations were: to get a Cambridge English certificate; to ascertain their level of English proficiency; to improve their access to better educational opportunities in the future; and to stimulate their interest in learning English. The major difference was that the Preliminary group had higher percentages in these motivations and revealed other differences that can be explained by their age and greater maturity. For example, they showed higher percentages for the motivation item concerning living or travelling abroad in the future. The Key for Schools children on the other hand had a higher percentage related to fulfilling their parents' expectations and for getting better jobs (see Table 14.2).

On the whole, the learners' motivation appeared to be instrumental and goal-oriented, as previous empirical studies have demonstrated (e.g. Shohamy 1993).

More than half of the children in both groups reported that they were anxious about taking the exams, and for both groups, the Speaking test was the part that made them most anxious, possibly due to its face-to-face, one-to-one format and their relative unfamiliarity with this testing method. It is noteworthy that the older children tended to be worried about certain aspects of the exams only (the listening test and "difficult questions"), while the Key for Schools children tended to be anxious about all components (see Table 14.3).

These findings indicate that the sources of test anxiety might lie in two aspects of the test itself – the format and the difficulty – as well as the test-takers' perceived weaknesses in certain language skill(s). It was not clear whether or not the children's anxiety impacted negatively on their test performances as this was not linked in the analysis.

Table 14.2 Purposes to take Key for Schools or Preliminary for Schools (%)

	KETfs	PETfs
To ascertain the level of my English proficiency	47	51
To get a Cambridge English certificate	46	64
To improve access to better educational opportunities in the future	35	41
To stimulate my interest in learning English	35	38
To fulfill my parents' requirement	30	29
To be able to live or travel abroad in the future	27	31
To improve opportunities for a better job in the future	22	15
To experience an international test	16	22
To emulate other children	11	12
To fulfill my teacher's requirement	10	6
To fulfill my school's requirement	7	10
Other	4	4

Table 14.3 Anxiety of taking the exams (%)

		KETfs	PETfs
Were you anxious about taking the exam?	Yes	53	53
	No	47	47
Which part made you feel most anxious?	Listening	13	24
	Reading & Writing	24	24
	Speaking	46	38
	All three parts	17	14
Why did you feel anxious?	Time pressure	5	11
	Difficult questions	19	29
	Unfamiliar question formats	14	8
	Unfamiliar topics	21	17
	Anxious in every test	31	24
	Other	10	11

14.3.3.4 Preparation for the exams

Most, but not all of the children surveyed, had taken test preparation courses before taking the exams themselves. On average, the Key for Schools children attended such courses for half a year *longer* than the Preliminary for School children, but a higher percentage of the Preliminary ones took such courses. However, there were large individual differences within each group (see Table 14.4). This was not unexpected based on previous findings related to individuals and variability in test preparation behaviours (Gu, 2007; Shih, 2007; Watanabe, 1996). In addition, and not surprisingly, the older children

reported spending more time every day on English learning than the younger ones.

Both groups shared the same top three purposes in taking test preparation courses: to develop their English proficiency; to raise their exam scores; and to stimulate their interest in learning English. However, the Preliminary for Schools children reported a wider and more balanced range of purposes (see Table 14.5).

Interview data showed that most children were attending test preparation courses at privately-owned training institutions. There were three types of English courses: integrative courses; last-minute test preparation courses; and combined language skills and exam preparation courses. It was apparent that the quality of the training courses varied considerably (for example, from teacher to teacher within the same institution and from institution to institution – for a detailed discussion, please refer to Gu and Saville, 2012).

14.3.3.5 The latest exam results

Re-sitting a Cambridge examination in China seems to be a common occurrence and probably more prevalent than in other parts of the

Table 14.4 Test preparation

Test preparation	**Yes**	**No**	**Number**	**Mean (year)**	**Std**
KETfs	69%	31%	93	1.7	1.74
PETfs	76%	24%	228	1.2	1.26
Daily hour	*No time*	*Less than 0.5 hour*	*0.5–1 hour*	*1–1.5 hours*	*More than 1.5 hours*
KETfs	8	12	40	18	22
PETfs	5	12	33	28	22

Table 14.5 Purpose of taking test-preparation classes (%)

	KETfs	**PETfs**
Developing my English proficiency	67	63
Raising my exam score	64	68
Stimulating my interest in learning English	44	42
Fulfilling parents' requirement	27	47
Emulating other children	20	26
Fulfilling teacher's requirement	13	32
Fulfilling school's requirement	12	21
Other	1	26

world. Both the survey and interview data revealed that many children took the exams more than once. This may be influenced by a Chinese approach to test-taking and determined by pressures from the wider societal context. Out of the 248 Key for Schools children, 171 had taken the exam before; their average score was just below the cut-score for passing, which was 70 points on the standardized reporting scale and about 30% of them were quite close to the pass mark (within 15 points). Of the 344 Preliminary for Schools children, 248 had taken the exam before and their average score was also lower than the cut-score for passing.

In comparing the average score for each part between the Preliminary and Key for Schools children, there was a noticeable drop in the average for Part Two Listening (see Table 14.6). This supported the finding that Listening was the component that many Preliminary for Schools children found the most difficult and were worried about.

14.3.4 Summary of Study 1

The test-takers for the Key and Preliminary for Schools examinations in the study were younger than the targeted groups in the rest of the world. This finding in itself may be indicative of the prevailing test-taking culture in China being applied to English and the fact that many parents in China are encouraging their children to obtain a recognised English language qualification at an early age while still at school.

However, despite their younger age and their anxiety about taking a relatively unfamiliar examination, the learners themselves reported very positive perceptions towards the exams and they entered for them with a variety of motivations and purposes in mind. Given the fact that they are still in school, it is not surprising that integrated motivations

Table 14.6 Key/Preliminary for Schools – most recent exam results

		N	Min.	Max.	Mean	Std. Deviation
KETfs	Total	171	25	93	69	13.83
	part 1 R & W	171	8	47	33	8.42
	part2 L	171	5	25	18	4.39
	part3 S	171	9	24	18	3.03
PETfs	Total	284	19	91	63	11.55
	part1 R & W	284	11	47	31	6.19
	part2 L	284	0	24	14	4.30
	part3 S	284	4	25	18	2.86

Note: The total scores for both examinations are 100 points. For Part1 R & W, the total score is 50 points and for Part2 L and Part3 S, the total score is 25 points respectively.

related to self-improvement appear to be stronger than instrumental ones (e.g. future employment).

Some potentially negative impacts of the exams are noteworthy; these relate to test anxiety and the extra workload for test preparation which the children experience (Yan, Gu and Khalifa, 2014). Both areas need to be looked at in more detail to determine possible effects of these points on the wellbeing of the children.

In summary, the findings from this study indicate that the relationship between macro and micro contextual factors needs be examined in more detail in order to develop a better understanding of the contextual dynamics. For example, if there are strong influences from Chinese society that are likely to determine how learners go about taking tests, how can potentially negative impacts be anticipated and avoided when they enter for international exams, such as the Cambridge ones? If typical test-taking behaviour is based on different assumptions about assessment than those underpinning the international examination, how can the gap be mediated?

In addressing such questions, one way forward might be to provide better-targeted communication and support for stakeholders based on the findings of impact studies. This would entail test providers and their stakeholders working collaboratively to ensure that the test constructs and model of assessment are well understood by all concerned, and that the international testing system is appropriately adapted to meet the local needs (cf. Cambridge's Maxims of Communicate and Support-Milanovic and Saville, 1996).

14.4 Study 2: The impact of the business certificates vanage and higher

In the next part of the chapter, we describe Study 2 and look into the impact of the Business Certificates – Vantage (CEFR B2) and Higher (CEFR C1) – on university students in Chongqing. The researchers adopted a similar approach to the previous study but with a focus on older learners in a different geographical and educational context.

14.4.1 Research procedures, methods and data collection

This impact study was conducted by the China Team in Chongqing in December 2012 and they went through six preparatory steps over a period of two months leading up to that date in order to achieve a better understanding of the contextual features and to design/refine the research instruments. The steps were as follows. Step One: To

understand the test and its local uses better; Step Two: To attend the BEC Centre Exam Managers meetings and to get to know the local administrators and the network of local centres; Step Three: To attend one BEC teacher training in Chengdu, Sichuan Province to learn how they prepared learners for BEC; Step Four: To observe BEC Higher and BEC Vantage speaking examiners being trained and certificated and to conduct focus-group interviews with eight speaking examiners in Chongqing; Step Five: To conduct informal interviews and communication with ten BEC Vantage and Higher test-takers and potential test-takers at Chongqing University in November 2012. The questions were mostly open-ended, focusing on their perceptions, motivations and learning and test preparation processes; Step Six: To finalize the questionnaires and interviews to be used in the main study.

All the above steps contributed to the design and revision of the questionnaires and interviews for the Chinese Vantage and Higher test-takers. These questionnaires and interviews focused on exams as a whole, including the Reading, Writing, Listening and Speaking subtests. The test-takers questionnaires had three parts: Part One sought test-takers' demographic information (nine items); Part Two covered test-takers' comments on and perceptions of the tests (five items); Part Three was about their test-preparation processes (ten items).

The BEC test-taker questionnaires had five item types, as in the ones for Key and Preliminary for Schools. There were altogether 15 semi-structured interview questions for the test-takers. Most of them were open-ended questions, for example: *Why did you take BEC Vantage/Higher? What do you think of the timing, number of items and difficulty level of BEC Vantage/Higher? What was the biggest difficulty you have come across during your BEC Vantage/Higher preparation? In what ways have you benefited from your BEC Vantage/Higher preparation?*

The last step was the collection of the data from the test-takers using the adapted questionnaires and interview procedures. The Vantage and Higher test-takers' questionnaire data was collected from test-takers immediately after they had taken the written tests. Altogether 397 Vantage and 272 Higher test-takers were surveyed. From that group, 36 also took part in a five–to-ten minute interview in pairs. These interviews were conducted immediately after the speaking tests.

14.4.2 Findings and discussion

Again, the findings are presented and discussed in relation to the following aspects: 1) the test-taker characteristics, 2) their perceptions of

the exams, 3) their motivations for taking the exams, 4) their preparation for the exams, and 5) the influence of the exams on them as learners.

14.4.2.1 Test-taker characteristics

The two groups of test-takers shared similar biographical patterns. On average, they were 21 and 22 years old respectively and over 90% of them were currently studying at universities for a bachelor's degree. Not unexpectedly, a dominant proportion of them were females (Vantage 87% and Higher 80%). In addition to the widely accepted (though not necessarily correct) impression that females are better at language learning than males, the reality in China is that females generally have more difficulty than male peers in finding graduate-level jobs due to traditional biases or prejudice against female graduates. Females may, therefore, feel that they have to work harder to prepare themselves for the job market, and obtaining an internationally recognized certificate, such as BEC, to demonstrate their English language competence may be one way of doing so (Zhang and Yin, 2012).

The test-taker characteristics showed that in Chongqing where there is a population of 36 million and around 50 universities and colleges, the test-takers came from three types of university, college or department: (1) finance, economics and management, (2) international studies, and (3) education or teacher training ("Normal" universities). The test-takers in both groups came from a range of majors, but mainly Economics, English Literature and Management. Management and Economics are now popular due to China's economic development and the current social needs and so offer good prospects for employment. The test-takers' jobs or job intentions were similar as well, with international trade (over 60%), education (over 30%) and finance (about 30%) ranked as the top three. In these professions a functional level of English proficiency (B2 or above) is likely to be needed for international communication.

14.4.2.2 Perceptions of the exams

On the whole, the test-takers surveyed were not very familiar with the exams they were preparing for, particularly with the speaking and writing tests (which have a format that is less commonly used in Chinese language examinations). Only slightly over 20% of Vantage and 30% of Higher test-takers reported that they 'definitely' or 'to a large extent' knew the rating criteria of the speaking and writing tests. The BEC Higher test-takers were somewhat more familiar with the Writing and Speaking rating scales perhaps due to their previous test-taking experience(s) with Vantage.

According to the interviews, few students bothered to access the official websites of Cambridge English Language Assessment to learn about the test-taking procedures or the rating system. Occasionally, one or two reported that they did so, but they said they easily got lost due to the large amount of information available on the websites. Instead, most of them depended on other people around them, such as teachers or peers, for relevant information, even though the information itself might have been inaccurate or incomplete. This confirms the view that if intended impacts are to be achieved much more needs to be done to get the relevant information to the targeted stakeholders, even when information is readily available via websites and other media. This finding also concurs with the findings from the Key/Preliminary for Schools project.

As a whole, however, both groups of BEC test-takers held very positive perceptions of the exams (see Table 14.7).

Specifically, both groups made positive comments about all test papers for their appropriate timing, number of items and difficulty levels – with the exception of the difficulty of Listening in BEC Higher for which the test-takers were less sure or about the appropriateness of the difficulty level. There were also some differences between the two groups in their opinions about Reading (see Table 14.8).

During the interviews, the test-takers commented on these perceptions; for the listening they referred to the fast speech rate, unfamiliarity with the accents of the speakers and to listening to a loud speaker rather than through headphones; for the reading they commented on time pressure and the inclusion of too many new business-related words.

14.4.2.3 Motivations for taking the exams

The two groups shared the following motivations in taking the examinations: English language learning; job hunting; understanding one's

Table 14.7 Perceptions of BEC Vantage and higher (%)

Item	BEC Vantage	BEC Higher
Its content is consistent with real-life business English	82	76
It's a good indicator of one's business English proficiency	70	61
It's helpful in future education or employment	72	68
Its directions are clear	88	88
Its way of score reporting is appropriate	81	73

Table 14.8 Comments on each paper of BEC (%)

	Appropriateness	BEC Vantage	BEC Higher
Reading paper	The timing	74	57
	The number of items	89	59
	The difficulty level	74	58
Writing paper	The timing	88	79
	The number of items	90	87
	The difficulty level	86	78
Listening paper	The timing	81	74
	The number of items	85	80
	The difficulty level	58	33
Speaking paper	The timing	73	75
	The number of items	77	78
	The difficulty level	72	61

Table 14.9 Reasons for taking BEC Vantage and higher (%)

Item	BEC Vantage	BEC Higher
Facilitating English learning	68	64
Job hunting	43	52
Understanding one's own Business English proficiency	42	39
Obtaining an English certificate from a world-recognized authority	40	38
Meeting the need of current job	23	19
Experiencing a world-recognized test	17	17
College requirement/teachers' recommendation	14	6
Its appropriate difficulty level	8	7
Peer influence	6	6
Overseas study	4	5
No special reason	3	2
Living or travelling abroad	3	3
Employer's requirement	0	1
Other	0	1

own level of Business English proficiency; and obtaining an English certificate from a world-recognized authority. The main difference lay in the fact that more of the Vantage test-takers were taking the exam for "facilitating English learning" while the Higher test-takers were taking it specifically for "job-hunting" (see Table 14.9).

The BEC test-takers mainly got to learn about the exams from their classmates (about 70%), their teachers (BEC Vantage 51% and BEC

Higher 36%) or from the Internet/TV/Radio (about 20%). A small number also found out about the exams purely by chance. In one interview, a male Vantage test-taker who had worked for three years since university graduation talked about his story – how he got to learn about the test and why he made the decision to prepare and take the test himself.

One year ago, I went with my boss to a job market to recruit new employees. … That was the first time I heard about this test and got to know its value. So I went back, learned more about the test through the Internet, and started to prepare for it in my spare time. I thought if I got the certificate, it would be very helpful in my career promotion or for job-hopping.

In fact, less than half of their studies or jobs were "definitely" or "to a large extent relevant" to *business English* specifically (BEC Vantage 32% and BEC Higher 46%). In their interviews, the test-takers reported a range of other reasons for taking the exams, such as for: love for English; the effectiveness of having a "life-time certificate"; support for their own English study; making good of their spare time; following what others did.

14.4.2.4 Preparation for the exams

About half the test-takers in each group spent two to three months preparing for their exams. During that time, a majority spent 1.5 to 2.5 hours each day on test preparation (see Table 14.10).

There were both similarities and differences between the two groups in their test-preparation activities. From among all the activities, the top one in both groups was *practicing past exam papers* (BEC Vantage 74% and BEC Higher 82%), a typically "cramming" approach that is common amongst Chinese learners in general. However, BEC Vantage test-takers did more of the following activities: study using textbooks and coaching materials, memorizing business English words, practicing business English writing, memorizing sentence patterns, practicing

Table 14.10 Time spent on test preparation for BEC (%)

Total length	**1 month or less**	**2–3months**	**4–6 months**	**7–12 months**	**1 year or more**
BEC Vantage	42	**53**	4	1	1
BEC Higher	42	**51**	3	0	1
Daily hour	*0.5 hour or less*	*0.5–1 hour*	*1.5–2.5 hours*	*3–4 hours*	*4 hours or more*
BEC Vantage	13	31	**40**	11	5
BEC Higher	12	21	**30**	20	8

mock test papers and studying grammar. The BEC Higher test-takers did more in practicing spoken English and listening to business broadcasts in English.

In the interviews, the test-takers also listed the materials they used to prepare for the exams. In addition to the past test papers, these included: model tests, coaching materials, oral English books, textbooks or teaching materials, English speaking news, English talk shows, movies and TV series.

Only 10% of the test-takers reported that they had participated in a test-preparation course or online tutorial course. Generally at universities in China, students have College English courses for the first two years and maybe English for their own major courses in the third year. Before taking the Vantage and Higher, most test-takers had already taken some other national English exams, such as the National College English Test or the Test for English Majors. Most Vantage and Higher test-takers had therefore already reached a certain level of language proficiency before they took the exam, and thus they chose to study "business English" mainly through self-learning (over 55%) or some major courses related to Business English (over 45%).

In the interviews, they reported specific difficulties in test-preparation, such as difficulty in finding partners to practice spoken English, inadequate availability of past papers, poor quality of locally-produced model papers or test-preparation materials, lack of opportunity for listening practice, inadequate business English vocabulary, no guidance on how to prepare for the test, poor time management and unfamiliar task types.

When asked about what kinds of test-preparation strategies and test-taking strategies they used, the test-takers reported some interesting individual differences, such as: doing the test practice according to the test task order or following an inverted order starting from the last to the first; doing the practice without calculating the time; reading books on question-answer skills or test-taking strategies; doing excessive amounts of practice; practicing skills separately; checking the answers while answering questions. Some of these warrant further investigation to understand why/how they are used and what the effects are on learning or success in the exams.

14.4.2.5 Influence of the exams on the test-takers as learners

While both groups strongly agreed on the positive influences of the exams, particularly in increasing their knowledge of business English, the Vantage test-takers as a whole showed more interest in business

English learning, improving English proficiency and in "gaining a sense of achievement" (see Table 14.11).

In their interviews, when asked what they had achieved through preparation for the test, the test-takers gave the following answers, which largely suggested positive impacts:

1. The test preparation improved their English learning in a variety of ways, such as: cultivating their interests in learning English; enhancing their capacity of speaking English; improving their listening skills; enhancing their reading proficiency which is beneficial to postgraduates' entrance English exam; getting them familiar with business writing; increasing their vocabulary; improving their ability to gain key information that helps improve English learning ability; and getting them to know more about foreign cultures.
2. It is beneficial to their work in business environments, as they become more familiar with business English.
3. It helped them mentally, such as: improving their test-taking mentality; strengthening their self-discipline ability; improving their self-learning capacity and enhancing their ability to work under pressure.
4. It helped them make new friends.

   Some of these points, such as those listed under 3, are not widely observed in feedback on Cambridge examinations from other contexts outside of China, and again may be related to a Chinese approach to learning and test-taking which is not only restricted to language examinations but is common in the wider milieu.

14.4.3 Summary of Study 2

In summary, many of the same features that were commented on in Study 1 are observable in Study 2. We can note that most BEC test-takers think highly of the exams and see a range of important benefits for them in their personal, educational and professional lives. They

Table 14.11 Possible influences on BEC preparation (%)

Item	BEC Vantage	BEC Higher
Increasing business knowledge	83	81
More interested in business English learning	77	67
Improving English proficiency	82	76
Gaining a sense of achievement	66	53

express both intrinsic and extrinsic motivations in taking the exams, but the higher the proficiency level, the more relevant the extrinsic and instrumental purposes become (e.g. for employment reasons or mobility outside of their educational local context).

While they perceive the exams positively, they seem to be less familiar with some aspects of the tests than would normally be expected in similar test-taker cohorts in other parts of the world. For example, they have lower familiarity with the assessment scales for the writing and speaking tests (as observed for the younger learners in Study 1) and again this probably stems from the fact that productive skills are typically not assessed in China using the same communicative approach that the Cambridge examinations employ.

Not unexpectedly there was a marked "test-oriented" approach in their test preparation processes, probably based on the transfer of test preparation and test-taking behaviours from their previous experiences of taking high-stakes examinations – as discussed above. It is likely that some of these behaviours do not lead to effective language learning or successful outcomes in the Cambridge examinations and should be investigated in follow-up studies so that better guidance can be provided in future.

14.5 Conclusions

The two small-scale studies reported here focused on two different contexts in China where well-known international examinations were being used to support English language learning. Although the educational contexts and geographical regions were different, the focus in both cases was on the Chinese test-takers as language learners with a view to having a better understanding of their perceptions, motivations and behaviours in preparing for the Cambridge exams. What is clear from the two studies is that socio-cultural pressures from Chinese society (the macro context) impact on the perceptions and behaviours of all learners to some extent – whether children in school (who are strongly influenced by their teachers and parents) or young adults embarking on their careers. Global examination providers, therefore, need to gain better understandings of these local socio-cultural factors so that they can communicate more effectively with their Chinese stakeholders, and in so doing ensure that basic information related to their examinations is communicated effectively. This is a fundamental starting point for ensuring that the intended impact is achieved. In keeping with the impact research methodology suggested by Cambridge English (Saville,

2010), these limited case studies, therefore, need to be followed up with more research to investigate the useful insights gained so far. In other words, like other aspects of test validation, this needs to be an iterative and on-going process.

Acknowledgement

The first author would like to acknowledge the support she received for writing this chapter from the Fundamental Research Funds for the Central Universities in China (No. 0205005201030).

References

Alderson, J. and Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback, *Language Testing, 13*(3), 280–297.

Alderson, J. and Wall, D. (1993). Does washback exist?, *Applied Linguistics, 14*(2), 115–129.

Andrews, S. (1995). Washback or washout? The relationship between examination reform and curriculum innovation. In Nunan, D., Berry, V. and Berry, R. (eds.) *Bringing about change in language education* (pp. 67–81). Hong Kong: University of Hong Kong.

Andrews, S., Fullilove, J. and Wong, Y. (2002). Targeting washback: A case-study, *System, 30*(2), 207–223.

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

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

Bailey, K. (1996). Working for washback: A review of the washback concept in language testing, *Language Testing, 13*(3), 257–279.

Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. In Cheng, L., Watanabe, Y. and Curtis, A. (ed.) *Washback in language testing: Research contexts and methods* (pp. 113–128). New Jersey: Lawrence Erlbaum Associates.

Cambridge ESOL (2011a). Cambridge English: Key for Schools. Handbook for Teachers, http://www.teachers.cambridgeesol.org/ts/digitalAssets/117394_Cambridge_English_Key__KET_for_Schools_Handbook.pdf

Cambridge ESOL (2011b) Cambridge English: Preliminary for Schools. Handbook for Teachers, http://www.teachers.cambridgeesol.org/ts/digitalAssets/117385_Preliminary_for_Schools_Handbook.pdf

Cambridge ESOL (2011c) *Cambridge English making an impact*. Cambridge: Cambridge ESOL.

Cambridge ESOL (2012) *Cambridge English: Business Vantage Certificate Handbook for Teachers*, http://www.teachers.cambridgeesol.org/ts/digitalAssets/118035_Cambridge_English_Business__BEC__Handbook.pdf

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge: Cambridge University Press.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *A dictionary of language testing*. Cambridge: Cambridge University Press.

Graddol, D. (2013). *Profiling English in China – The Pearl River Delta*. Cambridge: Cambridge University Press.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge: Cambridge University Press.

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

Gu, X. (2011). *A longitudinal study of the CET washback*, research report to the National Foundation of Philosophy and Social Science of China.

Gu, X. and Saville, N. (2012). Impact of Cambridge English: Key for Schools and Preliminary for Schools – Parents' perspectives in China, *Research Notes,* 50, 48–56.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns, *Language Testing, 14*(3), 295–303.

Hawkey, R. (2006) *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.

Hawkey, R. (2009) A study of the Cambridge Proficiency in English (CPE) exam washback on textbooks in the context of Cambridge ESOL exam validation. In Taylor, L. and Weir, C. (ed.) *Language testing matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008*. (pp. 326–343). Cambridge: Cambridge University Press.

Hayes, B. and Read, J. (2004) IELTS test preparation on New Zealand: Preparing students for the IELTS Academic Module. In Cheng, L., Watanabe, Y. and Curtis, A. (eds.) *Washback in Language Testing: Research Contexts and Methods* (pp.97–111). New Jersey: Lawrence Erlbaum Associates.

Hughes, A. (1989) *Testing for language teachers*. Cambridge: Cambridge University Press.

Kunnan, A. (2004) Test fairness. In Milanovic, M. and Weir, C. (eds.) *European language testing in a global context* (pp. 27–28). Cambridge: Cambridge University Press.

Liu, X. and Gu, X (2013) A review of empirical washback studies worldwide over the past two decades, *Foreign Language Assessment and Teaching,* 1, 4–17.

Messick, S. (1989) Validity. In Linn, R. (ed.) *Educational measurement* (pp. 13–103). New York: American Council on Education and Macmillan.

Messick, S. (1996) Validity and washback in language testing, *Language Testing, 13*(3), 241–265.

Milanovic, M. and Saville, N. (1996) *Considering the impact of Cambridge EFL examinations*, internal report. Cambridge: Cambridge ESOL.

Papp, S. and Nicholson, G. (2011). Vocabulary acquisition in children and Cambridge ESOL's wordlist for tests for young learners aged 9–14, *Research Notes,* 46, 13–22.

Qi, L (2004). *The intended washback effect of the National Matriculation English Test in China: Intentions and reality*. Beijing: Foreign Language Teaching and Research Press.

Qi, L. (2007). Examining the intended and actual washback of the proofreading subtest in the National Matriculation English Test, *Curriculum, Teaching Material and Method, 27*(10), 43–46.

Saif, S. (2006) Aiming for positive washback: A case study of international teaching assistants, *Language Testing, 23*(1), 1–34.
Saville, N. (2009) *Developing a model for investigating the impact of language assessments within educational contexts by a public examination provider*, unpublished PhD thesis. Luton: University of Bedfordshire.
Saville, N. (2010) Developing a model for investigating the impact of language assessment, *Research Notes,* 42, 2–8.
Shih, C. (2007) A new washback model of students' learning, *The Canadian Modern Language Review/La Revue canadienne des languesvivantes, 64*(1), 135–162.
Shih, C. (2009). How tests change teaching: A model for reference, *English Teaching: Practice and Critique, 8*(2), 188–206.
Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning, *The Modern Language Journal, 76*(4), 513–521.
Shohamy, E. (1993). The power of tests: The impact of language tests on teaching and learning. *NFLC Occasional Papers*. Washington, DC: The National Foreign Language Center.
Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests*. London: Pearson Education.
Shohamy, E., Donitsa-Schmidt, S. and Ferman, I. (1996). Test impact revisited: Washback effect over time, *Language Testing, 13*(3), 298–317.
Tang, X. (2005). A study of the backwash effect of language testing, *Foreign Languages and Their Teaching, 7*, 55–59.
Tsagari, D. (2009). Revisiting the concept of test washback: Investigating FCE in Greek language schools, *Research Notes,* 35, 5–10.
Wall, D. (1997). Impact and washback in language testing. In Clapham, C. and Corson, D. (ed.) *Language testing and assessment* (pp.291–302). Amsterdam: Kluwer Academic Publishers.
Wall, D. (2005). *The Impact of high-stakes examination on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge: Cambridge University Press.
Wall, D. and Alderson, J. (1993). Examining washback: The Sri Lankan impact study, *Language Testing, 10*(1), 41–69.
Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research, *Language Testing, 13*(3), 318–333.
Watanabe, Y. (2004). Teacher factors mediating washback. In Cheng, L., Watanabe, Y. and Curtis, A. (eds.) *Washback in language testing: Research contexts and methods* (pp. 129–146). New Jersey: Lawrence Erlbaum Associates.
Weir, C. (2005). *Language testing and validation: An evidence-based approach*. London: Palgrave Macmillan.
Yan, Q., Gu, X. and Khalifa, H. (2014). Impact of Cambridge English: Key for Schools on young learners' English learning: Voices from students and parents in Beijing, China, *Research Notes,* 58, 44–50.
Zhang, K. and Yin, S. (2012). Theory and evidence of female graduates' job search: A study based on data from 63 universities, *Chinese Journal of Population Science, 32*(1), 94–101, 112.

Index