Advancing the Field of Language Assessment

Papers from TIRF doctoral dissertation grantees

For a complete list of titles please visit: www.cambridge.org/elt/silt

Also in this series:

Changing Language Teaching through Language Testing: A washback study
Liying Cheng

The Impact of High-stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory
Dianne Wall

Impact Theory and Practice: Studies of the IELTS test and *Progetto Lingue 2000*
Roger Hawkey

IELTS Washback in Context: Preparation for academic writing in higher education
Anthony Green

Examining Writing: Research and practice in assessing second language writing
Stuart D. Shaw and Cyril J. Weir

Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference, May 2005
Edited by Lynda Taylor and Cyril J. Weir

Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams
Roger Hawkey

Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008
Edited by Lynda Taylor and Cyril J. Weir

Components of L2 Reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners
Toshihiko Shiotsu

Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual
Edited by Waldemar Martyniuk

Examining Reading: Research and practice in assessing second language reading
Hanan Khalifa and Cyril J. Weir

Examining Speaking: Research and practice in assessing second language speaking
Edited by Lynda Taylor

IELTS Collected Papers 2: Research in reading and listening assessment
Edited by Lynda Taylor and Cyril J. Weir

Examining Listening: Research and practice in assessing second language listening
Edited by Ardeshir Geranpayeh and Lynda Taylor

Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference, July 2011
Edited by Evelina D. Galaczi and Cyril J. Weir

Measured Constructs: A history of Cambridge English language examinations 1913–2012
Cyril J. Weir, Ivana Vidaković, Evelina D. Galaczi

Cambridge English Exams – The First Hundred Years: A history of English language assessment from the University of Cambridge 1913–2013
Roger Hawkey and Michael Milanovic

Testing Reading Through Summary: Investigating summary completion tasks for assessing reading comprehension ability
Lynda Taylor

Multilingual Frameworks: The construction and use of multilingual proficiency frameworks
Neil Jones

Validating Second Language Reading Examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference
Rachel Yi-fen Wu

Assessing Language Teachers' Professional Skills and Knowledge
Edited by Rosemary Wilson and Monica Poulter

Second Language Assessment and Mixed Methods Research
Edited by Aleidine J Moeller, John W Creswell and Nick Saville

Language Assessment for Multilingualism: Proceedings of the ALTE Paris Conference, April 2014
Edited by Coreen Docherty and Fiona Barker

Learning Oriented Assessment: A systemic approach
Neil Jones and Nick Saville

Advancing the Field of Language Assessment

Papers from TIRF doctoral dissertation grantees

Edited by

MaryAnn Christison
University of Utah

and

Nick Saville
Cambridge English Language Assessment

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781316634486

First published 2016

Printed in XXXX

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Christison, MaryAnn, author. | Saville, Nick.
Title: Advancing the field of language assessment : papers from TIRF doctoral
dissertation grantees / edited by MaryAnn Christison and Nick Saville.
Description: Cambridge, United Kingdom : Cambridge University Press, [2016] |
Series: Studies in language testing; 46 | Includes bibliographical
references and index.
Identifiers: LCCN 2016028993 | ISBN 9781316634486 (alk. paper)
Subjects: LCSH: English language--Ability testing--Evaluation. | English
language--Study and teaching--Research. | English language--Study and
teaching--Foreign speakers--Evaluation. | Language acquisition--Ability
testing--Evaluation. | Language and languages--Ability testing--Research.
| Language and languages--Study and teaching--Research.
Classification: LCC P118.7 .A38 2016 | DDC 418.0071--dc23 LC record available at https://lccn.loc.gov/2016028993

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate. Information regarding prices, travel timetables, and other factual information given in this work is correct at the time of first printing but Cambridge University Press does not guarantee the accuracy of such information thereafter.

Contents

Acknowledgements

As editors we would like to take this opportunity to thank a number of individuals who have contributed to this volume in the Studies in Language Testing (SiLT) series. Our sincerest appreciation goes to The International Research Foundation for English Language Education (TIRF) Doctoral Dissertation Grant (DDG) awardees, who are the authors of the individual chapters (see 'Notes on contributors'). Because their research has taken place in many different contexts, they have helped us craft a volume that provides global perspectives on issues on language testing.

We are pleased to be co-editors of the first publishing effort between TIRF and Cambridge English Language Assessment featuring TIRF Doctoral Dissertation Grants and would like to acknowledge Mike Milanovic for planting the seed for the SiLT project before the conclusion of his final term of service on the TIRF Board of Trustees and his retirement from Cambridge English Language Assessment. Thanks also go to Cyril J Weir, joint Series Editor of SiLT with Nick Saville, for his editorial feedback and advice and for the timely manner in which he made his comments available.

We also want to recognize the editing and production team at Cambridge English Language Assessment and Cambridge University Press, especially John Savage and Evelina Galaczi, for their attention to every detail related to the publishing process. It has been a pleasure working with you. A special thank you also goes to Kathleen M Bailey, TIRF President, who has written the epilogue for this volume.

Series Editors' note

This volume of Studies in Language Testing features 11 papers based on doctoral dissertations funded by TIRF – The International Research Foundation for English Language Education (www.tirfonline.org). TIRF was established in 1998 by a vote of the TESOL Board of Directors and is now an independent foundation that raises and distributes funds to promote research on key issues in language education.

In 2001, TIRF determined the foundation's main research priorities, one of which dealt with language assessment and the role that language assessment plays in social or educational contexts. The following topic areas were identified as areas of relevance: measurement issues related to new techniques or the innovative use of existing techniques; development of testing materials; validation of testing tools; investigation of test impact; use of technology in the administration of language assessment procedures; and reliability and validity of regional or local language assessment procedures.

To date TIRF has supported 75 doctoral dissertation grantees from 20 different countries, many with financial support from Cambridge English Language Assessment. Eighteen of these studies have been related to language assessment, and the chapters in this volume have been chosen from that body of work. As explained in the Preface, the Editors have organized the volume into three major categories: assessing the productive skills in post-secondary contexts; assessing young learners in school contexts, and test takers' perceptions of tests in local contexts.

This collection of shorter papers based on doctoral theses is something of an innovation for the Studies in Language Testing (SiLT) series, but it is very much in keeping with the editorial stance that has been taken since its inception in 1995. One of the core purposes of the series is to enable the language testing community to benefit from research that makes a significant contribution to the field, but that otherwise might not reach publication. Rigorous criteria are used to select PhDs for inclusion in the series, and these have also been applied to the papers in this volume. The criteria include the following:

- making a contribution to knowledge
- having a sound theoretical basis
- being well referenced to the literature
- being research based
- being executed with care and thoroughness

- demonstrating analysis and interpretation that is well founded; and
- having the style of an academic monograph.

In the series to date, 11 full PhD theses have been published in the SiLT series. The first was by Antony Kunnan on test taker characteristics and test performance (SiLT 2) and the next by James E Purpura on learner strategy use and performance (SiLT 8). Caroline Clapham documented the development of *IELTS (International English Language Testing System)* and looked in particular at the effect of background knowledge on reading comprehension (SiLT 4), while Kieran O'Loughlin compared direct and semi-direct tests of speaking (SiLT 13) and Angela Hasselgreen looked at testing the spoken English of young Norwegians (SiLT 20). Dianne Wall and Liying Cheng both investigated aspects of test washback and impact, with Cheng carrying out a study on the classroom in Hong Kong (SiLT 21) and Wall studying its effects on the classroom in Sri Lanka (SiLT 22). Tony Green investigated the impact of the *IELTS* Writing subtest on English for Academic Purposes pedagogy (SiLT 25). Toshihiko Shiotsu examined the component of L2 reading ability in the context of Japanese learners of English (SiLT 32) and Lynda Taylor investigated how far testing reading through summary tasks enabled us to get closer to measuring the underlying construct of reading ability more faithfully and comprehensively (SiLT 39). Most recently we published a revised version of Rachel Wu's PhD thesis on *Validating Second Language Reading Examinations* (SiLT 41).

Through these publications we have enabled high-quality doctoral research to reach a wider audience than would normally be expected. The publication of this volume containing 11 short synopses of research supported by the TIRF program is a continuation of that tradition. By including synopses of well-conceived and systematically executed PhD research we hope to share with all language testers (experienced or emerging), a wider than usual selection of state-of-the-art research on diverse language testing issues across a broad spectrum of contexts.

Methodologically the papers have much to offer readers, particularly to those who are embarking on PhDs or research projects themselves. The volume provides a useful point of reference for nascent researchers to engage with at an early stage of their journey. The collective value of these papers to the different parts of this process is something we would wish to emphasize.

We are grateful to MaryAnn Christison for her expertise and unstinting support in bringing this volume to fruition and to Kathleen M Bailey (Chair of the TIRF Board of Trustees and President of the foundation) for her insightful contribution in the epilogue.

Nick Saville
Cyril J Weir
January 2016

Preface

We are pleased to present Volume 46 of the Studies in Language Testing (SiLT) series, which features research on language assessment from the recipients of The International Research Foundation for English Language Education (TIRF) Doctoral Dissertation Grants (DDGs). The volume is the result of a partnership that involves TIRF, Cambridge English Language Assessment, and Cambridge University Press. A total of 11 researchers who were DDG awardees contributed to this volume. In addition, Kathleen M Bailey, who is President of TIRF, wrote the epilogue chapter, and the two co-editors, MaryAnn Christison and Nick Saville, are members of the TIRF Board of Trustees. Nick Saville is also Director of Research and Thought Leadership at Cambridge English Language Assessment.

As co-editors of this volume, we enthusiastically support this project and the TIRF/Cambridge English Language Assessment partnership for a number of reasons. First, we both have a close association with TIRF as members of the Board of Trustees. Second, we believe in and wish to support TIRF's mission, which is 1) to implement a research and development program that will generate new knowledge and inform and improve the quality of English language teaching and learning; 2) to promote the application of research on practical language problems; 3) to collect, organize, and disseminate information and research on the teaching and learning of English; and 4) to influence the formation and implementation of appropriate language education policies, recognizing the importance of local/transnational languages and cultures worldwide, and of English as an international language. We believe that by publishing SiLT Volume 46, all four of the points in TIRF's mission statement are covered. Finally, Cambridge English Language Assessment is a leader in language assessment worldwide, placing quality and fairness at the center of the process of assessment and test development. Over 20,000 universities, employers, and governments worldwide have accepted and use tests produced by Cambridge English Language Assessment, and Cambridge English Language Assessment has partnered with TIRF in granting DDGs since 2008. Cambridge University Press, joint publisher of this volume with Cambridge English Language Assessment, is a well-established and well-respected publishing house that is committed to disseminating and promoting quality research; consequently, it is a pleasure for TIRF to collaborate with them for the first time on this project.

Audience for the book

The chapters in this volume have been chosen to appeal to a wide audience, including emerging and established scholars in the field of language assessment. To this end, we have designed the contents of this book to include research on current issues in language assessment from different contexts, and research that presents different perspectives on research design and on methods for collecting and analyzing data.

Purpose of the book

All of the chapters in this SiLT volume focus on the applied nature of the research on language assessment, and each chapter includes a section that focuses on implications. We believe that the volume can serve as a valuable text for graduate seminars in English language assessment in applied linguistics, education, Teaching English to Speakers of Other Languages (TESOL) and Teaching English as a Foreign Language (TEFL). Because the volume presents a global perspective on the research on language assessment, the individual chapters are also useful as supplemental readings for courses on second language (L2) methodology and curriculum design, teacher development in English Language Teaching (ELT), as well as courses on language assessment. As a reference volume, it is appropriate for individual scholars, test developers, graduate and undergraduate students, and researchers.

Organization of the book

Before we introduce the organization of this volume of SiLT and discuss the individual chapters, it is important for us to explain a little about the DDGs and how the chapters were selected. A major vehicle for TIRF in carrying out its mission has been a set of grant-funding initiatives on specific research priorities, including the DDGs (details are available on TIRF's website, www.tirfonline.org). DDGs are for doctoral students who have advanced to candidacy (i.e. they have completed such activities as coursework, exams, developed their research proposal, and are working on their dissertation research). TIRF sponsors an annual DDG competition that funds applications that are ranked the highest by external reviewers and have the potential of being completed within a year.

It would have been impossible for us to include chapters from all DDG awardees who have been funded under the TIRF Research Priority on Language Assessment since its inception in 2008. As editors, our job has been to craft a volume that is balanced and includes different views on language assessment. We were particularly interested in including research that addressed the importance of contextual factors. To this end, each section focuses on a different context: post-secondary, school, and local contexts. In addition, we were interested in looking at how these contexts intersected with specific topics at the forefront of English language teaching, learning, and assessment, namely, the assessment of productive language skills (i.e. oral language development and writing), the assessment of young learners, and the importance of including test taker perceptions in the assessment process. To further unify the volume, we include an epilogue.

Creating an edited volume that is unified and coherent is always something of a challenge; consequently, we asked the 11 chapter authors to follow a uniform format in their organizational structure. Although there is some variation across chapters, we used similar headings when and where appropriate, so it is likely that you will find the following headings in the chapters to guide you: 1) motivation for the research, 2) data collection, 3) data analysis, 4) results, 5) discussion, and 6) implications. The implications section of the chapters has been conceived of quite broadly to include implications for policy, practice, and/or future research. By creating a similar organizational structure for these chapters, we hoped to promote consistency and coherence for you as the reader, while at the same time allowing authors the freedom to report on the unique aspects of their research.

This volume of SiLT contains a preface, 11 chapters that focus on empirical research, and an epilogue. The volume begins with a note by the Series Editors, Nick Saville and Cyril J Weir. The purpose of this note is to provide you with some background on SiLT and on how TIRF entered into its partnership with Cambridge English Language Assessment.

There are four different sections to this volume. The 11 chapters are divided into three sections that are separated by the different roles that context plays in creating language assessments that are responsive to the needs of learners. To this end, Section 1 contains five chapters that are devoted to assessing the productive skills in post-secondary contexts. Section 2 consists of three chapters that address issues related to assessing young learners in school contexts, while the three chapters in Section 3 explore language assessment concerns in local contexts. Section 4 contains the epilogue chapter.

Section 1 presents research on assessing English as a Second Language (ESL) and EFL learners' productive skills in post-secondary contexts and includes chapters that focus on topics of concern to a number of different stakeholders, such as the assessment of L2 writing, international teaching assistants' speech, and the use of automated essay scoring. It begins with Chapter 1 by Kristen di Gennaro, which is entitled 'Comparing international and US resident second language learners' performances in five domains of writing'. In her chapter, di Gennaro reports on a study designed to identify differences across two groups of L2 writers – resident and non-resident – in the five domains of the writing construct: grammatical, cohesive, rhetorical, sociopragmatic, and content control. Findings from her study complement and challenge previous research on this topic, as they reveal similar performances across the two groups in several domains. Chapter 2 is written by Cecilia Guanfang Zhao and is entitled 'The role of voice in L2 argumentative writing: The development and validation of an analytic rubric'. In this chapter the author formally investigates whether the strength of author voice in written texts can be reliably measured and, if so, how it can be done. Using a mixed methods approach, Zhao develops and validates an analytic rubric that measures voice strength in L2 writers' argumentative essays. Semire Dikli's Chapter 3, 'Use of an automated essay scoring system in a multi-draft ESL writing class', investigated the extent to which ESL writers used feedback from automated essay scoring (AES) versus teacher essay scoring across five traits of writing. Chapter 4, 'Effects of pragmatic task features on temporal measures of Chinese ESL and EFL spoken request production', is written by Lixia Cheng. Cheng's research looks at Chinese English learners in two different post-secondary contexts and examines the effects of pragmatic task features through a composite of power, distance, and rank of imposition (PDR) on spoken English performance of requests. In the final chapter in this section we turn our attention to assessing the speech of international teaching assistants (ITAs). Chapter 5 is written by Ching-Ni Hsieh and is entitled 'ESL teachers' versus American undergraduates' judgments of international teaching assistants' accentedness, comprehensibility, and oral proficiency'. Hsieh investigates how two groups of raters – ESL teachers who are trained raters and American undergraduates who are untrained raters – evaluate the oral proficiency, accentedness, and comprehensibility of the speech of ITAs.

In Section 2 the focus of the chapters changes from post-secondary contexts to school contexts, and from adults to young learners. Beth Clark-Gareca investigates classroom content tests in Chapter 6, 'Elementary English language learners and classroom content tests'. She examines how 50 English language learners (ELLs) in 10 mainstream Grade 4 classrooms perceive the role of language, testing, accommodations, and grading practices in science and math classrooms. In Chapter 7, 'Exploring relationships between multi-word vocabulary, transparency, and literacy development', Sara A Smith examines the role that multi-word vocabulary knowledge plays in the oral language and reading development in children. The final chapter in this section is Chapter 8, 'Teacher perspectives on social language assessment' by Kimberly K Woo. The author focuses on the dichotomy of social and academic language and calls into question traditional assumptions about what comprises social language and its influence on the development of academic language proficiency. In an educational landscape wherein greater priority is often placed on the use of language for academic rather than social purposes, Woo's research is especially important for test developers.

In Section 3 our attention shifts to local contexts and to test takers' perceptions of tests. In Chapter 9, ‘The implications of test taker perceptions for test validity in community college settings’, Tasha Darbes investigates assessment and placement practices with immigrant populations in US community college settings. In her study, Darbes examines the psychological and social impacts of assessment through the analysis of students’ causal thoughts relative to their performance on placement exams. In Chapter 10, ‘Washback and the reformed CET-4: Insights from students’, Zhiling Wu examines washback from the National College English Test Band 4 (CET-4) in three local university contexts in China. These university contexts represent three levels of prestige in the Chinese hierarchy of higher education. The study used students’ perceptions to determine the overall effect of the CET-4 at each university, as well as its effect on classroom instruction. The final chapter in Section 3 is Chapter 11 by Nick Zhiwei Bi, ‘The impact of strategic processing on lexico-grammar test performance’. In this chapter, Zhiwei Bi investigates Chinese EFL lexico-grammatical ability in a local context in China using a large-scale longitudinal study. The focus of Zhiwei Bi’s study is on helping test developers understand more about the ways in which strategic processing is related to L2 test performances.

The volume ends with a short epilogue by Kathleen M Bailey. In this epilogue, Bailey both summarizes and synthesizes the information provided in these chapters and comments on what she sees as the major contributions of the volume and future directions for research in language assessment.

Conclusion

From its inception, TIRF has supported different types of research and researchers who work in many different contexts. In this way, TIRF and Cambridge English Language Assessment are a natural partnership. You will note that in this SiLT volume, the chapter authors have presented research that is consistent with this broad orientation to research. Some chapter authors have used qualitative research methods, such as interviews and observations. Other chapters use questionnaires, standardized tests, and other specific tasks to generate quantitative data. A number of other chapters have used mixed methods in the design of their studies and used both qualitative and quantitative data to answer their research questions. We see the diversity of research practices in the chapters in this volume as consistent with TIRF’s orientation to research, the research funded by Cambridge English Language Assessment, and the research published by Cambridge University Press.

As the editors, we wish to express our gratitude to chapter authors. Because the chapter authors were DDG awardees, their chapters were based on their dissertation research; consequently, they had to undertake the arduous task of selecting specific data from their dissertations and reworking them for the chapter-length contributions. In addition, they had to craft a chapter that would fit the intended reading audience for this volume, which is very broad. As editors, we recognize that this task is difficult, so we want to acknowledge and congratulate these young scholars on their work.

As editors, we are pleased to have participated in the creation of the first SiLT volume on language assessment, resulting from the TIRF/Cambridge English Language Assessment partnership. It has been enjoyable to work with the authors of the chapters and interesting for us, as editors, to work together for the first time on a project. We are pleased to further the mission of TIRF, the TIRF/Cambridge English Language Assessment partnership, and research on English language assessment.

MaryAnn Christison
Nick Saville
January 2016

Notes on contributors

Kathleen M Bailey works in the TESOL/TFL Program at the Middlebury Institute of International Studies in Monterey (MIIS). Her research interests include teacher development and supervision, language assessment, and classroom research. She is the President of The International Research Foundation for English Language Education (TIRF) and the Vice President of the American Association of Applied Linguistics (AAAL).

Nick Zhiwei Bi is an associate professor of Applied Linguistics in the Department of English at the University of Shanghai for Science and Technology. His PhD dissertation on test takers' strategic competence in the lexico-grammar test performance was awarded an International Research Foundation for English Language Education (TIRF) Grant in 2013. His research interests focus on grammar and pragmatics assessment, self-regulated learning, and L2 learners' strategic processing in language use and test taking.

Lixia Cheng is a graduate of the PhD Second Language Studies program at Purdue University, Indiana. She has considerable experience in language testing, curriculum development, and teaching introductory linguistics and English as a Second Language (ESL) courses. She is currently Assistant Director of Testing for a new English for Academic Purposes (EAP) program: Purdue Language and Cultural Exchange (PLaCE).

MaryAnn Christison is a professor in the Department of Linguistics at the University of Utah. She is a member of The International Research Foundation for English Language Education (TIRF) Board of Trustees and was President of Teachers of English to Speakers of Other Languages (TESOL) International Association 1997–1998. Her research interests include teacher education, language and the brain, and classroom-based assessments.

Beth Clark-Gareca PhD, is currently a lecturer in the Applied Linguistics and Teachers of English to Speakers of Other Languages (TESOL) programs at State University of New York. Her work focuses on the preparation of pre-service teachers pursuing certification in English as a New Language (ENL). Her research interests include content-based assessment in classroom contexts, second language acquisition, and K-12 teacher education.

Tasha Darbes received her PhD in Teaching English to Speakers of Other Languages (TESOL) from New York University. She is currently an assistant professor at Pace University, New York. Her research interests are in how immigrant-origin students experience testing and placement practices and how these practices intersect with bilingual proficiencies and identities.

Kristen di Gennaro is Assistant Professor and Director of Composition at Pace University, New York City. Her research interests include writing assessment and pedagogy, with a focus on second language writers. Her work has appeared in various journals including *Assessing Writing*, *Language Testing*, *Journal of Basic Writing* and *Writing & Pedagogy*.

Semire Dikli is an Associate Professor of English for Academic Purposes (EAP) at Georgia Gwinnett College (GGC). Dr Dikli has taught courses related to English as a Second/Foreign Language (ESL/EFL) in numerous locations for several years. Her research interests include ESL/EFL writing and Automated Essay Scoring (AES).

Ching-Ni Hsieh is Research Project Manager II at the Center for English Language Learning and Assessment at the Educational Testing Service, New Jersey. Her research has focused on speaking assessment, test development and validation, and the assessment of young language learners. Ching-Ni holds a PhD in Second Language Studies from Michigan State University.

Nick Saville is Director of Research and Thought Leadership at Cambridge English Language Assessment. He is on the Board of Trustees for The International Research Foundation for English Language Education (TIRF) and is also a Board member of Cambridge University's Institute for Automated Language Teaching and Assessment (ALTA). Before joining Cambridge English he taught at the University of Cagliari (Italy) and worked in Japan. He is currently joint editor of the Studies in Language Testing (SiLT) series.

Sara A Smith received her PhD in Applied Linguistics from the University of Oxford, followed by postdoctoral research at Harvard Graduate School of Education, Massachusetts. In 2014, she joined California State University East Bay as an assistant professor in the Department of Human Development and Women's Studies. Her research interests include the lifelong cognitive implications of bilingualism and assessing young learners.

Kimberly K Woo received her doctorate in TESOL from the Multilingual Multicultural Studies program in the Steinhardt School of Culture, Education, and Human Development at New York University. She previously earned an MA in Childhood Education at New York University, and taught in New York City public schools. She is currently at Teachers' College, Columbia University, New York City.

Zhiling Wu received her PhD from the Composition and Teachers of English to Speakers of Other Languages (TESOL) program at Indiana University of Pennsylvania, and is a recipient of an International Research Foundation for English Language Education (TIRF) Doctoral Dissertation Grant. Currently she is teaching adult English as a Second Language (ESL) students in Houston, Texas. Her research interests include washback effects, learner autonomy, second language testing, and second language acquisition.

Cecilia Guanfang Zhao is currently an associate professor in the School of English Studies at Shanghai International Studies University, China. Her research interests are in the areas of second language assessment, writing, and Teaching English to Speakers of Other Languages (TESOL®). Her dissertation research, which is reported on in this volume, won a 2012 Jacqueline Ross Test of English as a Foreign Language (TOEFL) Dissertation Award, a 2011 Christopher Brumfit PhD/EdD Thesis Award, and a 2009 International Research Foundation for English Language Education (TIRF) Doctoral Dissertation Grant.

Section 1
Assessing the productive skills in post-secondary contexts

1 Comparing international and US resident second language learners' performances in five domains of writing

Kristen di Gennaro

Pace University, New York City, US

Motivation for the research

Higher education programs in the US typically require students to complete courses in writing or composition, that is, courses in which students focus on writing for university-level coursework. To accommodate the large numbers of students in these classes whose first language (L1) is not English, several institutions offer composition courses specifically for second language (L2) writers. Depending on the institution's placement procedures, L2 students enrolling in such courses may be international students who have recently arrived in the US with student visas after having completed secondary education in their home countries, or they may be long-term residents of the US who have completed secondary (or even primary) education in US schools. This group is often referred to as Generation 1.5. Most institutions of higher education enroll students in both categories. Recent statistics on international student enrollment indicate a dramatic rise in the number of international L2 students in US colleges and universities, for example the 'Open Doors Data' from the Institute of International Education (2014). Statistics on school-age students reporting a primary language other than English indicate a rise in US resident L2 learners as well (National Clearinghouse for English Language Acquisition 2008). Based on these demographics, L2 composition courses appear to serve an increasingly heterogeneous student body (di Gennaro 2012). Whether such differences warrant the creation of separate courses for different types of L2 learners is an ongoing debate within the community of L2 composition scholars (di Gennaro 2012, 2013, Doolan 2013, 2014, Matsuda 2008).

Literature review

International L2 learners, that is, students who have arrived in the US after having completed secondary education in their home countries, are likely to have acquired much of their knowledge of English in classroom environments. Traditionally, such instruction has been described as favoring written English and formal grammar instruction over spoken English and conversational fluency. International learners' completion of secondary school in their home countries assumes advanced literacy in their L1s which, some scholars suggest, may lead them to transfer grammatical rules and organizational preferences from their L1s into their written academic English (see Reid 2006, Thonus 2003). Conversely, US residents who are L2 learners most often attend US post-secondary institutions after having completed secondary or even primary education in the US, thereby, having acquired English primarily through immersion in English-speaking environments, including academic classroom experiences. These experiences result in conversational fluency, as well as familiarity with process approaches to teaching writing. Nevertheless, they may lack an awareness of the differences between informal and academic registers (di Gennaro 2008, 2009). Given differences in prior instruction and exposure to English, L2 writing scholars (see Ferris 2009, Reid 2006, Roberge, Siegal and Harklau 2009) claim that international and resident L2 learners will have noticeably different strengths and weaknesses in their writing abilities, and thus require different types of instruction to improve their writing (see Mikesell 2007).

Until very recently, empirical support for such claims was scant and limited to small-scale studies. For example, in a detailed case study, Leki (1999) noted that the US resident L2 student who was the focus of her qualitative study excelled at informal communication in English, yet did not do well in courses where the focus was on using grammar skills in writing. Frodesen and Starna (1999), who examined multiple writing samples and conducted interviews with two students over the course of several years, suggested that their participants' errors reflected their different backgrounds, and recommended different courses of action for each student. Specifically, they recommended that the long-term US resident L2 learner in their study could benefit more from a mainstream (i.e. L1) composition course than a course directed at L2 learners. The other student, who was a recently arrived L2 learner, would likely prefer a composition course created specifically for L2 learners. Bosher (1998), whose data included interviews, stimulated recalls and text analyses of three students, found that the international L2 participant in her study attended more to content and organization in writing than to other aspects of writing proficiency; conversely, the representative resident L2 participant attended more to surface-level language issues and generating text, and did not appear to focus much on content, discourse, or the overall purpose of the text. A second resident L2 learner demonstrated writing processes more similar to the international participant than to the other resident learner. While these studies are limited in their sample sizes, their findings highlight that L2 students placed in similar composition courses may have very different strengths and weaknesses in completing academic writing tasks.

Moving beyond small-scale case studies, Bosher and Rowekamp (1998) examined a series of factors to determine which of them best predicted resident L2 writers' success compared to those of international learners in post-secondary education. The study included 56 participants divided into two groups based on whether they had completed secondary school in the US or in their home countries. Participants who had completed secondary school in their home countries scored significantly higher on the objective section of the Michigan English Language Assessment Battery, while the US secondary school graduates scored significantly higher on the listening section. No significant differences appeared across the two groups in terms of their composition scores. In a similar study, Muchinsky and Tangren (1999) found that their 13 resident L2 participants excelled on the Michigan Test of Aural Comprehension, while their nine international L2 learners scored equally well on this test and the Michigan Test of English Language Proficiency, and significantly better than the resident group for the latter test. In terms of participants' writing scores, the international L2 group's scores were higher than those of the resident L2 group. Such studies indicate that differences exist across the two groups in terms of their academic strengths and weaknesses; however, the differences with regard to their writing are inconsistent.

Perhaps the first large-scale, systematic quantitative study comparing international and resident L2 learners' writing is Levi's (2004) unpublished doctoral dissertation. Based on analyses of 140 participants' writing in search of statistically significant differences in errors in writing, Levi found that international L2 and resident L2 participants produced similar numbers of lexico-grammatical and rhetorical errors. When errors were divided into subcategories, however, differences between the two groups emerged. Similarly, Mikesell (2007) compared grammatical error patterns across international L2 and resident L2 students' writing samples with a specific focus on past participle errors. Mikesell found that both groups produced the same percentage of errors, but they differed in terms of error types. When linguistic context was taken into account, the international L2 learners' errors stemmed mainly from producing the correct form but using it in an inappropriate context, while the resident L2 learners' errors were related primarily to producing an incorrect form. Continuing this line of research focusing on errors, Doolan (2013, 2014) also found statistically significant differences between international and resident L2 groups in terms of error patterns in their writing. Interestingly, while results from these quantitative studies suggest that international and resident L2 learners differ with regard to their writing, the researchers propose different solutions: Levi (2004) recommends creating writing courses for resident L2 learners separate from those for international learners, Mikesell (2007) proposes different types of grammar instruction for each, and Doolan (2013, 2014) recommends treating resident L2 learners as native English speakers and not as L2 learners at all.

While the studies by Levi (2004), Mikesell (2007), and Doolan (2013, 2014) provide much-needed empirical evidence regarding differences in the writing ability of international and resident L2 writers, they are inadequate. By focusing on learners' errors, such studies are limited in that they reflect a deficit perspective of each group's writing ability, highlighting learners' shortcomings rather than their potential strengths. Perhaps more importantly from an assessment perspective, by focusing almost exclusively on learners' grammatical performance in writing, these studies reflect an impoverished construct definition of writing ability. A more holistic view of learners' writing would permit a focus on each group's strengths and weaknesses. Moreover, it is possible that the two groups differ in aspects of their writing other than in grammatical (in)accuracies.

Adopting a comprehensive construct definition of writing ability, di Gennaro (2009) examined writing placement samples from 97 students (54 international L2 and 43 resident L2) who were scored on five different components of writing ability (grammatical, cohesive, rhetorical, sociolinguistic, and content control) along with essay length. Results showed the two groups differed only with regard to rhetorical control and essay length. While this study reflects an improvement over previous studies in terms of construct representativeness, the definitions of the two learner groups could have been more rigorous. Specifically, international and resident L2 learners were distinguished only in terms of location of high school completion, without consideration of participants' length of residence in the US, which allowed some long-term resident participants who had completed high school overseas to qualify as international L2 participants. More recently, di Gennaro (2013) adopted a more precise distinction between the two groups, including both high school location and length of residence, along with a fine-tuned definition of writing ability. The latter study, which included 134 participants (67 in each group), found that the international L2 learners scored slightly higher than the resident L2 learners in overall writing ability, and a bias analysis revealed that the two groups differed statistically only with regard to grammatical control. When the two groups were analyzed separately, results showed they had opposing strengths and weaknesses in grammatical control and sociopragmatic control. Grammatical control resulted in being the easiest among the five components for the international L2 group, and sociopragmatic control was the easiest for the resident L2 learners. Grammatical control was the second-most difficult component for the resident L2 learners, as was sociopragmatic control for the international L2 learners. Based on these findings, di Gennaro (2013) agrees with Levi (2004) and Mikesell (2007) in concluding that both international and resident L2 learners demonstrate a need for L2 writing instruction at the post-secondary level. Rather than segregate resident

L2 learners from international L2 students as these scholars propose, di Gennaro (2013) agrees with Matsuda (2008) in suggesting that programs can provide instruction relevant to both types of L2 learners in the same courses.

Research questions

The current study builds upon di Gennaro (2013) in that it continues the search for empirical evidence confirming (or not) that differences exist in the writing ability of international and resident L2 participants. Drawing upon the same dataset as in di Gennaro (2013), the current study subjected the data to additional analyses, providing another opportunity for differences (or similarities) to emerge. Only by analyzing both groups together for each individual component can such expectations be confirmed (or not). Thus, for the current study, five whole-group analyses were conducted: one for each individual component of writing ability. Examining results from whole-group analyses for each component can reveal how each group performed with respect to the other for each individual component, rather than how each group performed with respect to itself across all five components (as in di Gennaro 2013). The research questions addressed in the current study were:

1. How does the writing performance of international L2 writers compare to that of resident L2 writers in five separate components of writing ability, namely grammatical, cohesive, rhetorical, sociopragmatic, and content control?
2. What implications do the findings have for writing program administrators in post-secondary writing contexts?

Data collection procedures

Participants

Studies comparing international and resident L2 learners typically differentiate the two groups based on participants' educational background (see Bosher and Rowekamp 1998, di Gennaro 2009, Doolan 2013, Levi 2004, Muchinsky and Tangren 1999) or length of residence (see Bitchener and Knoch 2008, Connerty 2009). To strengthen the distinction between the two groups, the current study used both criteria for classifying participants: international L2 participants had completed high school in their home countries and lived in the US for a maximum of three years; resident L2 participants had completed high school and resided in the US for a minimum of three years. These criteria guaranteed that no participant could qualify for both groups. Participants who met one criterion but not the other were excluded from the study.

A total of 134 learners were included in the study: 67 international and 67 resident L2 learners. Participants represented 29 different L1s, with the most prominent being Chinese (61), Spanish (11), Korean (10), Russian (7), and Arabic (7). The median length of residence was less than one year for the international group and six years for the resident group. The median age was 19 years for both groups, as all participants were first-year students at the same post-secondary institution in the US.

Three experienced instructors of post-secondary writing courses for L2 students served as raters. All raters had graduate degrees in Teaching English to Speakers of Other Languages (TESOL) or Applied Linguistics, and had rated placement exams for students entering post-secondary writing courses for several years.

Instruments

All participants responded to the same prompt, instructing them to write an argument essay for or against the point of view that anyone who wants to attend college should be accepted. Five rubrics were developed to score participants' responses, one for each component of the writing construct (see the Appendix). Grammatical control referred to a writer's adherence to lexical and morphosyntactic rules at the sentence level. Moving beyond the sentence, cohesive control referred to the writer's ability to overtly connect ideas within and across clauses and sentences. Rhetorical control differed from cohesive control in that it referred to the writer's ability to organize ideas and supporting information at the discourse level rather than at the sentence level. Organizational cues considered part of rhetorical control might not be overt, as they are in cohesive control. Sociopragmatic control encompassed features categorized as sociolinguistic or pragmatic awareness; that is, it was related to the writer's choice of register, stance and tone within the context of the writing task. Finally, content control was defined as the extent to which a writer elaborated on the topic by providing supporting evidence of the type expected in post-secondary writing contexts.

Participants produced writing samples in class, as part of first-day procedures in their composition courses. They had 45 minutes to read and respond to the writing prompt. Participants were asked if they would be willing to share their writing samples as part of the current study. Participants who agreed completed a demographic information form and signed a consent form.

Three raters were trained to use the five rubrics designed for the current study to evaluate participants' writing and assign scores from 0–5 for each participant and in each of the five components separately. Raters did not have access to information about test takers' backgrounds during the rating process. Following procedures for a fully crossed rating design, each rater read and evaluated all 134 essays in all five components. The sum of the five component scores from each rater produced three composite ratings per participant. These summed ratings were then averaged to yield one score for each participant. Pearson product-moment correlations for raters ranged from 0.679 to 0.818. While moderate, all correlations were statistically significant at the 0.01 level. Since correlations only refer to agreement of examinee rankings and not agreement in the actual scores, Cronbach's coefficient alpha was also calculated as an additional reliability estimate. The resulting alpha value of 0.898 for the overall group of participants indicated very high internal consistency reliability for the ratings.

Data analysis

The analysis of writing ability is problematic, as raters' judgments are needed to evaluate participants' writing, and yet raters are not part of the writing construct. To account for such construct-irrelevant factors, many researchers of writing ability adopt a many-facet Rasch measurement (MFRM) as a statistical tool because MFRM calculates participants' ability levels after taking into account external factors, such as rater severity; therefore it produces a more accurate depiction of educational performance than inferential statistics do. Another advantage of MFRM is that it transforms participants' scores from ordinal scales to equal-interval scales, a process that inferential statistical procedures cannot do (Bond and Fox 2007). MFRM is also considered sample-independent, allowing findings to be generalizable to a larger population (Sudweeks, Reeve and Bradshaw 2005). For these reasons, MFRM was used to analyze and compare the two groups' writing performances for the current study.

Five separate MFRM analyses were performed: one for each individual component of the writing construct. Analyses were conducted with the FACETS computer program (Linacre 2009), which converts participants' raw scores into an equal-interval logit scale for each component. The resulting logit scales allowed for comparisons across groups within each component because participants' converted scores have the same frame of reference (Bond and Fox 2007).

Results

For each MFRM analysis, the FACETS program produces a visual summary in the form of a map, illustrating the dispersion of data. Maps produced from the analyses in the current study are presented in Figures 1.1 to 1.5, which include summaries for grammatical, cohesive, rhetorical, sociopragmatic, and content control. The column on the left in each figure is the equal-interval logit scale that is produced after all facets of the measurement procedure have been taken into account. Depending on participants' performance, the logit scale may have a greater or smaller range. Indeed, the logit scale in Figure 1.1, representing grammatical control, has a greater range (−10 to 10) than the scale in Figure 1.2 representing cohesive control (−8 to 6).

The wider of the two columns in Figures 1.1 to 1.5 displays the 134 participants. Each ‘I’ or ‘R’ represents one participant: participants identified with ‘I’ are international L2 learners; participants identified with ‘R’ are resident L2 learners. Participants’ placement in this column corresponds with each one’s logit score, or ability level, for that component. Participants placed higher in the column are described as having greater ability than participants placed lower in the column. For example, in Figure 1.1, the ‘I’ and ‘R’ at the top of the participant column indicate that each group had one participant who stood out as having greater ability in the component of grammatical control than the rest of the group. The three ‘R’s at the bottom of the same column indicate that the three participants with the lowest scores (and, therefore, the least ability) in grammatical control were all resident L2 participants.

Figure 1.1 FACETS summary for grammatical control

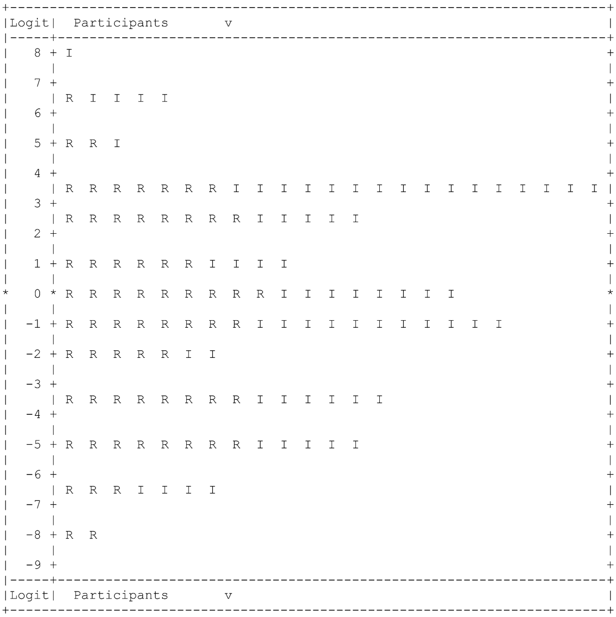
```
+----------------------------------------------------------------------------------------------+
|Logit|  Participants                                                                          |
|-----+----------------------------------------------------------------------------------------+
|  10 + I  R                                                                                   +
|     |                                                                                        |
|   9 +                                                                                        +
|     |                                                                                        |
|   8 +                                                                                        +
|     | R  R  I  I                                                                             |
|   7 +                                                                                        +
|     |                                                                                        |
|   6 +                                                                                        +
|     | R  R  R  R  R  R  R  I                                                                 |
|   5 +                                                                                        +
|     |                                                                                        |
|   4 +                                                                                        +
|     | R  R  I  I  I  I  I  I  I  I                                                           |
|   3 +                                                                                        +
|     | R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I  I  I  I                          |
|   2 +                                                                                        +
|     |                                                                                        |
|   1 +                                                                                        +
|     |                                                                                        |
*   0 * R  R  R  R  R  R  R  R  R  I  I  I  I                                                  *
|     |                                                                                        |
|  -1 +                                                                                        +
|     |                                                                                        |
|  -2 + R  R  R  R  R  R  R  R  R  I  I  I  I  I  I                                            +
|     |                                                                                        |
|  -3 +                                                                                        +
|     | R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I  I  I  I  |
|  -4 +                                                                                        +
|     |                                                                                        |
|  -5 +                                                                                        +
|     |                                                                                        |
|  -6 + R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I                                +
|     |                                                                                        |
|  -7 +                                                                                        +
|     |                                                                                        |
|  -8 + R  R  I  I  I  I  I  I  I  I                                                           +
|     |                                                                                        |
|  -9 +                                                                                        +
|     | R  R  R                                                                                |
| -10 +                                                                                        +
|-----+----------------------------------------------------------------------------------------+
|Logit|  Participants                                                                          |
+-----------------------------------------------------------------------------------------------
```

Figure 1.2 FACETS summary for cohesive control

```
+------------------------------------------------------------------------------------+
|Logit|  Participants                                                                |
|-----+------------------------------------------------------------------------------+
|   6 +                                                                              +
|     |                                                                              |
|     | I  I                                                                         |
|   5 +                                                                              +
|     |                                                                              |
|     | I  I  I  I                                                                   |
|   4 +                                                                              +
|     |                                                                              |
|     | R  R  R  R  R  R  R  R  R  I  I  I  I                                        |
|   3 +                                                                              +
|     |                                                                              |
|     |                                                                              |
|   2 + R  R  R  R  R  I  I  I  I  I  I  I  I  I                                     +
|     |                                                                              |
|     |                                                                              |
|   1 +                                                                              +
|     | R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I                         |
|     |                                                                              |
*   0 *                                                                              *
|     |                                                                              |
|     | R  R  R  R  R  R  R  R  R  R  R  I  I  I  I  I                               |
|  -1 +                                                                              +
|     |                                                                              |
|     | R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I |
|  -2 +                                                                              +
|     |                                                                              |
|     |                                                                              |
|  -3 +                                                                              +
|     | R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I                            |
|     |                                                                              |
|  -4 +                                                                              +
|     |                                                                              |
|     | R  R  R  R  I  I  I  I  I  I                                                 |
|  -5 +                                                                              +
|     |                                                                              |
|     | R  R  R  R  R  I  I  I                                                       |
|  -6 +                                                                              +
|     |                                                                              |
|     |                                                                              |
|  -7 + R  R  R  I  I                                                                +
|     |                                                                              |
|     |                                                                              |
|  -8 +                                                                              +
|-----+------------------------------------------------------------------------------+
|Logit|  Participants                                                                |
+------------------------------------------------------------------------------------+
```

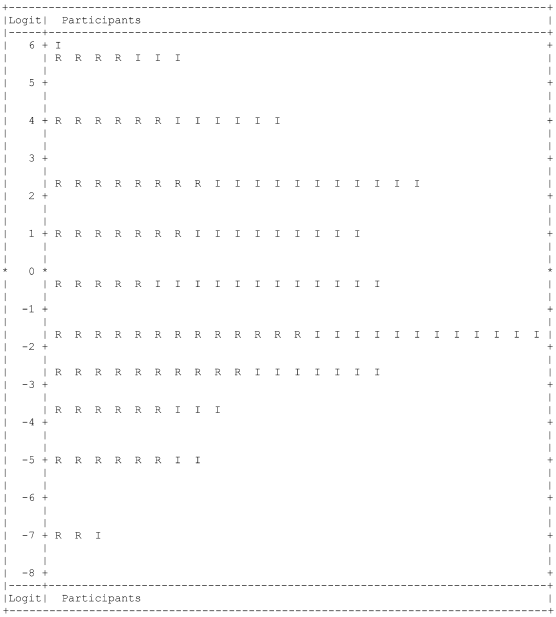
A glance at the maps for each component reveals that representatives from the international L2 group are consistently among the highest scoring participants in each component, a position shared with representatives from the resident group for the components of grammatical, sociopragmatic, and content control. Conversely, the lowest scoring participants on each map are consistently from the resident L2 group, with international participants sharing this position for the components of cohesive and sociopragmatic control. Apart from the extreme scores, participants from both groups achieved a wide range of logit scores, with neither group appearing particularly stronger or weaker than the other. The majority of participants from both groups cluster in the middle of the scale, indicating a normal distribution of scores.

Figure 1.3 FACETS summary for rhetorical control



```
+---------------------------------------------------------------------------+
|Logit|  Participants       v                                               |
|-----+---------------------------------------------------------------------+
|   8 + I                                                                   +
|     |                                                                     |
|   7 +                                                                     +
|     | R  I  I  I  I                                                       |
|   6 +                                                                     +
|     |                                                                     |
|   5 + R  R  I                                                             +
|     |                                                                     |
|   4 +                                                                     +
|     | R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I |
|   3 +                                                                     +
|     | R  R  R  R  R  R  R  R  I  I  I  I  I                               |
|   2 +                                                                     +
|     |                                                                     |
|   1 + R  R  R  R  R  R  I  I  I  I                                        +
|     |                                                                     |
*   0 * R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I                   *
|     |                                                                     |
|  -1 + R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I             +
|     |                                                                     |
|  -2 + R  R  R  R  R  I  I                                                 +
|     |                                                                     |
|  -3 +                                                                     +
|     | R  R  R  R  R  R  R  R  I  I  I  I  I  I                            |
|  -4 +                                                                     +
|     |                                                                     |
|  -5 + R  R  R  R  R  R  R  R  I  I  I  I  I                               +
|     |                                                                     |
|  -6 +                                                                     +
|     | R  R  R  I  I  I  I                                                 |
|  -7 +                                                                     +
|     |                                                                     |
|  -8 + R  R                                                                +
|     |                                                                     |
|  -9 +                                                                     +
|-----+---------------------------------------------------------------------+
|Logit|  Participants       v                                               |
+---------------------------------------------------------------------------+
```

Figure 1.4 FACETS summary for sociopragmatic control



```
+---------------------------------------------------------------------------------+
|Logit|  Participants                                                             |
|-----+---------------------------------------------------------------------------+
|   6 + I                                                                         +
|     | R  R  R  R  I  I  I                                                       |
|     |                                                                           |
|   5 +                                                                           +
|     |                                                                           |
|     |                                                                           |
|   4 + R  R  R  R  R  R  I  I  I  I  I  I                                        +
|     |                                                                           |
|     |                                                                           |
|   3 +                                                                           +
|     |                                                                           |
|     | R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I                   |
|   2 +                                                                           +
|     |                                                                           |
|     |                                                                           |
|   1 + R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I                            +
|     |                                                                           |
|     |                                                                           |
*   0 *                                                                           *
|     | R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I  I                         |
|     |                                                                           |
|  -1 +                                                                           +
|     |                                                                           |
|     | R  R  R  R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I  I |
|  -2 +                                                                           +
|     |                                                                           |
|     | R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I                         |
|  -3 +                                                                           +
|     |                                                                           |
|     | R  R  R  R  R  R  I  I  I                                                 |
|  -4 +                                                                           +
|     |                                                                           |
|     |                                                                           |
|  -5 + R  R  R  R  R  R  I  I                                                    +
|     |                                                                           |
|     |                                                                           |
|  -6 +                                                                           +
|     |                                                                           |
|     |                                                                           |
|  -7 + R  R  I                                                                   +
|     |                                                                           |
|     |                                                                           |
|  -8 +                                                                           +
|-----+---------------------------------------------------------------------------+
|Logit|  Participants                                                             |
+---------------------------------------------------------------------------------+
```

Figure 1.5 FACETS summary for content control

```
+---------------------------------------------------------------+
|Logit|  Participants                                           |
|-----+---------------------------------------------------------+
|   8 +                                                         +
|     |                                                         |
|     | R  R  I                                                 |
|   7 +                                                         +
|     |                                                         |
|     |                                                         |
|   6 + R  I  I                                                 +
|     |                                                         |
|     |                                                         |
|   5 +                                                         +
|     | R  I  I  I  I  I                                        |
|     |                                                         |
|   4 +                                                         +
|     | R  R  R  R  I  I  I  I  I  I  I  I  I  I                |
|     |                                                         |
|   3 +                                                         +
|     |                                                         |
|     | R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I          |
|   2 +                                                         +
|     |                                                         |
|     | R  R  R  R  R  I  I  I  I  I  I  I  I  I  I  I          |
|   1 +                                                         +
|     |                                                         |
|     |                                                         |
*   0 * R  R  R  R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I *
|     |                                                         |
|     |                                                         |
|  -1 +                                                         +
|     | R  R  R  R  R  R  R  R  R  I  I  I  I  I                |
|     |                                                         |
|  -2 +                                                         +
|     |                                                         |
|     | R  R  R  R  I  I  I  I  I  I                            |
|  -3 +                                                         +
|     |                                                         |
|     |                                                         |
|  -4 + R  R  R  R  R  R  R  R  R  I  I  I  I  I  I  I  I  I    +
|     |                                                         |
|     |                                                         |
|  -5 + R  R  R  R  I  I  I  I  I                               +
|     |                                                         |
|     |                                                         |
|  -6 + R  R  R  R                                              +
|     |                                                         |
|     |                                                         |
|  -7 +                                                         +
|     | R  R                                                    |
|     |                                                         |
|  -8 +                                                         +
|-----+---------------------------------------------------------+
|Logit|  Participants                                           |
+---------------------------------------------------------------+
```

While the FACETS maps provide an overview of participants' performance and dispersion in each component, Table 1.1 provides more precise statistics comparing the two groups' performances.

Table 1.1 presents information from each component analysis in terms of the mean, maximum, minimum, range, and standard deviation for each group. Statistics are presented in logit scores: the higher the logit score, the greater the ability level. As logit scales were all centered around 0, minimum scores and some mean scores have negative values.

As shown in Table 1.1, when comparing the mean score for each component across the two groups, the international L2 group's mean is consistently higher than the resident L2 group's mean. The maximum score for each component is also consistently higher for the international group with one exception: the groups shared the same maximum score for content control. For three of the five components, the minimum score is lowest for the resident L2 group; in the other two components, they share the same minimum score. The range for both groups is similar for grammatical and rhetorical control, larger for the international group in cohesive and sociopragmatic control, and larger for the resident group for content control. Standard deviations for each component are similar across the two groups.

The statistics presented in the lower portion of Table 1.1 provide additional information about the dispersion of the data for each component and each group. The separation ratio, which ranges from 0 to infinity, indicates the spread of scores within that component: the higher the separation ratio, the more dispersed the participants are within that component. The separation ratio for all but two cells in this row is above 2.00, indicating that participants reflected a range of ability levels. The next row refers to the strata (also referred to as the separation index). This statistic identifies the number of statistically distinct levels into which participants in that component can be separated. For example, the strata statistic of 4.26 in the first column indicates that international participants in the analysis for the component of grammatical control spread across more than four statistically distinct ability levels. The following statistic, the reliability of separation, can range from 0 to 1 and is similar to Cronbach's alpha in that it indicates the degree of reliability with which participants' logit scores, or ability levels, are distinct from one another. The chi-square tests for each analysis were significant, confirming that participants' logit scores within each analysis were distinct from one another.

Table 1.2 provides the difference, in logits, between the mean, maximum, and minimum scores across the two groups. These data illustrate that mean scores across the two groups were not very different from one another, as there is little more than one logit between means for the largest difference, which is in rhetorical control. The rows for maximum and minimum scores highlight that, when individual scores are examined, differences may be more noticeable than they are for mean scores.

Table 1.1 Summary statistics for international L2 and resident L2 groups within each component (in logit scores)

	Grammatical		Cohesive		Rhetorical		Sociopragmatic		Content	
	IL2*	RL2**	IL2	RL2	IL2	RL2	IL2	RL2	IL2	RL2
Mean	−1.25	−1.32	−0.66	−1.15	0.30	−0.89	0.29	−0.60	0.33	−0.74
Maximum	11.53	9.83	5.41	3.19	7.87	6.35	8.97	5.67	7.41	7.41
Minimum	−8.08	−9.55	−7.15	−7.15	−6.55	−8.19	−7.10	−7.10	−5.01	−7.33
Range	19.61	19.38	12.56	10.34	14.42	14.54	16.07	12.77	12.42	14.74
Standard deviation	4.56	4.46	3.10	2.92	3.57	3.38	2.93	3.25	3.31	3.44
Separation	2.94	2.31	2.14	1.98	2.66	2.36	1.87	2.18	2.40	2.55
Strata	4.26	3.42	3.19	2.98	3.88	3.48	2.82	3.24	3.54	3.74
Reliability of separation	0.90	0.84	0.82	0.80	0.88	0.85	0.78	0.83	0.85	0.87
Chi-square	666.80	597.80	471.20	392.40	537.60	471.00	296.90	406.60	580.70	567.40
Significance	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

**International L2 ***Resident L2*

Table 1.2 Difference in scores across groups within each component (in logit scores)

	Grammatical	Cohesive	Rhetorical	Sociopragmatic	Content
Mean	0.07	0.49	1.19	0.89	1.07
Maximum	1.70	2.22	1.52	3.30	0.00
Minimum	1.47	0.00	1.64	0.00	2.32

Implications and discussion

The comparisons of participants' placements along the logit scales for each component displayed in Figures 1.1 to 1.5, as well as the group statistics for each component as displayed in Tables 1.1 and 1.2, show that the international L2 group is, on average, slightly stronger than the resident L2 group for each of the five components of writing ability measured. International L2 participants also consistently produced the highest scores for each component, while resident L2 or Generation 1.5 participants produced the lowest scores in all five components, sharing the low score with the international group for two components.

In response to the first research question ('How does the writing performance of international L2 writers compare to that of resident L2 writers in five separate components of writing ability, namely grammatical, cohesive, rhetorical, sociopragmatic, and content control?'), these data suggest that the two groups are not very different from each other in terms of their writing ability, at least in terms of overall scores in the five components of writing evaluated in the current study.

The second research question is: 'What implications do the findings have for writing program administrators in post-secondary writing contexts?' Some scholars believe that resident L2 writers are comparable to native English L1 writers, and, thus, should not be placed in composition courses for L2 learners (see Doolan 2013, 2014). While the current study did not include a native L1 English group for comparison, results nonetheless confirm that the two L2 groups are comparable to each other in terms of their writing abilities. The implication of this finding is that if specialized writing instruction is offered to international L2 students when they enter a US post-secondary institution, it should be offered to resident L2 students as well. Perhaps if L2 composition courses were reframed as a benefit for L2 learners of all types, rather than a disadvantage holding students back, fewer students (and scholars) would resist considering resident students as potential candidates for L2 composition courses. While it might be favorable in some cases to distinguish resident L2 students from international L2 students in terms of their socioeconomic and educational backgrounds, empirical evidence from the current study supports the interpretation that it is a 'myth' that international and resident L2 learners cannot be taught in the same classroom (Matsuda 2008). Thus, a second implication is that creating composition courses to separate resident L2 writers from their international L2 peers for writing instruction is unnecessary if placement is based on students' writing ability. Finally, implicit in studies comparing resident L2 writers to native English L1 writers is the assumption that instruction for L2 learners is somehow insufficient or inferior to that for L1 writers. Results from the current study imply that international L2 learners will not necessarily lower the level of instruction for resident L2 learners. On the contrary, the presence of both international and resident L2 writers in the same course can be beneficial as their different backgrounds and experiences allow for greater depth and breadth of instruction for each population.

The growing number of resident L2 learners in post-secondary composition courses has increased the visibility of such students in recent years. As a result, the level of awareness among composition instructors and scholars that some L2 students are also long-term US residents has increased as well. While it is important to recognize that not all L2 learners are newcomers or international students who have completed secondary education in their L1s, the revelation that some L2 writers may also be long-term US residents appears to have distracted attention from the influence of their linguistic backgrounds on their writing ability. Instructors and scholars of L2 students acknowledge the diversity of L2 students with regard to their countries of origin and L1s, yet none currently suggest dividing students into separate writing courses based on their ethnic backgrounds. Students' length of residence in the US or the location of their secondary education, while important characteristics, may simply be additional facets contributing to the heterogeneity of the L2 student body. Unless studies can show that resident L2 learners' writing improves more when they are placed into composition courses for native English L1 writers than when they are placed into composition courses for L2 students, diminishing resident learners' L2 status seems ingenuous at best and irresponsible at worst, especially in cases where composition courses have been designed with L2 learners' strengths and weaknesses in mind.

Acknowledgements

I would like to acknowledge The International Research Foundation for English Language Education (TIRF) for providing me with a Doctoral Dissertation Grant (DDG) to conduct this study. I would like to thank the participants at the Language Testing Research Forum in Princeton, New Jersey, whose comments and questions helped me clarify the study's findings. I am especially grateful to James Purpura, Hansun Zhang Waring, Monika Ekiert and Phil Choong for their insightful feedback at various stages of the study.

References

Bitchener, J and Knoch, U (2008) The value of written corrective feedback for migrant and international students, *Language Teaching Research* 12, 409–431.

Bond, T G and Fox, C M (2007) *Applying the Rasch Model: Fundamental Measurement in Human Sciences*, Mahwah: Lawrence Erlbaum.

Bosher, S (1998) The composing processes of three Southeast Asian writers at the post-secondary level: An exploratory study, *Journal of Second Language Writing* 7, 205–241.

Bosher, S and Rowekamp, J (1998) The refugee/immigrant in higher education: The role of educational background, *College ESL* 8, 23–42.

Connerty, M (2009) *Variation in academic writing among generation 1.5 learners, native English-speaking learners and ESL learners: The discoursal self of G1.5 student writers*, unpublished PhD thesis, University of Birmingham.

di Gennaro, K (2008) Assessment of Generation 1.5 learners for placement into college writing courses, *Journal of Basic Writing* 27, 61–79.

di Gennaro, K (2009) Investigating differences in the writing performance of international and Generation 1.5 students, *Language Testing* 26, 533–559.

di Gennaro, K (2012) The heterogeneous second-language population in US colleges, *Teaching English in the Two-Year College* 40, 57–67.

di Gennaro, K (2013) How different are they? A comparison of Generation 1.5 and L2 international learners' writing ability, *Assessing Writing* 18, 154–172.

Doolan, S M (2013) Generation 1.5 writing compared to L1 and L2 writing in first-year composition, *Written Communication* 30, 135–163.

Doolan, S M (2014) Comparing language use in the writing of developmental Generation 1.5, L1, and L2 tertiary students, *Written Communication* 31, 215–247.

Ferris, D R (2009) *Teaching College Writing to Diverse Student Populations*, Ann Arbor: University of Michigan Press.

Frodesen, J and Starna, N (1999) Distinguishing incipient and functional bilingual writers: Assessment and instructional insights gained through second-language writer profiles, in Harklau, L, Losey, K and Siegal, M (Eds) *Generation 1.5 Meets College Composition: Issues in the Teaching of Writing to US-educated Learners of ESL*, Mahwah: Lawrence Erlbaum, 61–80.

Institute of International Education (2014) *Open Doors Data*, available online: www.iie.org/Research-and-Publications/Open-Doors/Data/International-Students

Leki, I (1999) 'Pretty much I screwed up': Ill-served needs of a permanent resident, in Harklau, L, Losey, K and Siegal, M (Eds) *Generation 1.5 Meets College Composition: Issues in the Teaching of Writing to US-educated Learners of ESL*, Mahwah: Lawrence Erlbaum, 17–43.

Levi, E I (2004) *A study of linguistic and rhetorical features in the writing of non-English language background graduates of US high schools*, unpublished PhD thesis, University of Pennsylvania.

Linacre, M (2009) *FACETS Rasch Measurement Computer Program*, Chicago: MESA Press.

Matsuda, P K (2008) Myth 8: International and US resident ESL writers cannot be taught in the same class, in Reid, J (Ed) *Writing Myths*, Ann Arbor: University of Michigan Press, 159–176.

Mikesell, L (2007) Differences between Generation 1.5 and English as a second language writers: A corpus-based comparison of past participle use in academic writing, *CATESOL Journal* 19, 7–29.

Muchinsky, D and Tangren, N (1999) Immigrant student performance in an academic intensive English program, in Harklau, L, Losey, K and Siegal, M (Eds) *Generation 1.5 Meets College Composition: Issues in the Teaching of Writing to US-educated Learners of ESL*, Mahwah: Lawrence Erlbaum, 211–234.

National Clearinghouse for English Language Acquisition (2008) *The growing numbers of limited English proficiency students (mini-poster)*, available online: www.ncela.us/files/uploads/9/growingLEP_0809.pdf

Reid, J (2006) 'Eye' learners and 'ear' learners: Identifying the language needs of international students and US resident writers, in Matsuda, P K, Cox M, Jordan, J and Ortmeier-Hooper, C (Eds) *Second-language Writing in the Composition Classroom: A Critical Sourcebook*, Boston: Bedford/St. Martin's, 76–88.

Roberge, M, Siegal, M and Harklau, L (2009) *Generation 1.5 in College Composition: Teaching Academic Writing to US-educated Learners of ESL*, New York: Routledge.

Sudweeks, R R, Reeve, S and Bradshaw, W S (2005) A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing, *Assessing Writing* 9, 239–261.

Thonus, T (2003) Serving Generation 1.5 learners in the university writing center, *TESOL Journal* 12, 17–24.

Appendix

Scoring rubrics

Grammatical control

5 Excellent control	Lexical, syntactic and graphical forms are always accurate. This includes (among other things): word forms, verb tenses and forms, word order, spelling.
4 Very good control	Lexical, syntactic and graphical forms are almost always accurate, but meaning is never obscured. This includes (among other things): word forms, verb tenses and forms, word order, spelling.
3 Sufficient control	Lexical, syntactic and graphical forms are sometimes inaccurate, but meaning is rarely obscured. This includes (among other things): word forms, verb tenses and forms, word order, spelling.
2 Limited control	Lexical, syntactic and graphical forms are often inaccurate and meaning is sometimes obscured. This includes (among other things): word forms, verb tenses and forms, word order, spelling.
1 Little control	Lexical, syntactic and graphical forms are mostly inaccurate, obscuring meaning throughout the essay. This includes (among other things): word forms, verb tenses and forms, word order, spelling.
0 No control	There are too few sentences to judge.

Cohesive control

5 Excellent control	Ideas are overtly linked throughout the essay. Use of cohesive devices (logical connectors, repetition, synonyms) is always accurate. Compound and complex sentences are used accurately to create clear connections across sentences and paragraphs.
4 Very good control	Ideas are overtly linked throughout the essay. Use of cohesive devices (logical connectors, repetition, synonyms) is almost always accurate. Compound and complex sentences are used almost accurately to create clear connections across sentences and paragraphs.
3 Sufficient control	Ideas are often overtly linked throughout the essay. Use of cohesive devices (logical connectors, repetition, synonyms) is mostly accurate, but may be lacking in places. Sometimes inaccurate use of compound and complex sentences leads to unclear connections across sentences and paragraphs.
2 Limited control	Ideas are sometimes overtly linked throughout the essay. Use of cohesive devices (logical connectors, repetition, synonyms) is often inaccurate or inadequate. An abundance of simple sentences may create a disconnected, staccato quality.

1 Little control	Ideas are rarely overtly linked throughout the essay. Use of cohesive devices (logical connectors, repetition, synonyms) is inaccurate or inadequate. An abundance of simple sentences creates a disconnected, staccato quality.
0 No control	There are too few sentences to judge.

Rhetorical control

5 Excellent control	The essay exhibits overall coherence (unity) through underlying logical organization and maintains a consistent point of view. The essay's structure (i.e. introduction, thesis, support, conclusion) conforms to the expectations of the genre and each part is connected to the same response/thesis.
4 Very good control	The essay exhibits mainly overall coherence (unity) through underlying logical organization and maintains a consistent point of view. The essay's structure (i.e., introduction, thesis, support, conclusion) mainly conforms to the expectations of the genre and each part is connected to the same response/thesis, but the ordering may not be logical.
3 Sufficient control	The essay exhibits some overall coherence (unity) through underlying logical organization and mostly maintains a consistent point of view. The essay's structure (i.e., introduction, thesis, support, conclusion) conforms somewhat to the expectations of the genre, but each part may not be connected to the same response/thesis, or the ordering may not be logical.
2 Limited control	The essay exhibits very little overall coherence (unity) through underlying logical organization and the writer's point of view appears inconsistent or vague. The essay's structure (i.e., introduction, thesis, support, conclusion) conforms very little to the expectations of the genre, and each part may not be connected to the same response/thesis, or the ordering may not be logical.
1 Little control	The essay exhibits no overall coherence (unity) through logical organization and the writer's point of view is inconsistent or vague. The essay's structure (i.e., introduction, thesis, support, conclusion) does not conform at all with the expectations of the genre, and each part may not be connected to the same response/thesis and the ordering is not logical. The essay is mostly writer-oriented and difficult to follow – the writer constantly needs to backtrack while reading to understand the writer's logic.
0 No control	There are too few sentences to judge.

Sociopragmatic control

5 Excellent control	Register is always consistent with academic writing (i.e., assumes an impersonal and formal relationship with the reader). Expression of stance always displays respect for members of the discourse community and reflects an appropriate attitude for an academic context.
4 Very good control	Register is almost always consistent with academic writing (i.e. almost always assumes an impersonal and formal relationship with the reader). Expression of stance almost always displays respect for members of the discourse community and almost always reflects an appropriate attitude for an academic context.
3 Sufficient control	Register is mostly consistent with academic writing, but may adopt a conversational register in places, such as through personal references and overt use of second person pronoun/generic 'you'. Expression of stance mainly displays respect for members of the discourse community and mainly displays an appropriate attitude for an academic context, but different viewpoints may be limited by such things as imperatives and generalizations, or the writer's attitude may appear overly assertive in places.
2 Limited control	Register is often inconsistent with academic writing (i.e. may assume a personal relationship with the reader and adopt a conversational register, such as through personal references and overt use of second person pronoun/generic 'you'). Expression of stance may limit different viewpoints by including such things as imperatives and overgeneralizations and the writer sometimes fails to reflect an appropriate attitude for an academic context (i.e. may be overly assertive).
1 Little control	Register is inconsistent with academic writing (i.e. assumes a personal relationship with the reader). Expression of stance limits different viewpoints by including such things as imperatives and overgeneralizations and writer fails to adopt an appropriate attitude for an academic context (i.e., may be overly assertive).
0 No control	There are too few sentences to judge.

Content control

5 Excellent control	The essay responds to and elaborates intellectually on the topic, and provides appropriate supporting examples and details. All information is relevant.
4 Very good control	The essay mostly responds to and elaborates intellectually on the topic, and provides appropriate supporting examples, but some support could be more developed. All information is relevant.
3 Sufficient control	The essay partly responds to and elaborates intellectually on the topic, but supporting examples and details are not always appropriate or could be more developed. The essay occasionally includes irrelevant information or frequent repetition.

2 Limited control	The essay elaborates very little on or strays from the topic; supporting examples and details are lacking. The essay includes irrelevant information or frequent repetition.
1 Little control	The essay does not elaborate on the topic at all; supporting examples and details are lacking.
0 No control	There are too few sentences to judge.

2 The role of voice in L2 argumentative writing: The development and validation of an analytic rubric

Cecilia Guanfang Zhao

Shanghai International Studies University, China

Motivation for the research

Despite its elusive nature and the debate among researchers about its viability as a pedagogical focus in writing instruction (e.g. Helms-Park and Stapleton 2003, Matsuda and Tardy 2007, 2008, Stapleton 2002), voice remains a key concept addressed in writing textbooks and most of the US state learning standards (Ramanathan and Kaplan 1996, Zhao and Llosa 2008). In fact, it is often also an integral part of the various rubrics used for the evaluation of students' writing across the contexts of high-stakes testing and classroom-based assessment at both secondary and post-secondary levels (DiPardo, Storms and Selland 2011, Jeffery 2007, Llosa, Beck and Zhao 2011, Zhao and Llosa 2008). The inclusion of this concept in writing textbooks, learning standards, and assessment rubrics seems to suggest that voice plays an important role in writing instruction and assessment. In reality, however, no empirical evidence has been offered to confirm or refute this proposition. The concept of voice is only loosely defined in the literature and impressionistically assessed in practice. Few attempts have ever been made, in the field of language assessment in general and writing assessment in particular, to formally investigate whether and how the strength of an authorial voice in written texts can be reliably measured. As a response, this study seeks to develop and validate an analytic rubric that can be used to capture voice in written texts and, at the same time, offer insight into how this seemingly intangible concept might be approached in writing instruction and assessment.

Literature review

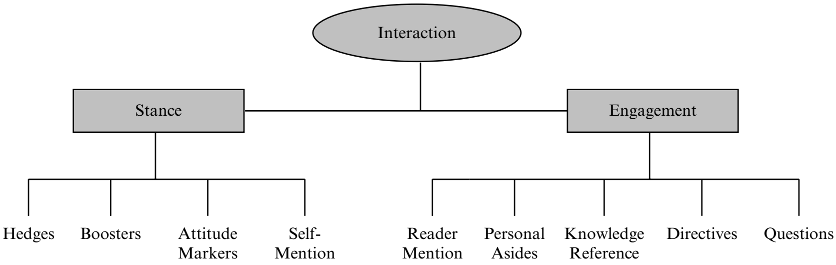
A review of the relevant literature shows that there is a plethora of abstract discussions and theoretical conceptualizations of the notion of voice.

Some researchers in the field of first language (L1) composition studies, for instance, argue that voice is an individual trait that writers can discover from within themselves through practices of expressive writing (e.g. Elbow 1968, 1981, 1994, 1999, Graves 1983, 1994, Holding 2005, Stewart 1969, 1972, 1992). Voice in a written text is hence regarded as 'the expression of the essential individuality of a particular writer' (Stewart 1992:283) and 'an ideal metaphor for individualism' (Elbow 1999:334). It is believed to be 'the fundamental quality of good writing' (Stewart 1992:283).

However, as the student population in the US becomes more and more diverse, both culturally and linguistically, this notion of voice laden with the mainstream US ideology of individualism has gradually come under scrutiny, particularly by scholars who believe that voice, and language by extension, is socially and culturally constructed. This emphasis on social context, therefore, leads to various alternative conceptualizations of voice in the literature, especially literature on second language (L2) writing. For example, both Ede (1992) and Bowden (1999) argue that writers often adopt different voices for different rhetorical situations, just as they would dress differently for different occasions. This conception is similar to Yancey's (1994) notion of 'multiple voices', which is also echoed in many other researchers' discussions and studies (e.g. Farmer 1995, Hirvela and Belcher 2001, Ivanič and Camps 2001, Kumamoto 2002, O'Leary 1993). Further complicating the definition of voice, Prior (2001) argues that voice should not be represented as either having a personal and individualistic nature or being socially embedded in a discourse community. Instead, he believes that voice is simultaneously personal and social. Drawing on all these insights, Matsuda (2001) presents a relatively more formal definition of voice. According to him, 'voice is the amalgamative effect of the use of discursive and non-discursive features that language users choose, deliberately or otherwise, from socially available, yet ever changing repertoires' (Matsuda 2001:40).

Regardless of the many attempts to capture the nature and characteristics of voice in writing, this concept remains a slippery construct that is difficult to define in theory and operationalize in practice. Most recently, Hyland (2008) proposes a more comprehensive model that sees voice in academic writing as essentially interaction between writers and readers. This interactional model of voice comprises two major dimensions, or 'systems' as he puts it. One is the writer-oriented *stance dimension*, which refers to how writers present themselves, their opinions, and their arguments through the use of four linguistically available elements: hedges, boosters, attitude markers, and authorial self-mention. The other is the reader-oriented *engagement dimension* that is realized through the use of five other linguistic- and discourse-level elements: reader mentions, personal asides, references to shared knowledge, directives, and (rhetorical/audience-directed) questions. Figure 2.1 visually represents the model, and definitions for the linguistic elements themselves; these are available in Appendix A for readers' reference.

Figure 2.1 Hyland's (2008) interactional model of voice



Hyland's interactional model of voice incorporates both the individualistic facet of the concept of voice, as represented by the stance dimension, and the interdependent facet of voice, as reflected in the engagement dimension. While acknowledging certain intangible aspects of the concept, this model, therefore, offers a systematic way of examining the construction and realization of voice in academic written discourse through the use of both linguistic- and discourse-level elements available in the English language.

While theoretical conceptualizations of voice proliferate in the literature, no empirical study has yet been done to translate any of these theoretical, often also rather abstract, conceptions of voice into research-friendly instruments or pedagogically useful tools that writing researchers and teachers could employ to either facilitate empirical research or inform writing pedagogy for the teaching and learning of voice. Although writing rubrics that contain voice as a criterion for evaluation do exist, voice in those existing rubrics is either evaluated holistically in very broad terms based solely on the reader's general impression (e.g. mature/immature in Yeh 1998, absent/distinctive in DeRemer 1998, and lifeless/compelling in various versions of the Six-Trait Scoring Rubric), or considered as a subcomponent under some other analytical dimensions on the rating scale (e.g. New York State Education Department 2007, Wolfe, Bolton, Feltovich and Niday 1996). As a result, the construct of voice, as operationalized in these rubrics, remains too elusive to be pedagogically useful.

Based on Weigle's (2002) review of different scale types (i.e. primary trait, multiple trait, holistic, and analytic) and Knoch's (2011) discussion of their appropriateness for use in a formative context, it seems that an analytic scale of voice, if available, would better inform pedagogy and facilitate the learning of this concept. In the literature, however, the only attempt to capture voice in the form of an analytic rubric is Helms-Park and Stapleton's (2003) Voice Intensity Rating Scale, which measures voice intensity in categories of assertiveness, self-identification, reiteration of the central point, and authorial presence and autonomy of thought. Nevertheless, even this scale is criticized for its construct underrepresentation (see Matsuda and Tardy 2007, 2008) and lack of formal validation (Zhao and Llosa 2008). Thus, in order to both fill the gap in the literature and demystify the seemingly intangible concept for student writers and writing teachers alike, the present study first develops an analytic voice rubric on the basis of Hyland's (2008) theoretical model and then validates it using empirical data.

Because issues of voice are especially prevalent and salient in the L2 writing community, this study examines the realization and assessment of voice in L2 writing in particular. Additionally, the study focuses on voice in argumentative writing as it is 1) a genre that commonly appears on high-stakes writing tests that could have a large impact on L2 students' educational careers (e.g. Jeffery 2009), and 2) a prevalent type of academic writing that is identified by researchers (e.g. Helms-Park and Stapleton 2003, Ramanathan and Kaplan 1996, Reid 2001) as 'a central component of university writing' (Helms-Park and Stapleton 2003:250).

Research question

Specifically, the study is guided by the following overarching research question: To what extent is the analytic voice rubric developed on the basis of Hyland's (2008) theoretical model of voice in academic written discourse a reliable and valid measure of the strength of an authorial voice in L2 argumentative writing?

Methodology

Materials

A total of 480 argumentative writing samples in response to two TOEFL$_{®}$ iBT independent writing prompts were used in this study with permission from Educational Testing Service (ETS), the copyright owner. Two hundred of such writing samples were used in the rubric development phase (Phase 1) of the study, and another 200 in the rubric validation phase (Phase 2). The rest of the writing samples were used for rater training purposes.

Instrumentation

A preliminary analytic voice rubric was developed based on Hyland's (2008) model of voice (see Figure 2.1). Each individual voice element in that model was translated into a separate analytic category in the rubric, evaluating the salience of that particular voice feature in a writing sample based primarily on frequency counts. In addition to those elements, central point articulation, which was found to be an important component of voice in previous research (Helms-Park and Stapleton 2003, Zhao and Llosa 2008), was added into the preliminary voice rubric as another category. For validation purposes, a category that captures the overall voice strength holistically based on raters' impressions was also added into the rubric. Hence, the preliminary voice rubric contained 11 categories; each was rated on a scale of 0–4, representing the range from the absence of a particular voice feature to extensive use of that feature in a writing sample. A copy of the preliminary voice rubric is presented in Appendix A.

Participants

Participants were six raters recruited from qualified PhD candidates at a major USA university. They all had extensive experience in L1 and/or L2 writing instruction and assessment. Four of them participated in Phase 1 of the study, whereas three raters participated from Phase 1, and two other raters participated in Phase 2 of the study. Five of the raters were native English speakers who were well aware of and quite familiar with the concept of voice. Only one rater who participated in the second phase of the study was a non-native speaker of English who was less familiar with the concept.

Data collection and analysis

Raters first went through a training session that provided them with an explanation of the structure and content of the voice rubric, as well as general instructions on how to use the rubric when rating the salience of voice elements in the writing samples. As a group, raters studied the definitions of the analytical categories in the rubric and the examples of how these voice categories were instantiated. They were also instructed to rate only the salience of voice features instead of writing quality, and that particular attention should be given to the coding of certain linguistic features that could be easily miscoded. For example, words such as 'could' and 'would', which were often used as hedging devices in the writing samples, could well be used in a subjunctive mood or as past tense that had nothing to do with hedging. Therefore, raters were instructed to examine these linguistic elements carefully when coding. During the training session, questions from raters about the rubric itself as well as the application of the rubric were discussed in a group until a consensus was reached. Once raters' voice ratings became more consistent (discrepancies in their ratings were, most of the time, no more than one point), rater training was completed. Raters then rated 200 TOEFL$_{®}$ iBT writing samples for voice strength. Each writing sample was double rated and the average of the two ratings was used in later analysis.

An in-depth think-aloud and post-think-aloud interview session with each of the four raters followed. Qualitative data collected in these sessions focused on raters' actual rating processes and their perception of the validity and applicability of the rubric. Each rater rated three writing samples during the think-aloud session – two were selected from among what they had actually rated with a purpose to check for intra-rater reliability, and a third one was used with all four raters to help shed light on inter-rater reliability. The think-aloud and interview sessions were audio-recorded and subsequently transcribed for later analysis.

Principal component analysis (PCA) was then employed to explore key components that define the construct of voice. Additionally, qualitative data from the think-aloud protocols and interviews were also analyzed to supplement the quantitative analysis and provide additional evidence on rubric reliability, applicability, and construct validity. The preliminary voice rubric was subsequently revised based on such analysis results. The revised rubric was then used to rate another set of 200 TOEFL$_{®}$ iBT writing samples for voice strength. Raters went through a similar training session. And, again, each writing sample was double rated. These voice ratings were then used in confirmatory factor analysis (CFA), PCA, and correlational analyses to test whether the findings from Phase 1 still held in Phase 2 data.

Results

Phase 1: Quantitative data analysis results

The percentage of agreement (ratings that were the same or within one point difference) between pairs of raters ranged from 88% to 93%, indicating high inter-rater reliability. In addition, ratings obtained from raters' think-aloud sessions also pointed to good intra- and inter-rater reliability. Specifically, data showed that the percentage of ratings observed to be the same or within one point difference on the two occasions were all around 91% for the four raters, indicating high intra-rater reliability. In the one writing sample that all the four raters worked with in the think-aloud sessions, eight of the 11 voice categories received the exact same ratings from all four raters, further suggesting a good level of inter-rater reliability.

Descriptive statistics for the voice ratings in Phase 1 (see Table 2.1) showed that the ratings for three voice categories – personal aside (C7), reference to shared knowledge (C8), and the use of (rhetorical and audience-directed) questions (C10) – were extremely skewed. The means for these categories were close to zero, suggesting that these language features were rarely identified in the TOEFL$_{®}$ iBT writing samples. Thus, these three categories were excluded from subsequent statistical analyses.

An examination of the correlation matrix with the remaining voice variables showed that most of the variables were weakly or moderately correlated. PCA was then performed. Using eigenvalue > 1 (Kaiser 1960) as a criterion for selecting principal components, two main components emerged. Table 2.2 shows the component matrix that contains factor loadings for the two components. Variables with factor loadings of 0.32 and above were interpreted, following Tabachnick and Fidell (2007). Hence, all the variables seemed to have loaded on one or the other of the two components, with the exception of the central point articulation (C5) and the use of directives (C9), which double-loaded on both components. Figure 2.2 also visually presents such results.

Table 2.1 Phase 1 analysis results: Descriptive statistics for voice ratings (N = 200)

	Minimum	Maximum	Mean	Standard deviation	Skewness	
					Statistic	Standard error
C1: Hedges	0.00	4.00	1.49	1.06	0.50	0.17
C2: Boosters	0.00	4.00	3.13	0.92	−1.17	0.17
C3: Attitude markers	0.50	4.00	2.64	0.91	−0.29	0.17
C4: Authorial self-mention	0.00	4.00	2.38	1.43	−0.22	0.17
C5: Central point articulation	0.00	4.00	1.59	0.78	0.35	0.17
C6: Reader reference	0.00	4.00	2.58	1.51	−0.52	0.17
C7: Personal aside	0.00	2.50	**0.41**	0.64	**1.48**	0.17
C8: Shared knowledge	0.00	2.00	**0.31**	0.48	**1.59**	0.17
C9: Directives	0.00	4.00	1.58	0.91	0.16	0.17
C10: Rhetorical questions	0.00	3.50	**0.32**	0.73	**2.38**	0.17
C11: Overall voice	0.00	4.00	2.18	0.94	−0.07	0.17

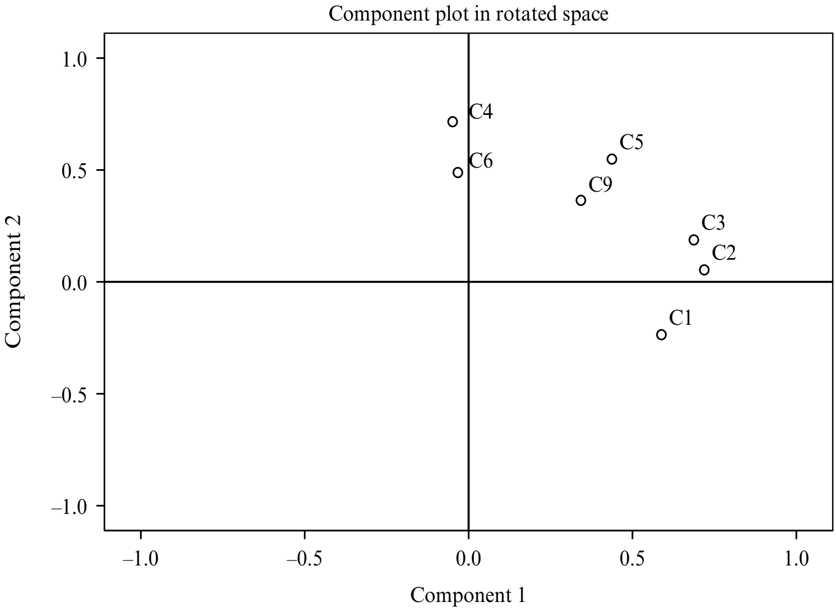
As shown, hedges (C1), boosters (C2), and attitude markers (C3) were more closely associated with one component, which was later interpreted as a writer's manner of presentation. In other words, a writer's use of these linguistic devices indicates whether the author's ideas and arguments are presented assertively, mildly, confidently, tentatively, enthusiastically, or maybe indifferently. The other component, consisting of authorial self-mention (C4) and direct reader reference (C6), was interpreted as writer and reader presence.

Table 2.2 Phase 1 PCA component matrix with factor loadings

	Component	
	1	2
C2: Boosters	0.72	0.05
C3: Attitude markers	0.69	0.19
C1: Hedges	0.59	−0.24
C4: Authorial self-mention	−0.05	0.72
C5: Central point articulation	0.44	0.55
C6: Reader reference	−0.03	0.49
C9: Directives	0.34	0.36

Note: PCA with varimax rotation.

Figure 2.2 Phase 1 component plot



Apparently, authorial self-revelation and direct reference to readers are also able to contribute to the realization of voice in written discourse.

As central point articulation (C5) and the use of directives (C9) double-loaded on the two components with more or less the same magnitude, they would normally be considered for elimination based on purely statistical criteria for determining the number of components. However, a more theoretically informed look at the nature of these two variables suggests that when considered together they form a dimension that is linguistically related to, but conceptually different from, the two identified components. To illustrate this point, consider the following thesis statement: *I* (C4) *firmly* (C2) believe that a *smart* (C3) choice for *us* (C6) to make is to study what *really* (C2) interests *us* (C6) instead of choosing subjects that *might* (C1) prepare *us* (C6) for a job or career. Obviously, within this articulated central point, all the other voice elements are involved. While voice can certainly be captured by all these linguistic-level elements, one should note that the presence or absence of a central point also makes a difference in terms of getting a clear voice across, especially in argumentative writing. This discourse-level component, however, is unlikely to be captured by coding the individual voice elements at the linguistic level alone. Based on such considerations, therefore, these two voice elements, central point articulation (C5) and the use of directives (C9), together were considered as yet another component – presence and clarity of ideas in the content. Quantitative analysis of voice ratings, therefore, have pointed to three major dimensions that together explain how voice is realized in written discourse – the presence and clarity of ideas in the content, manner of idea presentation, and writer and reader presence. Interestingly, these findings were corroborated by qualitative data gathered from the raters' think-aloud and interview sessions. The next section presents such findings in detail.

Phase 1: Qualitative data analysis results

The presence and clarity of ideas in the content in the conception of voice

When reflecting on how they arrived at the holistic and impressionistic ratings for overall voice strength (C11), raters reported that they would first look at whether the essay 'really [made] sense' and whether they were able to 'follow what was going on'. According to Rater 4, this is 'a minimum of voice' (Transcript R4:15). Likewise, Rater 1 also considered the presence of a central point as the most important element in realizing voice: 'if I don't know what the main point that the writer is trying to get across, no matter how friendly or funny they are, I'm going to lose patience and not going to want to engage' (Transcript R1:6).

Other than the presence of a clear idea, raters also pointed to the *uniqueness* of that idea as a major component of voice. Rater 1, for example, shared in the interview that she could hear a higher level of voice when the writer, whom she thought was a sophisticated thinker, 'wasn't afraid to challenge the prompt' (Transcript R1:10). Similarly, Rater 3 reflected that when she thought about voice, she was thinking more about 'something that is unique', and she 'really like[d] it when . . . for sophisticated writing, they had a different take, that they were able to put their own spin on the question' (Transcript R3:15). Rater 4 also believed that his holistic evaluation of voice strength was based on 'how much it stands out from others . . . [and whether] someone had taken a different stance or a slightly different approach' (Transcript R4:16).

In addition to clarity and uniqueness of ideas, raters also saw the creative use of examples and inclusion of details in argumentative writing as evidence of the writer's commitment to the topic under discussion, which in turn served as evidence of a strong voice. According to Rater 2, for example, 'the level of detail . . . shows a level of commitment and thought behind it, which again shows personality and commitment to the argument' (Transcript R2:17). While evaluating voice strength in a writing sample that received a unanimously high rating of overall voice strength, all the raters also noted its use of the directive phrase 'let us imagine' when giving examples to support the argument. They believed that this way of voicing was 'subtle' (Transcript R3:12), 'different' (Transcript R4:16), 'interesting and interactive' (Transcript R2:17), and at the same time, it was 'an open invitation for [readers] to engage' (Transcript R1:3).

The manner of idea presentation in the conception of voice

In addition to what the author says, qualitative data also revealed that raters' perceptions of voice strength were influenced by how the author says what he says. To Rater 4, for example, 'stating things confidently' could also add to voice in writing, especially academic writing. He also reflected on two different ways of stating things with confidence. Some of the writing samples, for example, used boosters frequently and were 'really hedge free, which to [him] spoke well in terms of voice as there was . . . definitely authority' (Transcript R4:11). But he also noticed how confidence and authority could still be found in writings that hedged in a more sophisticated way. As he reported, 'sometimes, there was hedging going on, but I did feel a lot of confidence in the writer and what they were talking about' and the writing was thus 'exuding authority' (Transcript R4:17).

Other than the use of language that shows confidence in presentation of ideas, raters also noted how word choice in general could contribute to the readers' perception of the level of voice strength in written texts. These word choices mainly involved the selection of linguistic markers of essay structure, attitude markers, verbs, and adjectives in general, which the raters thought could show personality in a piece of writing. Rater 1, for instance, thought that words such as 'firstly', 'secondly', and 'lastly' made the essay 'very predictable', and so readers would easily 'lose interest' (Transcript R1:4). If readers lost interest, they would very likely feel like 'drifting off' when reading the text, which to Rater 4 was 'a measure of "there wasn't a voice" [in the text]' (Transcript R4:14). All the raters were also quick to associate five-paragraph essays with voiceless essays. According to them, those were 'cookie-cutter essays' (R3) that were 'formulaic' (R4), 'robotic' (R2) and 'boring' (R1). In general, raters believed that the use of 'interesting vocabulary' was a good way of adding voice: 'those quirky words . . . they're trying to add a stylistic rhetorical flair to it . . . to mixed success. But . . . I think that reflects an authorial style, which for me gives it personality, and I consider that a part of their voice' (Transcript R2:16).

These data suggest that in addition to the presence of a clear and unique idea, the manner through which these ideas are presented is another way of realizing voice in written discourse. Such results were again consistent with the factor analysis results that identified the use of hedges (C1), boosters (C2) and attitude markers (C3) as a unique dimension of voice. A slight difference here in the qualitative data was that raters went beyond the boundaries of the rubric (i.e. focusing only on hedges, boosters and attitude markers) to reflect on how interesting and unique word choice, by extension, could contribute to a high level of voice in writing. Rater 3 actually explicitly stated at the end of the interview session that she 'wished that there had been a category for creative use of language' in the rubric so voice could be better captured (Transcript R3:19).

The writer and reader presence in the conception of voice

Furthermore, raters also agreed that if they could clearly sense the author behind the writing, they would tend to hear a stronger voice. For example, when rating the one writing sample that was used across all the think-aloud sessions, all the raters at different points started to comment on the identity of the writer based on the writing itself. Reading into the last few paragraphs of the writing, Rater 4, for instance, drew a conclusion about who the writer was: 'This person is obviously a computer person' (Transcript R4:11). Similarly, Rater 2 also reflected on her guess of the writer's identity in the post-think-aloud interview: 'it's a computer person; this is someone who is well rounded. They've read, you know' (Transcript R2:19). And Rater 3 even went on to draw a conclusion about the gender of the writer: 'His examples . . . his or her examples . . . I guess it's a man because he'd been talking about technology . . . but that example of Bill Gates and Windows, like millions of lines of code, is really able to create a vivid image in my mind through writing. And, um, so I think for that reason, I would give it a high [voice] score' (Transcript R3:12). It seems, therefore, that the envisioning of a writer's identity while rating for voice strength is quite prevalent among the raters. In fact, this phenomenon is not exclusive to the voice raters here; it has also been observed and documented in other studies that examined composition rating sessions and blind manuscript review processes (e.g. Cumming, Kantor and Powers 2002, Matsuda and Tardy 2007). So, to some extent, the construction of an author's identity is probably an integral aspect of these evaluative processes, despite that in certain situations, especially in some high-stakes testing situations, we would hope our raters to focus entirely and exclusively on the appraisal of test takers' products and performances rather than on the test takers themselves.

In addition to their reflection on how the image of the writer as a unique person contributed to their evaluation of the overall voice strength in a written text, raters also commented on how this authorial self-revelation could be qualitatively different. The sharing of one's personal background in the writing, for example, varied in terms of quality, thus contributing to different levels of voice strength as perceived by the raters. As Rater 4 noted, 'there were some personal anecdotes that . . . were really engaging or genuine, but a lot of them felt kind of generic' (Transcript R4:16). Likewise, Rater 2 shared the following in the think-aloud session: 'when I think of voice, I think of . . .when I read this, would I be able to imagine an individual, versus something formulaic or trying really hard to answer the question' (Transcript R2:4).

Other than this authorial presence, raters also commented on how they thought the pulling in of the reader was integral to the realization of voice in writing. Rater 2, for example, shared in the interview that whether the writer was 'trying to be engaging' (Transcript R2:15) was an important aspect of voicing to her. Rater 4 further elaborated on what this meant to him.

> As a reader, it is feeling like the writer is there talking to me as an individual rather than sort of a generic, faceless person in the distance . . . like in one of those five-paragraph essays where, you know, it is just words being plugged into a blank something . . . [and] they are thinking about the reader, and they are thinking about what would be interesting to the reader and how to engage the reader (Transcript R4:19).

Similarly, Rater 1 also commented on how she viewed the authorial presence and reader engagement as inseparable.

> I can't disconnect [authorial presence] from reader engagement because I think authorial presence is almost like hosting a discussion, so the author has to be welcoming, and there's a whole range of things they can do to help welcome the other reader to be involved in their text (Transcript R1:2).

This rater also elaborated on what she meant by 'a whole range of things' that could help engage the reader. For example, she pointed out that the use of reader pronouns is 'a very subtle way [of engagement]; by constantly saying things like "you and me", [the reader became] part of the writer's group' (Transcript R1:2). The above data, therefore, indicate that writer and reader presence, as identified in the factor analysis, was indeed another dimension that contributed to the realization of voice in a written text.

The qualitative data hence supported the three-dimensional conceptualization of voice coming out of the quantitative data analysis. The fact that the raters in this study were able to identify more or less the same major components of voice also seemed to indicate that despite the elusiveness often associated with the concept of voice, it might still be a measurable construct. As the preliminary voice rubric only focused on evaluating the salience of the voice elements identified in Hyland's (2008) model, it only measured how frequently these elements were used in a writing sample. Raters, however, pointed out that voice strength not only had to do with how often these features were used, but also had to do with how they were used in writing. Rater 2 shared explicitly with the researcher her view toward the use of frequency counts as the only criterion for evaluating voice strength: 'I think there needs to be a combination, because if you look strictly by the numbers, you can't say more is better . . . Sometimes the numeric aspect of it . . . makes me want to say, yes, but there's so much more to it' (Transcript R2:19–20). Likewise, Rater 4 also commented, 'just because you do something more does not make it better' (Transcript R4:20–21). Thus, it seemed necessary and important that a qualitative component be added into the voice rubric so as to make it a valid measure of the elusive construct of voice. And the qualitative data gathered from raters in this study shed some light on how this could be done. Based on these quantitative and qualitative analysis results, therefore, the preliminary rubric underwent a few major revisions. The next section outlines these changes and presents the revised rubric in detail.

Rubric revision

The preliminary voice rubric was revised so that 1) the three voice elements (i.e. personal aside, reference to shared knowledge, and the use of rhetorical and audience-directed questions) that were rarely identified in the empirical data were removed from the rubric; 2) the rubric was re-organized to reflect the three major dimensions identified in the previous analysis; and 3) a qualitative piece was added to each dimension in the rubric to better capture voice strength. The descriptors that were used to define levels of voice strength in this qualitative part of the rubric were based on raters' think-aloud and interview data presented earlier. Using the *Rubric for Rubrics* (Educational Testing Service 2006, as cited in Arter and Chappuis 2007) as a quality control tool, detailed descriptions for Levels 1, 3, and 5 in the qualitative part of the revised voice rubric were then constructed. Although Levels 2 and 4 were not explicitly defined, raters were instructed to view these levels as having qualities that were in between the two adjacent levels that were explicitly defined. The viability of such an approach is supported by previous rubric construction research and practice (e.g. Arter and Chappuis 2007, Lim 2011, Lumley 2006, Northwest Regional Educational Laboratory 2010). Table 2.3 offers a snapshot of what the revised voice rubric looks like for one of its dimensions, and the complete revised version of the rubric is presented in Appendix B. As shown, the revised rubric now includes both a frequency-based evaluation of the salience of the voice elements and a qualitative evaluation of overall voice strength under each identified dimension. This revised voice rubric was then used to rate for voice strength in another set of 200 TOEFL$_{®}$ iBT writing samples. Voice ratings obtained from this phase of the study were used to further validate the analytic rubric. The next section reports on the validation results.

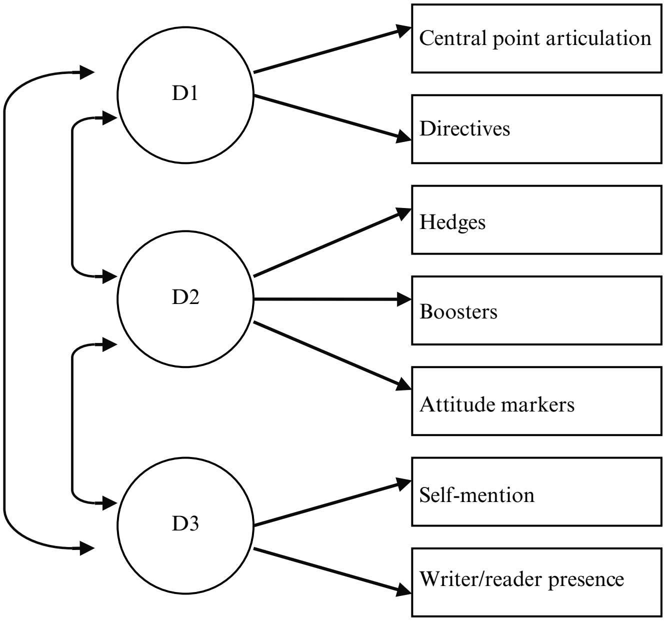
Table 2.3 Content-related dimension of the revised voice rubric

Dimension 1: Presence and clarity of ideas in the content			
C5: Central point articulation	C9: Directives	Overall voice evoked by the presence and clarity of ideas in the content	
A clear central point is articulated ____ times in the essay.	Directives are used ____ times in the essay.	5	-The reader feels a clear presence of a central idea (point of view) throughout the text. -The writing shows a strong commitment to the topic through full development of the central idea (point of view) with adequate use of effective examples and details. -The reader feels that he or she is being invited to participate in the discussion of the topic and the construction of an argument through the author's use of directive phrases when presenting ideas. -The idea (point of view) and the use of examples and details in the writing are unique, interesting, and engaging, indicating sophisticated thinking behind the writing.
		4	Not explicitly defined.
		3	-The reader feels that there is a central idea (point of view) in the text, but it is not fully developed. -The writing shows some commitment to the topic with proper use of some supporting examples and details. But the examples are not always appropriate or effective. -The reader occasionally feels that he or she is being invited to participate in the discussion of the topic; but more often, the reader feels a lack of interaction with the writer. -The idea (point of view) and the use of examples and details in the writing are safe and general, lacking uniqueness, sophistication, or thoughtfulness.
		2	Not explicitly defined.
		1	-The reader cannot find a consistent central idea (point of view) in the text. -The writing does not show any commitment to the topic; rather, it is only an attempt (or a failed attempt) to answer a question. No examples or details are used to develop the topic. -The reader feels that the writer is not concerned with the reader, and the writing is a confusing monologue instead of a clear dialogue between the writer and the reader. -The writing is generic and lifeless.

Phase 2: Validation of the revised voice rubric

After proper data screening and preparation, CFA was performed in EQS 6.1 for Windows (Bentler 1985) to test whether the model identified in Phase 1 held with Phase 2 data. The hypothesized model is presented in Figure 2.3. The three circles on the left represent the three identified voice dimensions: D1, presence and clarity of ideas; D2, manner of idea presentation; and D3, writer and reader presence. The rectangles on the right represent frequency-based ratings for the individual voice elements associated with each dimension. In the model, the three dimensions were also hypothesized to be correlated, as together they should tap into the construct of voice. It should also be noted that this round of data analysis involved only the frequency-based ratings of voice elements from 200 writing samples; qualitative dimensional ratings were not included.

Figure 2.3 Hypothesized CFA model



As Mardia's normalized coefficient was −4.56, indicating violations of multivariate normality, the Satorra-Bentler chi square robust estimation procedure was used (Satorra and Bentler 1988). An examination of the fit indices showed that they were only marginally acceptable – Satorra-Bentler χ^2 (12, N =200) = 21.96, p = 0.04, CFI = 0.75, RMSEA = 0.07, suggesting that the model did not fit the data well. An examination of the individual parameter estimates further revealed that none of the parameter estimates in the model were significant, whereas all the standard errors were. This again suggested that the hypothesized model did not fit the observed data well. Other CFA models were also tested. For example, the frequency-based ratings for the seven voice elements were tested to see if they together tap into one unified construct of voice directly. Results again showed very poor model fit.

Another round of analysis was thus performed to test whether the three qualitative ratings of dimensional voice strength were better able to measure the construct of voice. First, a CFA model was built to see if these three-dimensional ratings were tapping into the higher order construct of voice. This attempt, however, yielded a just-identified structural model, wherein 'the number of data variances and covariances equals the number of parameters to be estimated' (Byrne 2006:31). According to Byrne (2006), 'a just-identified model is not scientifically interesting because it has no degrees of freedom and, therefore, can never be rejected' (2006:31). PCA was performed in SPSS version 16.0 to explore if the qualitative voice dimensional ratings together were measuring the construct of voice.

The correlation matrix in Table 2.4 shows that all three-dimensional ratings are highly and significantly correlated, suggesting a close relationship between the three dimensions. The component matrix in Table 2.5, in addition, reveals that there is clearly only one component extracted. And the high factor loadings (all above or close to 0.9) also indicate that each of the three voice dimensions is highly correlated with this extracted component. As Tabachnick and Fidell (2007) point out: 'the greater the loading, the more the variable is a pure measure of the factor' (2007:649). And according to Comrey and Lee (1992), loadings above 0.71 are considered excellent. Thus, with the three loadings above or close to 0.9, the three voice dimensions all can be considered good measures of this extracted component, i.e. the construct of voice.

Table 2.4 PCA results: Correlation matrix

	D1	D2	D3
D1: Presence and clarity of ideas		0.77***	0.65***
D2: Manner of idea presentation			0.70***
D3: Writer and reader presence			

****Correlation is significant at the 0.0001 level (2-tailed)*

Table 2.5 PCA results: Component matrix

	Component 1
D1: Presence and clarity of ideas	0.90
D2: Manner of idea presentation	0.92
D3: Writer and reader presence	0.87

Table 2.6 shows the eigenvalues and proportions of variance explained by the components. Again, using eigenvalue > 1 (Kaiser 1960) as a criterion for selecting principal components, it is clear that only one component emerged (with an eigenvalue of 2.42). And this one component is able to account for 81% of the total variance in qualitative dimensional voice ratings, suggesting again that the three qualitative dimensions together do tap into the underlying construct of voice.

Table 2.6 PCA results: Eigenvalues and proportions of variance explained by the components

Component	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	2.42	80.63	80.63	2.42	80.63	80.63
2	0.36	11.92	92.55			
3	0.22	7.45	100.00			

To further test whether the three-dimensional ratings truly outperform the frequency counts as a measure of voice strength in written texts, regression analyses were also conducted. First, the three-dimensional voice ratings were used to predict the holistic and impressionistic ratings of overall voice strength. Results showed that the model was significant ($F = 593.52$, $df = 3{,}196$, $p < 0.0001$), and the three-dimensional ratings were able to account for about 90% ($R^2 = 0.90$) of the variance in raters' holistic assessment of overall voice strength in the TOEFL$_®$ writing samples. Moreover, the three-dimensional ratings in the model were all statistically significant predictors of the holistic ratings (see Appendix C for details).

Conversely, when using the seven frequency-based voice element ratings to predict holistic overall voice ratings, the model was able to explain only 35% ($R^2 = 0.35$) of the total variance in the raters' holistic evaluation of voice strength. Of the seven voice elements, only C5 central point articulation, C2 use of boosters, and C6 reader reference were statistically significant predictors of the holistic voice ratings (see Appendix C for details). Such results suggest that while the frequency of certain voice elements does influence raters' perceptions of voice strength, it alone only captures a small fraction of what raters see as voice, hence falling short of being a good measure of this construct.

Results from factor analysis and regression analysis, therefore, all confirm that the frequency-based ratings for the seven voice elements do not measure voice as well as do the voice dimensional ratings. This finding suggests again that a rater's perception or feeling of overall voice strength in a writing sample is less related to how many times individual voice elements are used but more related to how they are used. Future users of the voice rubric should bear in mind that the inclusion of frequency counts in the revised voice rubric is for validation purposes and does not entail that the assessment of voice strength must be based on the quantification of individual voice elements. In fact, validation results here already show that these linguistic elements in the rubric can add voice to a written piece only when they are used properly. In other words, instead of counting the frequency of discrete voice elements, looking at how these elements are used qualitatively better captures voice in written discourse.

Discussion

In this study, Hyland's (2008) interactional model of voice was used as the theoretical basis for the development of an analytic voice rubric. Although a rubric developed on the basis of a comprehensive theoretical model can be more generally applicable than one that is purely empirically derived (e.g. Brindley 1998, Knoch 2011, Shohamy 1996, Turner and Upshur 1996, Upshur and Turner 1999), it may lack content relevance when used in a particular performance assessment setting. As shown in this study, while the model identifies as many as nine different linguistic- and discourse-level elements that add voice to a written text, not all of them were sufficiently instantiated in the TOEFL$_{®}$ writing samples. The lack of close alignment between what is proposed in the theoretical model and what is observed in the empirical data here in this study, however, deserves our careful interpretation.

We should note that Hyland's model was developed based on his examination of voice-related language features in a large corpus of published academic articles, whereas data used in this study were short writing samples produced by TOEFL$_{®}$ iBT test takers within a limited time and in response to a single prompt. Published academic texts are almost always evidence-based analytic writing that involves the use of various source texts and the writer's interpretation of and conversation with these source texts. This characteristic inevitably demands more interactions between the writer and the source texts (or, in a sense, between the writer and the intended readers, as any academic text published in a particular field is in essence a conversation with the members of that particular community). The TOEFL$_{®}$ iBT independent writing tasks, however, are timed and opinion-based types of writing that require no use of source texts. The difference in the nature and purpose of the writing, the length of the written products, as well as the writers' level of language proficiency, thus explain why certain voice elements that signal more sophisticated interactions between a writer and a reader (i.e. personal asides, reference to shared knowledge, and rhetorical or audience-directed questions) were rarely observed in the data used in this study.

There are a number of implications based on these results. First, as much as we want a general measure of voice that could be used in different assessment or educational settings, the evaluation of voice is not context-independent. Rather, we need to take into consideration the characteristics of the writing task, the genre and level of the writing, and the audience for whom the writing is intended. In different contexts, the realization of voice may be different; therefore, the criteria used to evaluate voice may vary. In evidence-based research writing for an academic discourse community, for example, the use of hedging devices is believed to be a key feature of academic texts (e.g. Hinkel 2002, 2004, Hyland 1994, 1998, 2000a, 2000b), especially when the writing includes interpretations of data and research results. In fact, according to Hyland's (2008) corpus analysis, 'hedges [are] the most frequent feature of writer perspective in the corpus, reflecting the critical importance of distinguishing fact from opinion and the need for writers to present their claims with appropriate caution and regard to colleagues' views' (2008:12–13). Hence, the use of hedges in this context, although toning down the force of an argument/claim, still positively contributes to the realization of an authorial voice as it 'function[s] to convey deference to one's readers and/or an openness to alternative viewpoints' (Tang 2006:78). On the other hand, in the context of the TOEFL$_{®}$ independent writing assessment that elicits L2 writers' opinion-based arguments in response to a single prompt, voice raters tended to associate the use of hedges in these writing samples with a lack of confidence in the L2 writer, or a lack of a clear stance on a particular topic under discussion. It follows, therefore, that voice raters in this study favored definitive language more as a way of realizing voice and demonstrating writer confidence.

A second implication, based on our observation that TOEFL$_{®}$ test takers used less sophisticated voicing strategies when compared to advanced writers of academic texts, is that L2 writers' ability to employ voice-related features in their writing probably depends on their general language proficiency, particularly their facility with English syntax. The personal aside category, for example, is defined as insertions by the author that often appear in the middle of a statement as personal comments or reflections. As these insertions briefly interrupt the flow of an argument, they also tend to engage the reader by initiating some kind of dialogue between the author and the reader. Below is an example of the use of personal asides (*italicized*) in one of the TOEFL$_{®}$ writing samples.

> Going back to the example of the artist, art as a hobby is, *in my opinion*, much more enjoyable as a hobby without the pressure to sell. Should the hobby artist come to fame and start to make a fortune (*which is not very likely as we all know*), a switch to a full-time art career can be made at that time without the first frugal years.

Obviously, proper use of such language features in writing requires relatively more advanced syntactic knowledge in the first place. Not surprisingly, therefore, while personal asides are frequently employed by advanced writers in published academic articles, they were rarely identified in writing samples produced by TOEFL® test takers.

Thus, a potentially useful way to address the concept of voice in either writing assessment or writing instruction may be to take a developmental approach. At different L2 proficiency levels, student writers could be taught to use different voicing strategies appropriate at that level, moving from the simplest and most straightforward ways of voicing, such as the use of first person pronouns, gradually to more complex and sophisticated ways such as the use of personal asides. Although the concept of voice is complex and, hence, difficult to teach, especially to novice L2 writers, using such a developmental approach, writing instructors could break the concept down to its components in their instruction, and choose to present the more accessible elements first before moving on to more difficult ones that require higher levels of L2 proficiency. Of course, instruction of voice should not dwell on how frequently these features should be used in a piece of writing. Rather, teachers should always explain how these features could be used to achieve a strong and effective voice in a particular writing context, considering the nature, purpose, and audience of the writing.

Another implication for voice assessment and instruction stems from the observed relationship between the holistic voice ratings and the three qualitative voice dimensional ratings. Correlational and regression analyses between holistic and analytic voice ratings show that about 90% of the variance in the holistic ratings can be explained by the three qualitative dimensional ratings of voice. Such a result suggests that, despite its elusive nature, voice can still somehow be broken down to smaller components that are relatively more concrete and discrete, hence easier to understand. In the rater interview sessions of this study, when asked whether voice is something that could only be looked at holistically or something that could be broken down to different components, three of the four raters reported that before they participated in this study, they all would ‘tend to look at it holistically’ (Transcript R3:18), based on their own ‘feelings’ (Transcript R4:19) or ‘personal definition of voice’ (Transcript R2:18). But these personal feelings and beliefs are often too abstract and unstable to be articulated clearly, which probably contributes to the notorious elusiveness of this concept and the difficulty teaching or assessing it. The rater interview data from this study do suggest that raters are quite aware of this problem:

> I take the piece more holistically in terms of “can I imagine this person?” . . . [But] I feel like my personal definition of voice is so ephemeral that, you know, if you had asked me how I would measure it, I don’t even think I would know, necessarily, in a rubric form (Transcript R2:18).

The rating experience, however, made them ‘realize that there were some things that make [voice] up’ (Transcript R4:19). Rater 2’s reflection on her view toward the concept of voice before and after the rating sessions is representative of what other raters have shared.

> If you had asked me before doing this [rating voice strength using the preliminary voice rubric], I would have said, well, that I think that you need to look at [voice] holistically. Trying to look at it by component has been a very interesting experience and I do think that there are certain linguistic markers that help (Transcript R2:19).

Hence, the presence of an analytic rubric does help make the seemingly intangible and mysterious concept of voice more accessible, even to the experienced writing teachers themselves. The fact that the analytic dimensional voice ratings accounted for about 90% of raters' holistic and impressionistic evaluation of overall voice strength further suggests that these three dimensions in the rubric capture voice well.

Some may argue that if the holistic measure of voice does equally well, if not better, in terms of capturing an authorial voice in written texts, it is probably unnecessary, and meaningless, to develop an analytic voice rubric. Nevertheless, holistic and analytic rubrics serve different assessment purposes. An analytic voice rubric, with its identified voice dimensions and elements, could better help writing instructors and L2 writers explain and learn this complex concept. It could also be of use when writing instructors want to diagnose the kinds of difficulties students face when learning to write with an effective authorial voice. Breaking the concept down to smaller components in an analytic rubric is, therefore, pedagogically more useful than a vague, intuition-based holistic rubric. Of course, in other assessment settings where the purpose of assessment is not to inform pedagogy or diagnose strength and weaknesses, but to evaluate a piece of writing in a summative manner, a holistic impression-based measure of voice may not only serve the purpose well but also outperform an analytic rubric (as it is less labor-intensive when compared to analytic scoring), especially when there is evidence that the holistic and the analytic voice ratings are highly correlated and share a large amount of variance between them.

Finally, the voice dimensions identified in this study also provide a plausible alternative conception of voice. Despite the fact that many of the individual voice elements included in the analytic voice rubric also appeared in Helms-Park and Stapleton's (2003) Voice Intensity Rating Scale, their rubric focuses on the authorial stance aspect of voicing and fails to include the reader engagement aspect. As Hyland (2008) argues, and our rater think-aloud and interview data confirm, writer stance and reader engagement are essentially 'two sides of the same coin' (2008:8). Helms-Park and Stapleton's (2003) voice rubric, thus, only captures one of these two sides. Moreover, although Hyland's (2008) model of voice is comprehensive, it fails to address more global features of a text such as the presence, clarity, and uniqueness of a central point being communicated to the readers. Qualitative data analysis in this study, however, shows that this content-related aspect of voice could be the most important part of the voicing strategy, at least in argumentative writing. The three-dimensional voice rubric coming out of this study, therefore, incorporates all these critical components of voice. Its first dimension examines the presence, clarity, and uniqueness of a personal idea that demonstrates an author's commitment to a particular stance. The second dimension then looks at the manner in which the author presents his or her idea – assertively, mildly, confidently, tentatively, enthusiastically, or maybe indifferently. The third dimension focuses more on authorial presence and reader presence, examining the extent to which an author reveals him- or herself explicitly and pulls the readers into the construction of an argument or opinion. With these identified dimensions and the qualitative descriptors that define them, writing teachers and L2 writers will be better able to explain and understand this slippery concept that is thought by many to be unlearnable and unteachable to L2 student writers.

Conclusion

As a formal attempt to capture and define voice, this study developed and validated an analytic rubric that measures the strength of an authorial voice in L2 argumentative writing. The presence of such an instrument will help contribute to future research on voice-related issues in the fields of L2 writing, L2 writing assessment, and writing in general. It could, for example, allow for more empirical investigations of the relative importance of voice in writing instruction and assessment. As seen in the literature, writing researchers have long been debating about the importance, or lack thereof, of voice in writing instruction, especially in L2 writing instruction; yet, as Helms-Park and Stapleton (2003) also observe, there is little, if any, research that has empirically examined the relationship between voice and writing quality, possibly due to the elusive nature of the concept itself and the absence of an instrument that captures voice in written discourse. Consequently, no empirical evidence has been offered to confirm or refute the teaching of voice as a worthy pedagogical focus in L2 writing classrooms, despite the considerable debate at the theoretical level in the literature.

The presence of this analytic voice rubric can, therefore, enable researchers to conduct empirical investigations to better understand what voice is and what it does in argumentative writing. Only with such empirical evidence can we determine, with more confidence, the relative importance of voice in L2 writing instruction and assessment. Additionally, the presence of this rubric can demystify, even if only to some extent, the seemingly unlearnable concept of voice for both L2 writing teachers and novice L2 writers. By using such a rubric, writing instructors can better help their students, especially L2 writers, to write with a strong authorial voice and use that voice appropriately and effectively in their writing.

Acknowledgements

The preparation of this manuscript was sponsored by the Shanghai Pujiang Program (11PJC095) and an internal research grant from Shanghai International Studies University (HX181091). The study was also supported by The International Research Foundation for English Language Education (TIRF) Doctoral Dissertation Grant (DDG), a TOEFL® Small Grant for Doctoral Research in Second or Foreign Language Assessment from Educational Testing Service, and a Dean's Grant for Student Research from New York University's Steinhardt School of Culture, Education, and Human Development. The author wishes to thank the participants in this study and Dr Lorena Llosa.

References

Arter, J A and Chappuis, J (2007) *Creating and Recognizing Quality Rubrics*, Upper Saddle River: Pearson Education Inc.

Bentler, P M (1985) *EQS for Windows*, Encino: Multivariate Software.

Bowden, D (1999) *The Mythology of Voice*, Portsmouth: Heinemann.

Brindley, G (1998) Describing language development? Rating scales and second language acquisition, in Bachman, L F and Cohen, A D (Eds) *Interfaces between SLA and Language Testing Research*, Cambridge: Cambridge University Press, 112–114.

Byrne, B M (2006) *Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming*, Mahwah: Lawrence Erlbaum.

Comrey, A L and Lee, H B (1992) *A First Course in Factor Analysis*, Hillsdale: Lawrence Erlbaum.

Cumming, A, Kantor, R and Powers, D E (2002) Decision making while rating ESL/EFL writing tasks: A descriptive framework, *The Modern Language Journal* 86, 67–96.

DeRemer, M (1998) Writing assessment: Raters' elaboration of the rating task, *Assessing Writing* 5 (1), 7–29.

DiPardo, A, Storms, B A and Selland, M (2011) Seeing voices: Assessing writerly stance in the NWP Analytic Writing Continuum, *Assessing Writing* 16 (3), 170–188.

Ede, L (1992) *Work in Progress: A Guide to Writing and Revising* (2nd edition), New York: St Martin's Press.

Elbow, P (1968) A method for teaching writing, *College English* 30 (2), 115–125.

Elbow, P (1981) *Writing with Power*, New York: Oxford University Press.

Elbow, P (1994) What do we mean when we talk about voice in texts? in Yancey, K B (Ed) *Voices on Voice: Perspectives, Definitions, Inquiry*, Urbana: National Council of Teachers of English, 1–35.

Elbow, P (1999) Individualism and the teaching of writing: Response to Vai Ramanathan and Dwight Atkinson, *Journal of Second Language Writing* 8 (3), 327–338.

Farmer, F (1995) Voice reprised: Three études for a dialogic understanding, *Rhetoric Review* 13 (2), 301–320.

Graves, D H (1983) *Writing: Teachers & Children at Work*, Portsmouth: Heinemann.

Graves, D H (1994) *A Fresh Look at Writing*, Portsmouth: Heinemann.

Helms-Park, R and Stapleton, P (2003) Questioning the importance of individualized voice in undergraduate L2 argumentative writing: An empirical study with pedagogical implications, *Journal of Second Language Writing* 12 (3), 245–265.

Hinkel, E (2002) *Second Language Writers' Text: Linguistic and Rhetorical Features*, Mahwah: Lawrence Erlbaum.

Hinkel, E (2004) *Teaching Academic ESL Writing: Practical Techniques in Vocabulary and Grammar*, Mahwah: Lawrence Erlbaum.

Hirvela, A and Belcher, D (2001) Coming back to voice: The multiple voices and identities of mature multilingual writers, *Journal of Second Language Writing* 10 (1–2), 83–106.

Holding, M (2005) Liberating the student's voice: A teacher's story of the college essay, *English Journal* 94 (4), 76–82.

Hyland, K (1994) Hedging in academic writing and EAP textbooks, *English for Specific Purposes* 13 (3), 239–256.

Hyland, K (1998) *Hedging in Scientific Research Articles*, Amsterdam: John Benjamins Publishing Company.

Hyland, K (2000a) Hedges, boosters and lexical invisibility: Noticing modifiers in academic texts, *Language Awareness* 9 (4), 179–301.

Hyland, K (2000b) 'It might be suggested that. . .': Academic hedging and student writing, *Australian Review of Applied Linguistics* 16, 83–97.

Hyland, K (2008) Disciplinary voices: Interactions in research writing, *English Text Construction* 1 (1), 5–22.

Ivanič, R and Camps, D (2001) I am how I sound: Voice as self-representation in L2 writing, *Journal of Second Language Writing* 10 (1–2), 3–33.

Jeffery, J (2007) *Discourses of writing in high-stakes direct writing assessments*, paper presented at the National Reading Conference, Austin, November 2007.

Jeffery, J (2009) Constructs of writing proficiency in US state and national writing assessments: Exploring variability, *Assessing Writing* 14 (1), 3–24.

Kaiser, H F (1960) The application of electronic computers to factor analysis, *Educational and Psychological Measurement* 20, 141–151.

Knoch, U (2011) Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing* 16 (2), 81–96.

Kumamoto, C D (2002) Bakhtin's others and writing as bearing witness to the eloquent 'I', *College Composition and Communication* 54 (1), 66–87.

Lim, G S (2011) *Meeting multiple validation requirements in rating scale development*, paper presented at the 2011 Language Testing Research Colloquium, Ann Arbor, June 2011.

Llosa, L, Beck, S W and Zhao, C G (2011) An investigation of academic writing in secondary schools to inform the development of diagnostic classroom assessments, *Assessing Writing* 16, 256–273.

Lumley, T (2006) *Assessing Second Language Writing: The Rater's Perspective*, Frankfurt: Peter Lang.

Matsuda, P K (2001) Voice in Japanese written discourse: Implications for second language writing, *Journal of Second Language Writing* 10 (1–2), 35–53.

Matsuda, P K and Tardy, C M (2007) Voice in academic writing: The rhetorical construction of author identity in blind manuscript review, *English for Specific Purposes* 26, 235–249.

Matsuda, P K and Tardy, C M (2008) Continuing the conversation on voice in academic writing, *English for Specific Purposes* 27, 100–105.

New York State Education Department (2007) *Comprehensive Examination in English: Scoring Key and Rating Guide*, available online: www.nysedregents.org/ComprehensiveEnglish/Archive/20070123scoringkey1.pdf

Northwest Regional Educational Laboratory (2010) *6+1 Trait® Rubrics*, available online: educationnorthwest.org/resource/464

O'Leary, M E (1993) *A voice of one's own: Born, achieved, or thrust upon one?* available online: archive.org/stream/ERIC_ED360633/ERIC_ED360633_djvu.txt

Prior, P (2001) Voice in text, mind, and society: Sociohistoric accounts of discourse acquisition and use, *Journal of Second Language Writing* 10 (1–2), 55–81.

Ramanathan, V and Kaplan, R B (1996) Audience and voice in current L1 composition texts: Some implications for ESL student writers, *Journal of Second Language Writing* 5 (1), 21–34.

Reid, J (2001) Advanced EAP writing and curriculum design: What do we need to know? in Silva, T and Matsuda, P K (Eds) *On Second Language Writing*, Mahwah: Lawrence Erlbaum, 143–160.

Satorra, A and Bentler, P M (1988) Scaling corrections for chi-square statistics in covariance structure analysis, in American Statistical Association (Eds) *1988 Proceedings of the Business and Economic Statistics Section*, Alexandria: American Statistical Association, 308–313.

Shohamy, E (1996) Competence and performance in language testing, in Brown, G, Malmkaer, K and Williams, J (Eds) *Performance and Competence in Second Language Acquisition*, Cambridge: Cambridge University Press, 138–151.

Stapleton, P (2002) Critiquing voice as a viable pedagogical tool in L2 writing: Returning spotlight to ideas, *Journal of Second Language Writing* 11 (3), 177–190.

Stewart, D C (1969) Prose with integrity: A primary objective, *College Composition and Communication* 20 (3), 223–227.

Stewart, D C (1972) *The Authentic Voice: A Pre-writing Approach to Student Writing*, Dubuque: Brown.

Stewart, D C (1992) Cognitive psychologists, social constructionists, and three nineteenth-century advocates of authentic voice, *Journal of Advanced Composition* 12 (2), 279–290.

Tabachnick, B G and Fidell, L S (2007) *Using Multivariate Statistics* (5th edition), Boston: Allyn and Bacon.

Tang, R (2006) Addressing self-representation in academic writing in a beginners' EAP classroom, *Journal of Language and Learning* 5 (2), 76–85.

Turner, C E and Upshur, J A (1996) Developing rating scales for the assessment of second language performance, in Wigglesworth, G and Elder, C (Eds) *The Language Testing Cycle: From Inceptions to Washback*, Australian Review of Applied Linguistics Series S 13, Melbourne: Australian Review of Applied Linguistics, 55–79.

Upshur, J A and Turner, C E (1999) Systematic effects in the rating of second language speaking ability: Test method and learner discourse, *Language Testing*, 16 (1), 82–111.

Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.

Wolfe, E W, Bolton, S, Feltovich, B and Niday, D M (1996) The influence of student experience with word processors on the quality of essays written for a direct writing assessment, *Assessing Writing* 3 (2), 123–147.

Yancey, K B (1994) Introduction: Definition, intersection, and difference – mapping the landscape of voice, in Yancey, K B (Ed) *Voices on Voice: Perspectives, Definitions, Inquiry*, Urbana: National Council of Teachers of English, vii–xxiv.

Yeh, S S (1998) Validation of a scheme for assessing argumentative writing of middle school students, *Assessing Writing* 5 (1), 123–150.

Zhao, C G and Llosa, L (2008) Voice in high-stakes L1 academic writing assessment: Implications for L2 writing instruction, *Assessing Writing* 13 (3), 153–170.

Appendix A

The preliminary voice rubric

	C1: Use of hedges	C2: Use of boosters	C3: Use of attitude markers	C4: Authorial self-mention	C5: Articulation of the central point
4	Hedges are used always or almost always in the author's claim. 8↑	Boosters are used always or almost always in the author's claim. 8↑	Attitude markers are used always or almost always in the author's claim. 8↑	First person pronouns are used extensively. 8↑	A clear central point is articulated more than three times in the essay. 4↑
3	Hedges are used frequently in the author's claims. 5–7	Boosters are used frequently in the author's claims. 5–7	Attitude markers are used frequently in the author's claims. 5–7	First person pronouns are used frequently. 5–7	A clear central point is articulated three times in the essay. 3
2	Hedges are used occasionally in the author's claims. 3–4	Boosters are used occasionally in the author's claims. 3–4	Attitude markers are used occasionally in the author's claims. 3–4	First person pronouns are used occasionally. 3–4	A clear central point is articulated twice in the essay. 2
1	Hedges are rarely used in the author's claims. 1–2	Boosters are rarely used in the author's claims. 1–2	Attitude markers are rarely used in the author's claims. 1–2	First person pronouns are rarely used. 1–2	A clear central point is articulated once in the essay. 1
0	Hedges are never used in the author's claims. 0	Boosters are never used in the author's claims. 0	Attitude markers are never used in the author's claims. 0	First person pronoun is never used. 0	A clear central point is not present. 0

	C6: Use of reader pronouns	C7: Use of personal asides	C8: References to shared knowledge	C9: Use of directives/ reader signposts	C10: Use of rhetorical/ audience-directed questions
4	Reader pronouns are used extensively. 8↑	Personal asides are used extensively. 6↑	There are extensive references to shared knowledge. 6↑	Directives are used extensively. 6↑	Rhetorical or audience-directed questions are used extensively. 6↑
3	Reader pronouns are used frequently. 5–7	Personal asides are used frequently. 4–5	There are frequent references to shared knowledge. 4–5	Directives are used frequently. 4–5	Rhetorical or audience-directed questions are used frequently. 4–5
2	Reader pronouns are used occasionally. 3–4	Personal asides are used occasionally. 2–3	There are occasional references to shared knowledge. 2–3	Directives are used occasionally. 2–3	Rhetorical or audience-directed questions are used occasionally. 2–3
1	Reader pronouns are rarely used. 1–2	Personal asides are rarely used. 1	There are a few references to shared knowledge. 1	Directives are rarely used. 1	Rhetorical or audience-directed questions are rarely used. 1
0	Reader pronouns are rarely used. 0	Personal asides are never used. 0	There are no references to shared knowledge. 0	Directives are never used. 0	Rhetorical or audience-directed questions are never used. 0

	C11: Overall authorial presence and reader engagement
4	The reader feels a strong sense of authorial presence in the writing.
3	The reader feels a fairly strong sense of authorial presence in the writing.
2	The reader feels a somewhat weak sense of authorial presence in the writing.
1	The reader feels a very weak sense of authorial presence in the writing.
0	The reader feels no sense of authorial presence in the writing.

Note: Frequency ranges to use as an indicator of the level of salience of these voice features are given with each element of the rubric; ↑= And above

Appendix B

The revised voice rubric

Dimension 1: Presence and clarity of ideas in the content

C5: Central point articulation	C9: Directives	Overall voice evoked by the presence and clarity of ideas in the content	
A clear central point is articulated ____ times in the essay.	Directives are used ____ times in the essay.	5	– The reader feels a clear presence of a central idea (point of view) throughout the text. – The writing shows a strong commitment to the topic through full development of the central idea (point of view) with adequate use of effective examples and details. – The reader feels that he or she is being invited to participate in the discussion of the topic and the construction of an argument through the author's use of directive phrases when presenting ideas. – The idea (point of view) and the use of examples and details in the writing are unique, interesting, and engaging, indicating sophisticated thinking behind the writing.
		4	– Not explicitly defined.
		3	– The reader feels that there is a central idea (point of view) in the text, but it is not fully developed. – The writing shows some commitment to the topic with proper use of some supporting examples and details. But the examples are not always appropriate or effective. – The reader occasionally feels that he or she is being invited to participate in the discussion of the topic; but more often, the reader feels a lack of interaction with the writer. – The idea (point of view) and the use of examples and details in the writing are safe and general, lacking uniqueness, sophistication, or thoughtfulness.
		2	– Not explicitly defined.

			1	– The reader cannot find a consistent central idea (point of view) in the text. – The writing does not show any commitment to the topic; rather, it is only an attempt (or a failed attempt) to answer a question. No examples or details are used to develop the topic. – The reader feels that the writer is not concerned with the reader, and the writing is a confusing monologue instead of a clear dialogue between the writer and the reader. – The writing is generic and lifeless.

Dimension 2: Manner of idea presentation

C1: Hedges	C2: Boosters	C3: Attitude markers	Overall voice evoked by the manner of presentation	
Hedges are used ____ times in the essay.	Boosters are used ____ times in the essay.	Attitude markers are used ____ times in the essay.	5	– The writer presents ideas and claims with language that shows authority and confidence. – The reader feels that the writer has a clear stance on and a strong attitude toward the topic under discussion. – The tone of the writing shows personality, adds life to the writing, and is engaging and appropriate for the intended reader. – Word choice, and language use by extension, is varied, often interesting, sophisticated, and eye-catching to the reader.
			4	– Not explicitly defined.
			3	– The writer presents ideas and claims somewhat mildly with frequent use of unnecessary hedges; only occasionally does the writing show some degree of authority and confidence. – The writer seems to have a stance on the topic under discussion, but no strong attitude is revealed in the writing. – The tone of the writing is appropriate for the intended reader and the purpose of the writing, but lacks personality and liveliness. – Occasional interesting word choice and language use may catch the reader’s attention, but the effect is inconsistent.
			2	– Not explicitly defined.

Dimension 2: Manner of idea presentation			
		1	– The writer presents ideas and claims very mildly, showing a lack of authority and confidence in what he/she is writing. – The writer seems indifferent and does not have a clear stance on or attitude toward the topic under discussion. – The writer writes in a monotone that does not engage the reader at all; oftentimes the reader finds him- or herself drifting off while reading the text. – Word choice or language use is flat, general, and dull, and thus unable to catch the reader's attention.

Dimension 3: Writer and reader presence			
C4: Authorial self-mention	**C6: Reader reference**	**Overall voice evoked by writer and reader presence**	
First person pronouns are used ____ times in the essay.	Reader pronouns are used ____ times in the essay.	5	– The writer reveals him- or herself in the writing either directly or indirectly, giving the reader a clear sense of who the writer is as a unique individual. – The reader feels that the writer is aware of and able to engage the reader effectively in a direct or subtle way. – The sharing of personal backgrounds and experiences, if any, is effective, genuine, and engaging to the reader.
		4	– Not explicitly defined.
		3	– The writer reveals him- or herself in the writing to some extent, leaving the reader with some sense of who he/she is. – The reader feels that the writer is aware of and trying to engage the reader in a way, but with limited success. – The sharing of personal backgrounds and experiences, if any, is genuine but not so engaging or effective to the reader.
		2	– Not explicitly defined.
		1	– The reader has little or no sense of who the writer is as a unique individual instead of a generic, faceless person. – The reader feels that the writer is not concerned with the reader or completely fails to engage the reader in any way. – The sharing of personal backgrounds and experiences, if any, is generic, ineffective, and even inappropriate, making the reader feel annoyed.

Overall voice strength	
5	The reader feels a very strong authorial voice in the writing.
4	The reader feels a fairly strong authorial voice in the writing.
3	The reader feels a somewhat weak authorial voice in the writing.
2	The reader feels a very weak authorial voice in the writing.
1	The reader cannot really feel the presence of an authorial voice in the writing.

Appendix C

Comparison of regression models predicting holistic overall voice ratings (N = 200)

	Model 1	**Model 2**
Intercept	−0.06 (0.08) −0.80	0.58 (0.28) 2.11*
C5: Central point		0.17 (0.04) 3.81***
C9: Directives		0.09 (0.04) 1.98
C1: Hedges		0.06 (0.05) 1.32
C2: Boosters		0.23 (0.05) 4.96***
C3: Attitude markers		0.05 (0.05) 1.05
C4: Self-mention		0.06 (0.05) 1.35
C6: Reader reference		0.19 (0.04) 4.36***
D1: Content-related	0.26 (0.04) 7.01***	
D2: Manner-related	0.42 (0.04) 9.97***	
D3: Reader and writer presence	0.40 (0.03) 12.69***	
Overall model fit	$R^2 = 0.90$ $F = 593.52$ *(df)* = (3, 196) $p < 0.0001$	$R^2 = 0.35$ $F = 14.86$ *(df)* = (7, 192) $p < 0.0001$

Note: Cell entries are estimated regression coefficients, (standard errors), and t-statistics.

**p<0.05, ***p<0.001*

3 Use of an automated essay scoring system in a multi-draft ESL writing class

Semire Dikli

Georgia Gwinnett College, Georgia, US

Motivation for the research

As an English as a Second Language (ESL) instructor, I have often struggled with finding time to provide written feedback to my students on essays that they write in my multi-draft writing classes. This struggle becomes even more challenging when I teach multiple-level classes or several sections of a writing class with large numbers of students. In the past, when I taught English as a Foreign Language (EFL), I also observed that EFL instructors were hesitant about grading students' writing assignments. In addition, writing is often not included in standardized English proficiency exams, so EFL instructors and students pay less attention to writing skills. Thus, whenever instructors experience a 'time crunch' in terms of being able to cover all of the material in the course syllabus, the writing sections in the textbooks seem to be the first ones to be omitted. Seeing these instructional practices play out over time motivated me to look for ways to facilitate the feedback process in writing classes. I wondered if advancements in technology could be used to assist teachers in responding to ESL and EFL students' compositions. Could computers aid ESL and EFL instructors in scoring essays and reduce their workload in grading essays?

There are several automated essay scoring (AES) systems on the market. These programs are able to score student essays and provide feedback in a matter of seconds. Some companies claim that their programs are in '98% agreement with human raters'. One of the developers of an AES system called PEG™ (Project Essay Grader), summarized the purpose of the system that he developed as 'concerned with improving writing practice and performance', and stated that 'PEG™ research has concentrated on how to improve student writing and simultaneously relieve the pressure of the extra work for teachers to grade such work' (Page 2003:48). For second language (L2) writing teachers who spend hours and hours each week responding to students' writing, this claim seems to be almost too good to be true. Is AES able to score students' writing as claimed? Could AES assist ESL/EFL writers and teachers? The present study, therefore, aims to explore whether an AES system can be used to facilitate the revision process in a multi-draft ESL writing classroom. It also investigated the extent to which the AES system was able to provide the same kinds of responses and feedback to writers as their teachers.

Automated essay scoring research

AES can be defined as a computerized system that evaluates and scores essays automatically within seconds. AES programs rely on various machine-learning methods including artificial intelligence (AI), natural language processing (NLP), and latent semantic analysis (LSA) to provide instant feedback and scoring. AI focuses on designing intelligent machines that have the ability to imitate the human mind, and NLP is one application of AI, which has been used to summarize texts and translate them into different languages for decades. Furthermore, LSA describes a word used in a sentence, passage, or essay based on the semantic associations (Landauder, Foltz and Laham 1998). There are many AES systems on the market, but the most popular ones are PEG™, Intelligent Essay Assessor™ (IEA) and its instructional application WriteToLearn, e-rater® and its instructional application Criterion®, and IntelliMetric™ and its instructional application MY Access!®. These AES systems are widely used by testing companies and educational institutions (Dikli 2006).

The developers and designers of AES systems claim that the systems have benefits, such as being a time and money saver in the administration of large-scale assessments (Shermis and Burstein (Eds) 2003), thereby reducing the teacher's work load (Myers 2003) and addressing issues related to subjectivity or teacher bias with assessment (Myers 2003). On the other hand, AES has been criticized for its inability to carry on meaningful interactions between the reader and writer as human scorers often do (Hamp-Lyons 2001), not assessing an essay as human raters do (Page 2003), and failing to count variables that might not be 'truly' important in essay grading (i.e. focusing on formal aspects of writing rather than organizational features) (Chung and O'Neil 1997, Page 2003).

Despite the criticisms, AES systems are extensively used by testing companies, universities, and public schools for large-scale high-stake assessment purposes because the AES scores are highly correlated with human scores (Attali 2004, Burstein and Chodorow 1999, Landauer, Laham and Foltz 2003, Landauer, Laham, Rehder and Schreiner 1997, Nichols 2004, Page 2003, Vantage Learning 2003b). While most of the studies to date have focused on the accuracy and reliability of the AES systems (Attali 2004, Burstein and Chodorow 1999, Deane 2013, Landauer et al 2003, Landauer et al 1997, Lee, Gentile and Kantor 2008, Nichols 2004, Page 2003, Ramineni 2013, Ramineni and Williamson 2013, Vantage Learning 2000a, 2000b, 2001b, 2002, 2003a, 2003b) limited research has been done in terms of the use of AES systems in classroom contexts (Attali, 2004, Chen and Cheng 2008, Dikli 2010, Dikli and Bleyle 2014, Elliot and Mikulas 2004, Grimes and Warschauer 2008, Grimes and Warschauer 2010, Vantage Learning 2004).

Furthermore, the majority of the research on AES focuses on native English-speaking (L1) writers, and only a small number of studies have been conducted in ESL/EFL contexts (Burstein and Chodorow 1999, Chen and Cheng 2008, Chodorow and Burstein 2004, Dikli 2010, Dikli and Bleyle 2014, Edelblut and Vantage Learning 2003, Vantage Learning 2001a). Developing companies have made some attempts to build models that involve languages other than English (Shermis and Burstein (Eds) 2003). For instance, Educational Testing Service (ETS) has been working on 'computer-based corpora' that characterize language variation, both for subgroups who use non-standard dialects of English and non-native speakers of English (Burstein and Chodorow 1999). Additionally, Vantage learning has developed programs that could provide non-native English-speaking students with feedback in their own languages (e.g. Spanish and Chinese).

Another concern is that there is limited AES research done by independent researchers. Companies developing the software sponsor the vast majority of studies, and most research findings are based on technical reports that are released by the software developers. For instance, the studies on IntelliMetric™ and MY Access!® are sponsored and conducted by Vantage Learning in the form of technical reports and are not published in refereed journals (see Vantage Learning 2000a, 2000b, 2001a, 2001b, 2002, 2003a, 2003b, 2004). Similarly, the research on Criterion® (the AES system developed by ETS) is conducted by the ETS research team (Attali 2004, Attali and Burstein 2004, Burstein and Chodorow 1999, Burstein, Chodorow and Leacock 2003, Chodorow and Burstein 2004, Lee et al 2008). There is a need for more independent studies, such as this one I report on in this chapter.

Research questions

The AES technology used in this study is MY Access!® (Version 6.0) by Vantage Learning. The four research questions were the following:

1. How did two ESL students who were exposed to AES feedback incorporate the feedback across multiple drafts of their papers?
2. How did two English L2 students who received written teacher feedback (TF) incorporate this type of feedback across multiple drafts of their papers?
3. What differences exist between the two feedback pairs (i.e. AES versus TF) in how they used the feedback they received to revise their drafts?
4. What were the student perceptions regarding the use of AES feedback versus written TF in each feedback pair?

Methodology

Participants

The participants were four adult ESL students who were receiving ESL instruction at an Intensive English Center at a university in the USA. The students were at low-intermediate English proficiency level, and they were from various linguistic backgrounds, including Spanish, Arabic, Turkish and Korean. Table 3.1 provides more information about the participants concerning their language and cultural backgrounds, age, gender, length of time in the USA, and their reasons for studying English. The participating teacher was a native-English speaker with a doctoral degree in ESL.

Table 3.1 Background information about the case study participants

Names (pseudonyms)	Pedro	Khalid	Songie	Selma
Country	Chile	Saudi Arabia	South Korea	Turkey
Native language	Spanish	Arabic	Korean	Turkish
Age range	24–29	30–35	24–29	24–29
Gender	Male	Male	Male	Female
Length of time in the US	< 6 months	6 months–1 year	< 6 months	< 6 months
Reason for studying English	Employment	Employment	Education	Employment

Data collection

A class of 12 students was divided into two groups. Approximately half of the students were exposed to computerized feedback and were called the AES group for the purposes of this study, while the other half received written feedback from the teacher and were called the TF group. Although there were 12 students in the class, the focus of this study was on four of the students who comprised the case study, two students from each group. The students were selected based on the holistic scores they received on the diagnostic essays that were administered at the beginning of the study. Students wrote three draft papers on each of the five prompts for five weeks. Khalid and Pedro received feedback from the teacher, so they are addressed as the TF pair in this study, whereas Songie and Selma were exposed to AES feedback and are referred to as the AES feedback pair. Purposeful sampling was used to assign the participants to the AES group and the TF group. The diagnostic essays were scored holistically by two different raters based on a 6-point scale using a holistic rubric that was generated by the MY Access!® program. The correlation between the raters was computed using a non-parametric correlation method (Spearman's rho) and the result was 0.93.

Three times a week the participants typed their essays using the MY Access!® program in separate computer labs; they wrote three drafts on five prompts. The essays were evaluated both holistically and analytically either by the computer or by the teacher. Scoring was based on the same point-scale holistic and analytic rubrics generated by the MY Access!® program. Both the AES and the TF group students typed their essays using the MY Access!® program in response to the same prompts with the MY Access!® program functioning similar to the Microsoft Word program for the TF group.

The data were collected from essays using the five traits in the analytic feedback rubric (i.e. focus and meaning, content and development, organization, language use and style, and mechanics and conventions). In addition holistic feedback scores were assigned to the essays either by the MY Access!® program or by the teacher. Other data collected included surveys (demographic and computer literacy surveys), interviews, and classroom observations. To ensure consistency, the feedback and revisions in the study were also independently coded by a second person using the same categories.

To answer the first and second research questions, the initial and subsequent drafts written by the case study students, as well as the related feedback provided either by the MY Access!® program or by the teacher were classified and analyzed. The five writing traits generated by the MY Access!® program were used as a basis to classify both the written TF and AES feedback in this study. Table 3.2 presents the classification of the indicators of writing quality for each trait.

Table 3.2 Indicators of writing quality on each trait of narrative rubric

Trait	Focus	Indicator
Focus and meaning	Content	Understanding of purpose and task, awareness of audience, and having a main idea.
Content and development	Content	Planning (having an outline), including sensory details and details that support the main idea. Using quotes and dialogs and paragraph-length entries (of 4 to 8 sentences).
Organization	Content	Including an introduction paragraph, body paragraphs, and a conclusion. Using transitions, and having a logical order of ideas.
Language use and style	Form	Sentence structure (length, simple versus complex sentences, using a variety of sentence beginnings), word choice (variety, using sensory details), and writing style (formal versus informal).
Mechanics and conventions	Form	Using correct grammar, mechanics, spelling and punctuation (format/spacing, punctuation, plural form, spelling, word form, clause, preposition, pronoun, verb tense, verb form (infinitive, auxiliary), preposition, and article use).

Data from the student drafts and the related feedback either from the teacher or from the AES system were compared and cross-referenced to explore how students incorporated the type of feedback they received into their revisions. The feedback points that students were presented with were classified as *usable written TF/AES feedback* and *other written TF/AES feedback*. Additionally, the revisions made by each student on each draft were categorized as follows:

a. changes that might be based on written TF/AES feedback
b. partial changes that might be based on written TF/AES feedback
c. changes that are not based on written TF/AES feedback, and
d. usable written TF/AES feedback that is not used.

Hyland's 1998 study sheds a light on the feedback and revision categories listed above. In the present study, the researcher adapted Hyland's (1998) feedback category *usable written TF/AES feedback* and added another feedback category called *other written TF/AES feedback* for unusable feedback, problem statement, or positive reinforcement. This category was partially implied (e.g. positive reinforcement) by Hyland (1998), but it was not included in his study. Finally, this study included a third feedback category called *usable written TF/AES feedback that was not used*, which was categorized as *points not acted* on by Hyland (1998).

To answer the third research question, the number of feedback points (i.e. either AES feedback or written TF) provided to students and the number, as well as the percentage of feedback points each student used in their revisions, were calculated and compared. The percentages provided an indication of how similar or different the students were in terms of number of feedback points they were presented and the usefulness of each type of feedback (i.e. AES feedback vs. written TF). The revision categories, which were determined based on the changes that students made on each draft, were matched to the writing traits in the analytic rubric (focus and meaning, content and development, organization, language use and style, and mechanics and conventions). For example, changes in terms of capitalization in student essays were categorized as mechanics and conventions.

The fourth research question was answered through an opinion survey, observations of students' writing processes, and interviews. These data sources not only provided information on the context in which the feedback was given, but they also assisted the researcher in explaining the possible reasons for the differences found in students' drafts.

Results

The MY Access!® program promoted process writing by letting all students (regardless of the type of feedback they received) save their essays in an online portfolio, which permitted students to make revisions in their previous drafts. Observation results showed that all participants were concerned about adding more details in their essays. The AES feedback pair went back and revised the mechanical and conventional aspects of the existing paragraphs based on the feedback they received from the MY Editor feature of the program. Similarly, the TF feedback pair revised a previous sentence if the teacher pointed out an error or suggested a change. Both feedback pairs also made changes that were not based on AES feedback or TF.

Document analyses revealed that there were differences in the type of revisions performed by the participants in terms of the five traits – focus and meaning, content and development, organization, language use and style, and mechanics and conventions. While the difference within each pair was quite small for most traits, the difference across pairs was much larger for the traits that focused on content (i.e. excluding the traits that focused on form – language use and style and mechanics and conventions).

Focus and meaning trait

The students who were exposed to AES feedback (Songie and Selma) received many more feedback points on focus and meaning compared to those that were presented with written TF (Pedro and Khalid). Songie received 34 usable feedback points on focus and meaning, and Selma received 35. Songie used five of them in her drafts, and Selma used three. In addition, they partially used another four of the feedback points; consequently, it seems that both students failed to use the majority of the feedback points on focus and meaning (N = 25 versus N = 28). On the other hand, the students who were exposed to the written TF did not receive a large number of feedback points on focus and meaning. Pedro did not receive any written TF on focus and meaning and Khalid received only seven feedback points. Nevertheless, he fully incorporated five of seven and partially used two of them in his drafts. Table 3.3 shows a summary of the extent to which the case study participants used focus and meaning feedback in their drafts.

Content and development trait

The students who were exposed to the AES feedback received three to four times more feedback points on content and development compared to the students who were exposed to the written TF. However, Songie and Selma chose not to use more than half of the feedback points generated by the MY Access!® program. Pedro received seven feedback points from the teacher on content and development, and he used only three of them. Khalid used four out of six feedback points he received on content and development. Table 3.4 provides a detailed summary of the information regarding feedback points that both feedback pairs received, full or partial changes they

Table 3.3 Feedback on the focus and meaning trait

Students	Usable AES feedback/ written TF		Changes that might be based on AES feedback/written TF		Partial changes that might be based on AES feedback/written TF		Usable AES feedback/written TF that is not used	
	TF	AES	TF	AES	TF	AES	TF	AES
Pedro								
Khalid	7		5 (71%)		2 (29%)			
Songie		34		5 (15%)		4 (12%)		25 (73%)
Selma		35		3 (9%)		4 (11%)		28 (80%)

Table 3.4 Feedback on the content and development trait

Students	Usable AES feedback/written TF		Changes that might be based on AES feedback/written TF		Partial changes that might be based on AES feedback/written TF		Usable AES feedback/written TF that is not used	
	TF	AES	TF	AES	TF	AES	TF	AES
Pedro	7		3 (43%)				4 (57%)	
Khalid	6		3 (50%)		1 (17%)		2 (33%)	
Songie		45		6 (13%)		12 (27%)		27 (60%)
Selma		45		4 (9%)		16 (35.5%)		25 (55.5%)

made based on the type of feedback they received, and usable feedback that was not used.

Organization trait

The written TF on organization and content and development traits were closely related to each other. When the teacher suggested a feedback point on organization, it directly affected the content and development of the text as well. For example, including a body paragraph is a feedback point related to organization. However, adding a body paragraph also impacts content and development. For example, by adding a body paragraph a writer includes more details in the text. Because of this close relationship, the written TF feedback points on organization and content and development were combined in the analysis. As a result, the information presented in Table 3.5 regarding the revisions that the TF pair made in organization and the revisions they made in content and development in Table 3.4 remained the same. However, the results were different for the AES feedback pair for both traits. See Tables 3.4 and 3.5 for details.

Language use and style trait

The students who received written TF and those who were provided with the AES feedback differed dramatically in terms of the feedback points they received on the language use and style trait. While Khalid did not receive any feedback on language and style, Pedro received five written feedback points that were usable on the same trait. Except for one, he incorporated all of them. On the contrary, the AES feedback pair, Songie and Selma, received more than five times more usable AES feedback points on language use and style. They both received exactly the same number of feedback points on language use and style (N = 27) and incorporated the same number of feedback points (19%) on this trait. Table 3.6 illustrates the extent to which the students used language use and style feedback in their drafts.

Mechanics and conventions trait

Even though both the teacher and the MY Access!® program (MY Editor feature) offered the majority of their feedback on mechanics and conventions, the MY Access!® program suggested more feedback points than the teacher. The students who were exposed to the written TF used almost all of the feedback points suggested by the teacher (86.5% and 93.5% respectively) on mechanics and conventions; the AES feedback pair incorporated fewer feedback points (79% and 62%) into their drafts. The number of feedback points the AES feedback pair did not use was greater than for the written TF pair. Pedro and Khalid failed to use less than 10% of the feedback while

Table 3.5 Feedback on the organization trait

Students	Usable AES feedback/ written TF		Changes that might be based on AES feedback/written TF		Partial changes that might be based on AES feedback/written TF		Usable AES feedback/written TF that is not used	
	TF	AES	TF	AES	TF	AES	TF	AES
Pedro	7		3 (43%)				4 (57%)	
Khalid	6		3 (50%)		1 (17%)		2 (33%)	
Songie		34		5 (15%)		11 (32%)		18 (53%)
Selma		36		6 (17%)		11 (30%)		19 (53%)

Table 3.6 Feedback on the language use and style trait

Students	Usable AES feedback/ written TF		Changes that might be based on AES feedback/written TF		Partial changes that might be based on AES feedback/written TF		Usable AES feedback/written TF that is not used	
	TF	AES	TF	AES	TF	AES	TF	AES
Pedro	5		4 (80%)				1 (20%)	
Khalid								
Songie		27		1 (4%)		4 (15%)		22 (81%)
Selma		27		1 (4%)		4 (15%)		22 (81%)

Songie and Selma did not use more than 20% of the feedback they were suggested. Table 3.8 displays the usable feedback points suggested both by the teacher and the MY Access!® program. It also shows how many of these feedback points the students used or did not use in their drafts.

Total feedback

The major difference between students who were exposed to the AES feedback (i.e. Songie and Selma) and the ones who were exposed to the written TF (i.e. Pedro and Khalid) was that the former received a larger number of usable feedback points than the latter on all traits. Songie and Selma received a total of 236 and 241 usable AES feedback points respectively; whereas, Pedro and Khalid received a total of 86 and 90 usable written TF points respectively. The feedback points that were suggested by the MY Access!® program were more than double those assigned by the teacher. Table 3.8 below provides the visual display of the usable feedback suggested either by the teacher or by the AES system, as well as data showing the extent to which the feedback was used by the students based on each trait. Table 3.8 also presents differences in the type of revisions performed by the participants.

Another important finding was that the students who were exposed to the AES feedback incorporated more partial changes into their drafts than those who were exposed to the written TF. Songie and Selma made 31 (13%) and 36 (15%) partial changes respectively in their drafts based on the AES feedback; whereas, Pedro and Khalid made only six (7%) and four (5%) partial changes respectively based on the written TF. Table 3.9 shows the extent of AES feedback or written TF that the case study participants used in their drafts.

One possible reason for differences across pairs in terms of the use of feedback in revisions is that the MY Access!® program provided extensive, redundant, and generic feedback, which discouraged the students from reading and using it. The MY Tutor feature of the program, which generated feedback on all five traits, provided generic and redundant feedback regardless of the amount of writing the AES feedback pair produced or how low or high their scores were. Songie and Selma received either exactly the same feedback or very similar feedback with slightly different wording when they submitted a draft to the MY Access!® program (see Figure 3.1 for an example). Unlike the written TF, the AES feedback was not cumulative. For example, even though Selma wrote an introduction paragraph in a previous draft, MY Tutor suggested that she include one in the subsequent draft. The MY Tutor feature generated lengthy feedback, as well. These characteristics considerably reduced the motivation of the AES feedback pair. While Selma stopped checking the MY Tutor feedback after the first prompt (Drafts 2 and 3), Songie used MY Tutor feedback in Prompt 1 (Drafts 2 and 3), Prompt 2 (Draft 2 only), Prompt 3 (Draft 3 only), and Prompt 4 (Draft 3 only). They

Table 3.7 Feedback on the mechanics and conventions trait

Students	Usable AES feedback/ written TF		Changes that might be based on AES feedback/written TF		Partial changes that might be based on AES feedback/written TF		Usable AES feedback/written TF that is not used	
	TF	AES	TF	AES	TF	AES	TF	AES
Pedro	74		64 (86.5%)		6 (8%)		4 (5.5%)	
Khalid	77		72 (93.5%)				5 (6.5%)	
Songie		96		76 (79%)				20 (21%)
Selma		98		61 (62%)				37 (38%)

Table 3.8 The summary of the usable TF/AES feedback suggested and the percentage used by the students

	Focus and meaning		Content and development		Organization		Language use and style		Mechanics	
	Usable feedback	Total feedback used	Usable feedback	Total feedback used	Usable feedback	Total feedback used	Usable feedback	Total feedback used	Usable feedback	Total feedback used
Pedro	n/a	n/a	7	43%	7	43%	5	80%	74	94.5%
Khalid	7	100%	6	67%	6	67%	n/a	n/a	77	93.5%
Songie	34	27%	45	40%	34	47%	27	19%	96	79%
Selma	35	20%	45	44.5%	36	47%	27	19%	98	62%

Table 3.9 The extent of total feedback used by the students

Student	Number of the total feedback points (usable and other)	Number of the usable feedback points	Changes (full)	Partial changes	Total changes
Pedro	110	86	71 (82.5%)	6 (7%)	77 (89.5%)
Khalid	108	90	82 (91%)	4 (5%)	86 (96%)
Songie	602	236	93 (39%)	31 (13%)	124 (52%)
Selma	604	241	75 (31%)	36 (15%)	111 (46%)

both pointed out the redundant and extensive nature of the AES feedback during the interviews. While Selma said, 'Too much reading . . . it's same . . . boring', Songie stated, 'Good, but long . . . very, very long.'

Figure 3.1 AES (MY Tutor feature) feedback on organization for the AES feedback pair

<table>
<tr><th>Organization feedback Songie received from the MYAccess!® program based on a score of 3 in Prompt 3/Draft 2</th><th>Organization feedback Selma received from the MYAccess!® program based on a score of 3 in Prompt 3/Draft 2</th></tr>
<tr><td colspan="2">Selma (or Songie), on a scale of one to six, your response to this assignment <b>was rated a 3 for organization</b>. Organization relates to your ability to present your ideas in a logical and ordered fashion.
Your organization is limited. Typically, a response at this level shows appropriate organizational structure (beginning, middle and end), but your writing is missing transitions and is not unified and consistent throughout.
Selma, with some attention to the following four ideas, you can make your structure and organization much better.
<b>Introduction</b>: Include an introduction that will do two things:
1. Catch the reader's attention and make it interesting and creative.
2. Tell what the main idea of your writing is.
Introductions are very important and will get your writing off to a good beginning and make the reader want to keep reading.
<b>Conclusion</b>: Include a conclusion. Does your writing end strongly and leave the reader something to think about? Make sure your writing does not just stop.
<b>Transitions</b>: <u>Transitions</u> are the glue that holds the ideas of your essay together. Use the right transitions that will move the reader from idea to idea. The right transitions will connect your details and help the reader.</td></tr>
</table>

The teacher, on the other hand, provided feedback based on the individual needs of each student for a particular draft. For example, unlike Khalid, Pedro received feedback on language use and style; however, he did not receive any feedback on focus and meaning from the teacher. The individualized and specific feedback allowed the written TF pair to follow teachers' suggestions closely. See Figures 3.2 and 3.3 for a comparison of the AES feedback (using the MY Tutor and MY Editor features) to written TF in terms of length and content.

Although the feedback that the MY Access!® program suggested was consistent, it was not usable at all times. There were a number of instances when the MY Editor feature suggested feedback points on grammar, mechanics, and conventions that were either confusing or incorrect (see Table 3.10 for examples). Just like MS Word's grammar and spelling feature, MY Editor failed to provide usable suggestions many times. In that case, students either selected the wrong suggestion or simply ignored the feedback. For example, Selma was confused when MY Editor captured an error but did not provide an explanation stating why that particular word was incorrect:

> One time I wrote “porfoy”. It said spelling error, but it didn’t say why. Today, I wrote “in conclusion” wrong, the computer gave different words. They were not right . . . I wrote about my cousin. It’s looks like speak. The computer said “especially if” shouldn’t be used together. Why? I can use it. I didn’t change.

Figure 3.2 Sample written TF and AES (MY Tutor) feedback on focus and meaning

Sample written TF for Khalid on focus and meaning in Prompt 3/Draft 2	**Sample AES (MY Tutor feature) feedback for Songie on focus and meaning in Prompt 3/Draft 2**
1. However, the sentence **‘but sometimes we don’t knew about the inventor, such as telephone, computer, and TV’** does not belong here because this essay is NOT about the inventor but only the inventions. 2. Also, what exactly is this essay going to be about? 3. What is each paragraph going to be about?	Songie, on a scale of one to six, your response to **this assignment was rated a ‘4’ for focus.** Focus relates to your ability to present a consistent, unified message and stay on topic. Your focus is adequate. Typically, a response at this level gives a fairly clear statement of purpose and wanders only slightly from the topic. Songie, your focus is adequate. Let's look at the following four points as you revise and think about how you can make your writing better. As you get ready to revise, read your essay out loud so you can hear your ideas and words. **Purpose:** What is the purpose of your writing? Are you narrating a story, informing, or persuading? Make sure you are clear on your purpose. **Main idea (controlling idea)**: Do you have a sentence that states the <u>controlling idea</u>, or the <u>main idea</u>, of your essay? Make sure you know exactly where this sentence is, because it determines what details you will use. **Audience**: Think about your audience. Are you writing to someone your age, or are you writing to an adult? Are you writing to someone familiar with your subject and ideas, or are you writing to someone who doesn't know about your subject and ideas? Make sure your details are all ones your audience will easily understand. Task: Did you do the task the prompt tasks you to do?

From time to time, MY Editor provided unnecessary feedback on mechanics and conventions, e.g. capturing every spelling and spacing error, which only made the screen look more crowded. As a result, the students felt overwhelmed with the number of suggestions provided. Songie stated the following when he was asked what he thought about the MY Editor feedback:

> Some part of MY Editor . . . the recommendation is not useful . . . if I don't understand, [I] pass . . . for example, I didn't use comma, spacing . . . it says . . . sometimes too much so pass . . . sometimes spell check gives a word but I don't understand and I don't change.

Figure 3.3 Sample written TF and AES feedback (MY Editor) on grammar, mechanics, and conventions

<table>
<tr><th>Sample written TF for Khalid on grammar in Prompt 5/Draft 2</th><th colspan="2">Sample MY Editor feedback for Songie on "misused words" in Prompt 3/Draft 1</th></tr>
<tr><td rowspan="3">I think that my life would become exciting and fast because it has a lot of things **to I do** (**for me to do**).</td><td>**Advice:**
Error:
Category:
Error label:
Suggestions:</td><td>Consider too instead of 'to'.
USAGE
Misused words
too</td></tr>
<tr><td>Explanation:</td><td>Misused words</td></tr>
<tr><td colspan="2">**Misused words**
The grammar checker flags words or phrases that are often used incorrectly or misunderstood because they are confused with similar words or phrases.
Examples:
Change: The rain did not **effect** the game.
To: The rain did not **affect** the game.
Change: He is good at creating the **allusion** he knows what he's talking about.
To: He is good at creating the **illusion** he knows what he's talking about.
Explanation:
The confused expressions have different meanings and are used in different contexts, so that most of the flags returned identify real errors ('elude to' instead of 'allude to'; 'sit the books on the chair' instead of 'set the books on the chair').</td></tr>
</table>

Table 3.10 Sample AES (MY Editor) non-usable feedback points

Prompt/Draft	Sample sentences with errors	Portion of the AES (MY Editor) feedback
Songie Prompt 3/Draft 2	We can enjoy <*Pronoun errors (ESL)*> with my family during <*Preposition errors (1)*> night time.	**Advice:** Depending on your meaning, it might be better to use **since** or **for** instead of 'during'. **Suggestion:** since~ for
Songie Prompt 5/Draft 1	All<*Punctuation errors*> is free(But<*Spelling errors*> I have to serve about five years).	**Advice:** The word 'free(But' is not in the Main or Personal Dictionary. **Suggestion:** Freeboot
Selma Prompt 4/Draft 1	His house was 25 minutes far by car to my apartment; however when the time was 12:00 am<*Subject-verb agreement errors 1*>, he came my apartment with birthday cake.	**Advice:** Consider **is** or **are** instead of 'am'. **Suggestion:** is~ are
Selma Prompt 4/Draft 3	I was suprized <*Spelling errors1*>.	**Advice:** The word 'suprized' is not in the Main or Personal Dictionary. **Suggestion:** suppressed~ surpassed

The results of the opinion survey provided insights about students' perceptions of the MY Access!® program (see Table 3.11 for more information). All students found the feedback they received helpful. Unlike the AES pair, the TF pair reported that they used a dictionary while typing their essays, but they did not specify whether it was a bilingual or monolingual dictionary. The students who were exposed to the AES feedback used the spell check feature in every draft, whereas those who were exposed to the written TF occasionally used this feature. Finally, both of the written TF students confirmed that they did not use the MY Editor or MY Tutor features of the MY Access!® program. The students who received written TF were not able to access these two features because they were in the off position for them at all times. On the other hand, the AES feedback pair frequently used the MY Editor feature. Selma reported that she sometimes used the MY Tutor feature, while Songie pointed out that he often used this feature. The field notes and interview notes supported the student answers regarding the use of the MY Editor feature. However, they slightly contradicted their responses in the opinion survey regarding the use of the MY Tutor feature. According to the field notes, Selma used the MY Tutor feedback on only two drafts, and Songie used it on five out of 15 drafts. This discrepancy might be because neither student was especially familiar with the name of the MY Tutor feature. During the interviews, they described it as the other feedback or the long one. On the other hand, both Songie and Selma were quite familiar with the MY Editor names because they used the feedback points it suggested on every draft.

Table 3.11 Case study participants' responses to the opinion survey (Part 1)

	Pedro	Khalid	Songie	Selma
How helpful the feedback was	Very helpful	Helpful	Very helpful	Helpful
The internet resources used while writing	Book dictionary [sic]	Dictionary	None	None
Frequency of the internet use	Never	Never	Sometimes	Never
Frequency of cut/copy/paste use	Never	Rarely	Always	Never
Frequency of spell checker use	Sometimes	Sometimes	Always	Always
Frequency of MY Editor use	Never	Never	Always	Always
Frequency of MY Tutor use	Never	Never	Always	Always
How much they think they improved	A lot	Moderate	A little	A little

In the second part of the opinion survey, students were asked to rank the traits (i.e. focus and meaning, content and development, organization, language use and style, and mechanics and conventions) they received on a Likert scale from 1 to 5 with a score of 5 indicating the most feedback and a score of 1 indicating the least. Table 3.12 displays the traits on which students received the most and the least feedback. The results of the opinion survey regarding students' perceptions of the feedback they received either from the teacher or the AES system and the essay analysis data are somewhat contradictory. The analyses of student essays showed that all students, regardless of the type of feedback they were exposed to, were mostly presented with mechanics and conventions feedback. Songie was aware of his problem with grammar because he reported this fact during the interview. It is possible that he did not understand what the errors he made in the mechanics and conventions trait entailed, or it is possible that he might have confused the language use and style trait with the mechanics and conventions trait.

Table 3.12 The traits students received the most and the least feedback on

Traits	Pedro	Khalid	Songie	Selma
Focus and meaning	3	5	4	3
Content and development	4	4	2	3
Organization	5	1	1	3
Language use and style	1	3	5	3
Mechanics and conventions	2	2	3	3

Note: 5 = the most; 1 = the least

Discussion

Research in AES predominantly includes native English-speaking populations; however, software development companies target non-native English-speaking populations for marketing purposes as well (Warschauer and Ware 2006). As a result of marketing, the development companies have included some additional features in the software in an attempt to address the needs of non-native English-speaking writers. For instance, both MY Access!® and Criterion® can provide feedback in various languages. However, including a multilingual feedback capacity in an AES system may not be an effective solution, as ESL students need feedback in English that is appropriate for their English proficiency levels and writing skills to improve their writing in English. Moreover, MY Access!® claims to flag errors that non-native speakers are likely to make; however, the findings in this study showed that the program did not recognize several errors that are common such as articles, preposition, word form, etc. The developing companies need to better understand how ESL/EFL students differ from native English-speaking students, and what type of support systems they need in writing classes.

The results of this study suggest that the AES system may be utilized as an additional tool for giving feedback in L2 writing classrooms so that teachers can promote process writing and increase student motivation by exposing them to different tools for writing essays. They can also be used as a means for students to practice writing independently outside the classroom and at their own pace. The AES system used in this study (i.e. MY Access!®) serves this purpose well. Nevertheless, teacher guidance is crucial when using this program in writing classes because of the limitations that MY Editor and MY Tutor features have (Dikli 2010). As mentioned previously, MY Editor provides direct suggestions on grammar, mechanics, and conventions, yet it also includes generic and standard examples and explanations that may cause students to correct an error without really understanding why they are correcting it. Furthermore, MY Editor sometimes provides incorrect or confusing feedback to L2 writers on the formal aspects of essay writing; consequently, it is imperative for teachers to notify students in advance that they might receive unusable feedback from the MY Access!® program. Teachers who decide to use AES systems in their classrooms are encouraged to help their students develop strategies for questioning the feedback they receive and using other resources, such as dictionaries, thesauruses and grammar textbooks to assist them in making decisions about whether to correct an error that is related to the form of the language. Because providing lengthy and generic feedback is a major limitation with the MY Tutor feature, it is absolutely crucial that teachers help students use and interpret the feedback that MY Tutor generates.

The results showed that unlike the MY Access!® program, the feedback provided by teachers is based on the individual needs of each student for the particular draft on which they were working. This finding supports the work of Chen and Cheng (2008). As Chen and Cheng (2008:107) state, AES feedback 'though delivered instantly and viewed as helpful in improving some formal aspects of writing, provides only formulaic, generic information that cannot address students' individual writing problems, particularly in the areas of coherence and idea development, whereas human feedback can attend to meaning, respond to the writer's thoughts, and give specific, personal comments. This finding has critical implications for teachers of writing. That is, quality rather than the quantity of feedback on writing is effective. While this implication concerns teachers of both L1 and L2 writers, it is particularly important for teachers of L2 writers, as non-native English speakers are prone to make errors while writing in English. As a result, L2 writers may become overwhelmed with the extensive amount of the language presented in the feedback and lose interest in responding to the feedback with which they are provided, especially feedback that is redundant and generic. It is vital for teachers to provide focused and specific feedback based on the individual needs of their students. Table 3.13 summarizes the benefits and drawbacks of the written TF and the AES feedback.

Table 3.13 Benefits and limitations of the written TF and AES feedback

	Written TF	AES feedback
Pros	● Provides individualized, specific feedback. ● Provides shorter feedback (not redundant or generic). ● Provides human interaction and sense of audience.	● Provides immediate feedback. ● Generates systematic feedback on numerous essays. ● Provides multilingual feedback. ● May increase motivation due to instant feedback and multiple submission opportunities. ● Is considered objective.
Cons	● Requires time to provide scoring and feedback. ● Might fail to provide feedback regularly (consistently). ● Can get tired so may limit the number of essays and/or drafts to be submitted.	● Provides generic, redundant and extensive feedback. ● May decrease motivation due to generic, redundant and extensive feedback. ● Lacks human interaction and may lack sense of audience. ● Fails to catch many ESL/EFL errors.

The demand for learning English is steadily increasing. Both ESL/EFL teachers and students from many different contexts and countries around the world could benefit from an effective online tool to help them to teach and learn to write in English. While instructional applications of AES systems such as MY Access!® have a potential for such assistance, they seem to fail to address various needs of ESL/EFL populations in writing in English. Therefore, it is imperative for developing companies to ensure that rigorous attempts are made to meet those needs by conducting more research in ESL/EFL contexts and including ESL/EFL professionals in their product development teams. Future studies looking more closely into automated feedback versus teacher feedback would certainly provide deeper insights regarding the nature of each feedback type. The results can be used to improve the feedback capacities of AES programs.

References

Attali, Y (2004) *Exploring the feedback and revision features of criterion*, paper presented at the National Council on Measurement in Education (NCME), San Diego, April 11–17 2004.

Attali, Y and Burstein, J (2004) *Automated essay scoring with e-rater V.2.0*, paper presented at the Conference of the International Association for Educational Assessment (IAEA), Philadelphia, June 13–18 2004.

Burstein, J and Chodorow, M (1999) *Automated essay scoring for nonnative English speakers*, available online: www.ets.org/Media/Research/pdf/erater_acl99rev.pdf

Burstein, J, Chodorow, M and Leacock, C (2003) *Criterion: Online essay* *evaluation: An application for an automated evaluation of student essays*, paper presented at the 15th Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, 2003.

Chen, C E and Cheng, W E (2008) Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes, *Language Learning & Technology* 12 (2), 94–112.

Chodorow, M and Burstein, J (2004) *Beyond Essay Length: Evaluating e-Rater's Performance on TOEFL Essays*, Princeton: Educational Testing Service.

Chung, G K W K and O'Neil, H F (1997) *Methodological approaches to online scoring of essays*, available online: www.cse.ucla.edu/products/reports/TECH461.pdf

Deane, P (2013) On the relation between automated essay scoring and modern views of the writing construct, *Assessing Writing* 18 (1), 7–24.

Dikli, S (2006) An overview of automated scoring of essays, *Journal of Technology, Learning, and Assessment* 5 (1), available online: ejournals.bc.edu/ojs/index.php/jtla/article/view/1640/1489

Dikli, S (2010) Nature of automated essay scoring feedback, *CALICO Journal* 28 (1), 99–134.

Dikli, S and Bleyle, S (2014) Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing* 22, 1–17.

Edelblut, P and Vantage Learning (2003) *An analysis of the reliability of computer automated essay scoring by IntelliMetric of essays written in Malay language*, paper presented at TechEX, November 3 2003.

Elliot, S and Mikulas, C (2004) *A summary of studies demonstrating the educational impact of the MY Access online writing instructional application*, paper presented at the National Council on Measurement in Education (NCME), San Diego, April 11–17.

Grimes, D and Warschauer, M (2008) Automated writing assessment in the classroom, *Pedagogies* 3 (1), 52–67.

Grimes, D and Warschauer, M (2010) Utility in a fallible tool: A multi-site case study of automated writing evaluation, *Journal of Technology, Language, and Assessment* 8 (6), 1–43.

Hamp-Lyons, L (2001) Fourth generation writing assessment, in Silva, T and Matsuda, P K (Eds) *On Second Language Writing*, Mahwah: Lawrence Erlbaum, 117–125.

Hyland, F (1998) The impact of teacher written feedback on individual writers, *Journal of Second Language Writing* 7 (3), 255–286.

Landauer, T K, Foltz, P W and Laham, D (1998) Introduction to latent semantic analysis, *Discourse Processes* 25, 259–284.

Landauer, T K, Laham, D and Foltz, P W (2003) Automated essay scoring and annotation of essays with the Intelligent Essay Assessor, in Shermis, M D and Burstein, J C (Eds) *Automated Essay Scoring: A Cross-disciplinary Perspective*, Mahwah: Lawrence Erlbaum, 87–112.

Landauer, T K, Laham, D, Rehder, B and Schreiner, M E (1997) *How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans*, paper presented at the 19th Annual Conference of the Cognitive Science Society, Stanford, 1997.

Lee, W Y, Gentile, C and Kantor, R (2008) *Analytic Scoring of TOEFL CBT Essays: Scores from Humans and E-rater (RR08-01)*, Princeton: Educational Testing Service.

Myers, M (2003) What can computers and AES contribute to a K-12 writing program? in Shermis, M D and Burstein, J C (Eds) *Automated Essay Scoring: A Cross-disciplinary Perspective*, Mahwah: Lawrence Erlbaum, 3–20.

Nichols, P D (2004) *Evidence for the interpretation and use of scores from an Automated Essay Scorer*, paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, April 12–16 2004.

Page, E B (2003) Project essay grade: PEG, in Shermis, M D and Burstein, J C (Eds) *Automated Essay Scoring: A Cross-disciplinary Perspective*, Mahwah: Lawrence Erlbaum, 43–54.

Ramineni, C (2013) Validating automated essay scoring for online writing placement, *Assessing Writing* 18 (1), 40–61.

Ramineni, C and Williamson, M D (2013) Automated essay scoring: Psychometric guidelines and practices, *Assessing Writing* 18 (1), 25–39.

Shermis, M D and Burstein, J (Eds) (2003) *Automated Essay Scoring: A Cross-disciplinary Perspective*, Mahwah: Lawrence Erlbaum.

Vantage Learning (2000a) *A Study of Expert Scoring and IntelliMetric Scoring Accuracy for Dimensional Scoring of Grade 11 Student Writing Responses (RB-397)*, Newtown: Vantage Learning.

Vantage Learning (2000b) *A True Score Study of IntelliMetric Accuracy for Holistic and Dimensional Scoring of College Entry-level Writing Program (RB-407)*, Newtown: Vantage Learning.

Vantage Learning (2001a) *A Preliminary Study of the Efficacy of IntelliMetric for Use in Scoring Hebrew Assessments (RB-561)*, Newtown: Vantage Learning.

Vantage Learning (2001b) *Applying IntelliMetric Technology to the Scoring of 3rd and 8th Grade Standardized Writing Assessments (RB-524)*, Newtown: Vantage Learning.

Vantage Learning (2002) *A Study of Expert Scoring, Standard Human Scoring and IntelliMetric Scoring Accuracy for Statewide Eighth Grade Writing Responses (RB-726)*, Newtown: Vantage Learning.

Vantage Learning (2003a) *Assessing the Accuracy of IntelliMetric for Scoring A District-wide Writing Assessment (RB-806)*, Newtown: Vantage Learning.

Vantage Learning (2003b) *How Does IntelliMetric Score Essay Responses? (RB-929)*, Newtown: Vantage Learning.

Vantage Learning (2004) *A Practitioner's Guide to Scientifically-based MY Access Research (RB-931)*, Newtown: Vantage Learning.

Warschauer, M and Ware, P (2006) Automated writing evaluation: Defining the classroom research agenda, *Language Teaching Research* 10 (2), 1–24.

4 Effects of pragmatic task features on temporal measures of Chinese ESL and EFL spoken request production

Lixia Cheng

Purdue University, Indiana, US

Motivation for the research

Language *tasks* have been defined as activities involving the use of language, each of which is designed for the purpose of 'achieving a particular goal or objective in a particular situation' (Bachman and Palmer 1996:44). Task effects on second language (L2) performance have been studied quite extensively, especially with respect to the relationships between psycholinguistic task features (e.g. familiar/unfamiliar information), performance conditions (e.g. length of pre-task planning time) and task response characteristics such as production complexity, accuracy, and fluency (CAF) (e.g. Elder, Iwashita and McNamara 2002, Foster and Skehan 1996, Iwashita, McNamara and Elder 2001, Robinson (Ed) 2011, Tavakoli and Foster 2008, Xi 2005). Skehan (1998:99) was one of the first researchers to discuss theories related to the impact of the cognitive characteristics of tasks on L2 learners' task performance when he proposed 'a three-way distinction (i.e. code complexity, cognitive complexity, and communicative stress) for the analysis of tasks'. Skehan's model, used primarily in pedagogical contexts, has to date not had a great impact in testing situations (e.g. Elder et al 2002, Iwashita et al 2001, Norris, Brown, Hudson and Yoshioka 1998).

Fulcher and Márquez Reiter (2003) comment that the 'lack of score sensitivity to variation in task' (Fulcher and Márquez Reiter 2003:325) in studies that examined task effects from a psycholinguistic perspective was due to the fact that 'learner ability accounts for most score variance, and task difference, even if significant, accounts for only a small part of score variance' (Fulcher and Márquez Reiter 2003:326). Based on their investigation of how first language (L1) background, social power, and rank of imposition impact the assessment of task outcomes, Fulcher and Márquez Reiter (2003) suggest that a pragmatic perspective is a viable alternative and can make task difficulty specific to individual speakers.

Review of the literature

This study used Speech Act Theory (Austin 1962). Pragmatic task features were manipulated with the purpose of detecting the impact, if there was any, of pragmatic conditions on task performance. Pragmatic task features were studied through a composite measure representing the additive effects of the three well-known pragmatic variables: Power (P), Distance (D), and Rank of imposition (R). Specifically, this study investigated whether PDR (of pragmatic tasks) has differential effects on L1 Chinese participants' oral English production of requests. In addition, it examined the extent to which PDR interacts with English language proficiency and the learning setting: English as a Second Language (ESL) or English as a Foreign Language (EFL).

This study grew out of my ongoing interest in examining the impact of task difficulty and variability on L2 performance. According to Bachman (2002), one of the two general approaches to task difficulty surfaced in language testing research is to explicitly identify 'difficulty features, which are essentially combinations of ability requirements and task characteristics' (Bachman 2002:463). Studies that fall under this approach may ask individuals to rate the expected difficulty levels of tasks, as well as solicit expert ratings of L2 task performances (e.g. Brown, Hudson, Norris and Bonk 2002, Elder et al 2002, Iwashita et al 2001, Norris et al 1998, Robinson 1995, 2001, Skehan 1998). Unfortunately, no systematic relationship has been found between *a priori* estimates of difficulty (i.e. difficulty features) and empirical indicators of difficulty (e.g. ratings of task difficulty or expert judgments of task performances).

In contrast to these largely failed attempts to extend Skehan's (1998) psycholinguistic model of task difficulty from pedagogical contexts to testing situations, Fulcher and Márquez Reiter's (2003) and Taguchi's (2007) studies suggest greater potential of a pragmatic perspective on task difficulty in speaking assessments.

Fulcher and Márquez Reiter (2003) examined the extent to which L1 cultural background and pragmatic task conditions (i.e. relative power between interlocutors and the level of imposition of the request) account for the variance in the assessment of task outcomes. Twenty-three L1 Spanish and 32 L1 English test takers were paired within their L1 group for performing six role plays and evaluating their own pair's communicative success in those role plays. Two weeks later, individual participants completed six questionnaires (one for each role play task), independently assigning probabilities of success to written transcriptions of request productions drawn from the entire pool. Primary findings from a three-way univariate analysis suggest that L1 background, the power differential between interlocutors, and the level of imposition of the request all have significant influence on the assessment of task outcomes. Among these, the influence of power differential on outcome assessment depends on task respondents' L1 background, and vice versa.

Similarly, Taguchi (2007) also examined the effects of task difficulty from a pragmatic point of view and asked participants to perform role plays. She analyzed 29 high and 30 low proficiency Japanese EFL participants' oral English production of requests and refusals. The dependent variables examined in relation to task performance include pre-task planning time, average speech rate (i.e. number of words produced per minute), and overall appropriateness of the requests and the refusals produced, as judged by L1 English expert raters using a six-point ordinal rating scale. Taguchi also manipulated the three variables (i.e. power, distance, and rank of imposition) and focused on two quite extreme situation types defined by the additive effects of the three variables: *PDR-high* and *PDR-low*. PDR-high denotes social situations where the speaker's social power is less than the hearer's, and the distance between the interlocutors and the degree of imposition associated with the speech act are both large. PDR-low denotes social situations where the speaker and hearer are power equals, and the distance between them and the degree of imposition of the speech act are both small.

To examine whether situation type (PDR-high or PDR-low) and English proficiency (high or low) have differential effects on pre-task planning time, speech rate, and the appropriateness of the EFL speech act production, Taguchi (2007) performed three repeated-measures Analysis of Variance (RM-ANOVA) tests, each with a Bonferroni-adjusted significance level (alpha = 0.05/3 = 0.017). Her primary findings include significant main effects of situation type and English proficiency, as well as a significant interaction between these two factors when appropriateness scores and speech rate are considered. With regard to pre-task planning time, however, only situation type has a significant main effect, and there is no significant interaction between situation type and English proficiency. Regarding the effects of situation type, in either proficiency group, PDR-high speech acts require a longer planning time and lead to slower production rated as less appropriate than PDR-low speech acts.

These two studies (Fulcher and Márquez Reiter 2003, Taguchi 2007) conclude that pragmatic task features (e.g. power differential, rank of imposition, or the composite measure PDR) have an influence on the assessment of task outcomes or temporal measurements and appropriateness ratings of task performance. In addition to the inspiration drawn from this line of research examining task difficulty from a pragmatic perspective, the present study was also motivated by 1) a reflection on my own pragmatic performances, especially when I first arrived in the USA, and 2) my observations of the challenges that international undergraduate students in my freshman composition class encountered when they were developing sociocultural awareness at the same time they were trying to improve their language proficiency. These experiences were confirmed by the research literature arguing that L2 pragmatic competence does not develop in parallel with grammatical competence, contrary to the common understanding of a strong correlation and proportional growth between grammatical and pragmatic competence (Bardovi-Harlig and Hartford 1990, 1993). Even learners with high grammatical competence still may have not mastered the nuances of the appropriate use of the L2 in social contexts and, thus, may come across as pragmatically inappropriate (Bardovi-Harlig 1999). Worse still, violations of pragmatic rules and principles may have even more serious consequences on L2 learners than grammatical errors. Indeed, pragmatic failure often has unfortunate consequences for the L2 speaker as such failure may be attributed to deficiencies as a person rather than deficiencies as a language learner (Edwards no date, Thomas 1983).

Given the fact that pragmatics constitutes an essential component of communicative competence (Bachman 1990), it is important that language educators and testers have some knowledge of the conditions and factors that cause difficulty for L2 pragmatic performance. Roever (2001) lists the effects of learning setting as one of the three major factors affecting the development of L2 pragmatics, along with more intrinsic factors such as L1 pragmatic transfer and learner aptitude for pragmatics. Second language settings (i.e. settings in which the target language is widely spoken outside of the classroom) are commonly believed to have an advantage over foreign language settings in facilitating the acquisition of routine formulas and increasing learners' awareness of pragmatic appropriateness due to 'much greater potential availability of input' (Roever 2001:15) from interlocutors and real-life models in the L2 settings. Kitao (1990) argues that Japanese ESL learners approximate native speaker norms better than Japanese EFL learners in their perception of politeness in requests. In addition, second language settings are often associated with less negative transfer of L1 pragmatics. Takahashi and Beebe (1993) assert that negative transfer of L1 pragmatic norms to L2 production of refusals is more prevalent in EFL contexts than in ESL contexts.

Despite the need in language pedagogy and assessment for a better understanding of the factors causing difficulty in pragmatic task performance, there is a paucity of research on the assessment of L2/FL (foreign language) pragmatic competence (Roever 2011). The present study contributes to L2 pragmatic assessment, an underexplored sub-area of L2 assessment, by probing into task effect from an experimental pragmatic perspective. Task performances were examined through the lens of temporal measures of oral English speech act formulation and production, using variables, such as response latency and speech rate. Speech rate is a temporal variable that is recognized as a good predictor of speaking fluency (Lennon 1990, Towell, Hawkins and Bazergui 1996). Response latency, on the other hand, has been utilized more often as an independent variable with levels (e.g. 3.5 minutes versus 0.5 minutes) (see Iwashita et al 2001). However, it was adopted in this study as a dependent variable for measuring the automaticity of speaking during the conceptualization and formulation phases in Levelt's (1989) model. While human evaluations of pragmatic appropriateness of speech act production would add an important dimension to the understanding of pragmatic task performances, it was deemed more appropriate to focus only on temporal measures in this study. This decision was made because human ratings follow a different measurement scale than temporal variables (i.e. ordinal versus interval scales).

Research questions

The fundamental question studied was which factors, intrinsic or extrinsic to learners of English, affect their spoken production of requests in English. Requests were selected because they are one type of speech act that has been widely researched and also are frequently used by ESL learners in academic and institutional settings in the USA and other English-speaking countries.

As indicated in the research literature (e.g. Roever 2001, Taguchi 2007), there are three factors that can have effects on learners' oral English production of requests – 1) pragmatic task features, 2) oral English proficiency, and 3) the learning setting. Among these three independent variables, pragmatic task features constituted the central variable of interest. It was operationalized with PDR, a composite measure of the additive effects of power, distance, and rank of imposition. PDR was a within-subjects factor with two levels: PDR-high and PDR-low. Each participant was asked to perform all of the PDR-high and PDR-low request elicitation tasks. The other two independent variables (i.e. English proficiency and learning setting) were between-subjects factors comprised of two levels: high/low proficiency or ESL/EFL, which means that a participant was classified into either condition of proficiency and either condition of learning setting.

The following questions guided this research on the effects of pragmatic task features, English proficiency, and learning setting on Chinese ESL/EFL spoken request production:

1. Do pragmatic task features (i.e. PDR) influence temporal measures of spoken request production in terms of response latency and speech rate?
2. Are there differences in mean response latency and mean speech rate between high and low proficiency English learners, and between participants who learn English in ESL and EFL settings?
3. Does the influence of English proficiency, if there is any, on response latency and speech rate depend on learning setting? And does the influence of learning setting, if there is any, on response latency and speech rate depend on learner proficiency?
4. Does the influence of pragmatic task features on response latency and speech rate depend on English proficiency or learning setting?

Data collection procedures

Participants

To examine the effects of English proficiency and learning setting on L1 Mandarin speakers' oral English production of requests, 20 high and 20 low speaking proficiency Chinese learners of English who had been in the USA for over three months were recruited from a large Research 1 (R1) public university in the Midwest. R1 is a designator used to refer to an institution of higher education in the US that engages in extensive research activity and offers doctoral degrees. In addition, 20 high and 20 low speaking proficiency Chinese learners of English who had never been in an English-speaking country for more than one month were recruited from two large 4-year universities in mainland China. The cell size of 20 was decided because *a priori* computation in G*Power 3.1.6 (Faul, Erdfelder, Lang and Buchner 2007), a free software program for statistical power analyses, suggested that a total sample size of 56 participants would be sufficient to detect a main effect of the within-factor (i.e. PDR) and a within-between interaction effect (i.e. between PDR and proficiency, or between PDR and learning setting) with an actual power of 0.8.

The age range of all Chinese participants was 19–35. It was not very likely to balance potential gender effects on request production due to two practical constraints: 1) more females than males voluntarily signed up for the study; and 2) the independent measure used for evaluating oral English proficiency screened out many initial-stage participants and placed more females in the EFL high proficiency group whereas more males on the EFL low proficiency side.

Instruments

The independent screening measure is part of the publicly available practice test of a computer-based semi-direct Oral English Proficiency Test (OEPT), locally developed and marked by the Oral English Proficiency Program (OEPP) at Purdue University. The OEPT has been implemented since 2001 to evaluate International Teaching Assistant (ITA) candidates' general speaking ability. The practice test of the OEPT represents an additional form comparable to the four operational test forms (see Purdue University Oral English Proficiency Program 2013).

The OEPT uses a six-point holistic rating scale of 35–40–45–50–55–60 with 50 being the cut score. In this study, two score levels, '35' and '50' were used as the criterion for low and high proficiency respectively. The two higher scores, '55' and '60', were not used because the very low number of Chinese students with either score in the OEPT test database suggested extreme difficulty in participant recruitment if '55' or '60' were used for the high proficiency screening. Table 4.1 is an excerpt from the OEPT holistic rating scale and displays the criterial differences between '35' and '50', the two English proficiency levels included in this study.

Table 4.1 Criterial differences between low and high oral English proficiency levels in this study (adapted from Purdue University Oral English Proficiency Program 2013:9)

OEPT level	**ITA certification result**	***General proficiency level*** *Requirements of listener* **Performance of speaker**
60	**pass**	
55		
50		***Adequate and ready for teaching undergraduates without support.*** *Acceptably small amount of listener effort required to adjust to accent/prosody/intonation.* **Consistently intelligible and comprehensible. Speaker may exert a little noticeable effort, but despite minor errors of grammar/vocab/stress/fluency, message is adequately coherent, with correct information, some lexico-syntactic sophistication, and displays of automaticity and fluency.**
45	**fail**	
40		
35		***Restricted – May need more than one semester of support.*** *Listening may require considerable effort.* **May be unintelligible or incoherent more than occasionally OR have marked deficiencies in at least three other areas: fluency, vocabulary, grammar/syntax, listening comprehension, articulation/pronunciation, and prosody. May have difficulty completing responses.**

To identify the eligible participants for the main study with pragmatics tasks, each proficiency screening test was double rated by the same two certified OEPT raters and was assigned to a third rater, if the first two raters disagreed.

After the screening test responses were rated, eligible L1 Chinese participants were invited back to record their responses to four request-elicitation tasks delivered via a computer-mediated semi-direct oral Discourse Completion Test (DCT). The oral DCT was the major part of a computer test program written in JavaScript by a computer programmer. A semi-direct test measure was favored over a more direct measure (e.g. role plays) in collecting oral English production of requests because computer-based test administration facilitates an accurate, direct capture of the measurement of variables, such as response latency. In addition, computer test administration can minimize inconsistency in speaking data collection potentially caused by using different individuals for the interlocutor role that would be necessary because of the scope of this study.

The four tasks in Table 4.2 were two exemplars each of two types of social situations: ‘PDR-high’ or ‘PDR-low’ (terms adapted from Taguchi 2007). A total of eight exemplars (four PDR-high and four PDR-low), including these four, had been given to L1 American English informants to rank as the most or least awkward and challenging. Two most and two least awkward and challenging situations (see Table 4.2) were selected and used in the oral DCT in this study. The decision to keep two situations each for PDR-high and PDR-low was due to the fact that there was no exact agreement among the L1 English informants concerning which situation was the most awkward and which was the least awkward. Besides, including two situations for either PDR condition minimized the possible effect of using a problematic task exemplar to elicit oral speech act production.

Table 4.2 The four task exemplars in the oral DCT used in the study

Pragmatic task type	Situation description	Situation abbreviation
PDR-low	You have an English exam tomorrow. You are studying in the school library to prepare for the exam. A good friend of yours is also studying in the library, sitting just a few seats away from you. Your pen just ran out of ink, so you go over to your friend and want to ask her to lend you a pen.	‘pen’
PDR-high	You are applying for a job in a company and want to make an appointment for an interview. You have heard that the manager is very busy and usually only schedules interviews in the afternoon from one to four o’clock. However, you have exams or teaching appointments during those hours every day this week. You are leaving a phone message for the manager to ask to schedule an interview in the morning.	‘manager’
PDR-low	It is Sunday afternoon. You and your younger brother are sitting on the couch in your living room, watching TV. Your brother has just stood up to get himself a snack. Since he is already up, you want to ask him to get you the remote control that is lying on the TV stand.	‘TV remote’
PDR-high	Your cousin called you half an hour ago to ask if you could help her out with a difficult situation. You agreed, even though you have an exam scheduled for today in just two hours. You are leaving a phone message for your professor to ask if you could take the exam a day late, even though that would mean your professor would have to write another version of the test just for you.	‘professor’

In addition to the four request elicitation situations, four politeness-unrelated tasks with the same format were inserted one after each pragmatic situation to function as distractors that could potentially divert participants’ attention from the main target of research. Moreover, the presence of these distractors could reduce practice effects and make it harder for participants to remember and recycle the routines they had used to respond to a previous pragmatic situation.

Each task description, occupying one screen, was displayed in both English and Chinese texts, as well as presented via audio recordings of an L1 English speaker reading out the scenario. Chinese translation was provided to ameliorate possible adverse effects on low proficiency Chinese ESL and EFL learners. On the two preceding screens, participants had been asked to perform a sound check on the headset provided to them, and read and listen to test directions.

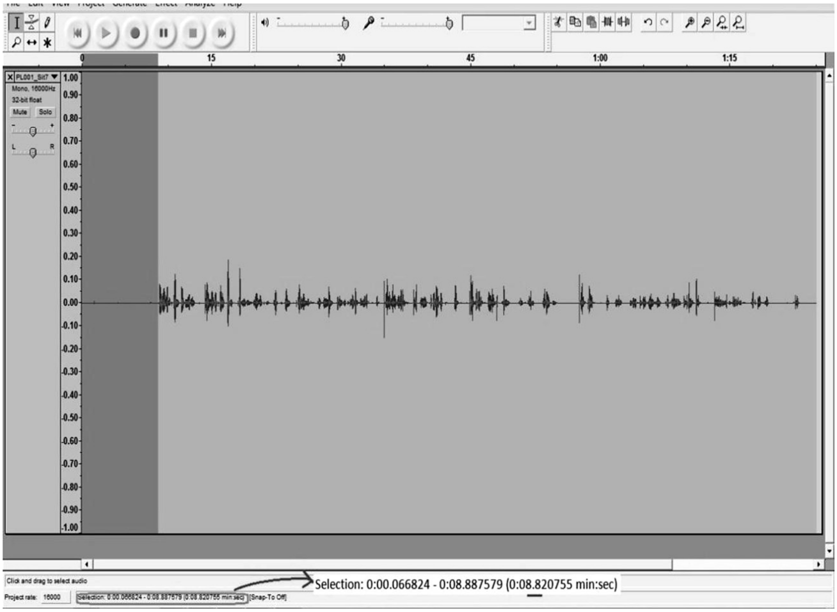
Pilot testing

The Java-programmed oral DCT (see the Appendix for the test script) was pilot-tested with eight Chinese and two Korean volunteers from an undergraduate introductory composition class at a USA university to evaluate the usability of the L2 pragmatic measure. These students' feedback provided a little insight into ESL learners' perception of the difficulty levels of the pragmatic tasks, the clarity of the situation descriptions, and the functionality of the Java test program. The comments they provided included that the 'pen' and 'TV remote' tasks were substantially less difficult than the 'manager' and 'professor' situations because the first two situations were much less awkward and were even described as 'real-world' and 'realistic' by the pilot phase participants. In general, the Java test program was considered straightforward and easy to operate, and the test directions and situation descriptions were clearly written. One problem, though, was that three participants were concerned about whether a one-sentence response would be adequate for either the 'pen' or 'TV remote' situation. To address this concern, two sentences were added to test directions: 'The response to some situations can be less than a couple of sentences. When you feel you have provided a sufficient response, click "stop".'

Data analyses

For the calculation of response latency and speech rate, automatic measurements were taken and recorded by the Java test program of *response latency* (or pre-task planning time) and total response time. Response latency was defined as the lapse in time between stimulus and response, which was normally the gap between when the audio prompt ended with 'you say to . . .' and when the participant pressed the 'record' button. However, it was possible that a participant used up the 120-second preparation time, at which time the 'record' button was automatically activated, accompanied by an audio alert stating 'Recording now' (see the test directions in the Appendix). If this was the case and the participant spoke right after the alert, their total latency time was recorded as 120 seconds. If participants waited a little longer after the audio alert to start speaking – whether or not they had used the entire preparation time – their latency time was calculated as the sum of their additional speech initiation latency time identifiable in Audacity® (a free sound editor and recording software program) and their pre-task planning time automatically captured by the Java test program. Figure 4.1 illustrates how Audacity® helped to identify additional response initiation latency time. The participant in Figure 4.1 paused for 8.8 seconds after the audio alert ‘Recording now’ went off. Response latency of one second or longer beyond planning time was flagged and considered in calculations of latency.

Figure 4.1 Using Audacity® to identify additional response initiation latency time



Selection: 0:00.066824 - 0:08.887579 (0:08.820755 min:sec)

Following Ginther, Dimova and Yang’s (2010) definition, *speech rate* is the number of syllables uttered per minute of response time, which includes the time spent on meaningful utterances and non-meaningful fillers (e.g. ‘um’, ‘uh’) as well as the time spent in silence. Hence, the formula for speech rate:

$$\text{Speech rate} = \frac{\text{total number of syllables uttered}}{\text{total response time}}$$

Before using the formula to compute speech rate in SAS® (a statistical analysis software program), the total number of syllables included in and the total response time spent on each task response were calculated following these procedures.

1. All 80 Chinese participants' spoken responses to the four sociopragmatic situations were transcribed using VoiceWalker, a free transcribing software program.
2. A tally was performed on each task response in relation to the number of syllables uttered.
3. The total response time in each task response was more precisely calibrated in Audacity® with initial silence excluded because these additional moments of silence immediately preceding the speech were already included in the calculation of response latency.

This study utilized a repeated measures (RM) or within-subjects design because all participants recorded their spoken responses to PDR-low as well as PDR-high tasks. RM designs are believed to have the following advantages: 1) They are economical especially when sample members are difficult to recruit; 2) They allow experiments to be conducted more efficiently due to less training of participants; 3) They can reduce the impact of individual differences as a potential confounding factor; 4) They possess more statistical power due to the reduction of unsystematic variability in the design (Keselman Huberty, Lix, Olejnik, Cribbie, Donahue, Kowalchuk, Lowman, Petoskey, Keselman and Levin 1998, Loerts 2008, Stevens 1996).

To examine the interval-scaled dependent variables (i.e. response latency and speech rate), two 2 × 2 × 2 RM-ANOVA tests were performed using a Bonferroni-adjusted significance level (alpha = 0.05/2 = 0.025). Prior to the RM-ANOVAs, data were summarized and examined to ensure that model assumptions were not violated. An RM-MANOVA was opted out because the research questions focused on how each individual dependent variable, not a linear combination of them, would be impacted by pragmatic task features, English proficiency, and learning setting.

Results

Descriptive statistics

Each Chinese participant's response latency scores and speech rates were averaged across the two situation exemplars for either pragmatic task type, PDR-low or PDR-high. For example, a participant's mean response latency in relation to PDR-low tasks was calculated by averaging their response latency scores in the 'pen' and 'TV remote' situations.

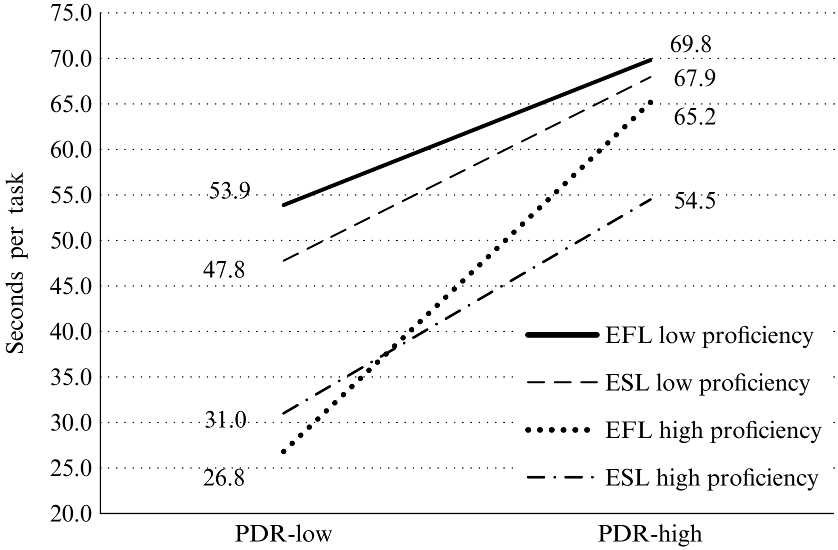
Table 4.3 and Figure 4.2 present the descriptive statistics and a comparative means plot for response latency across the four Chinese groups. Examination of the mean response latency measurements reveals an average increase of between 15.9 seconds and 38.4 seconds for every participant

Table 4.3 Descriptive statistics for response latency by task type

Variable	Participant group	N	Task type	Mean	SD	Min	Max	Lower 99% CI	Higher 99% CI
Response latency (Seconds per task)	EFL low proficiency	20	PDR-low	53.9	29.0	4.0	113.0	35.3	72.5
		20	PDR-high	69.8	41.3	7.5	123.0	43.4	96.2
	EFL high proficiency	20	PDR-low	26.8	22.2	7.5	96.5	12.6	41.0
		20	PDR-high	65.2	34.8	11.0	120.0	42.9	87.4
	ESL low proficiency	20	PDR-low	47.8	29.1	5.5	99.5	29.2	66.4
		20	PDR-high	67.9	45.0	10.0	137.5	39.1	96.7
	ESL high proficiency	20	PDR-low	31.0	17.5	4.5	65.0	19.8	42.2
		20	PDR-high	54.5	26.6	5.5	110.5	37.4	71.5

Note: SD = Standard deviation; CI = Confidence interval

Figure 4. 2 Comparative means plot of response latency across pragmatic task types



group when the pragmatic task switches from PDR-low to PDR-high. Among the groups, the EFL high proficiency group demonstrates the largest increase, i.e. 38.4 seconds, a little more than half a minute. The EFL high proficiency group is clearly distinguishable from the rest of the groups also because its 99% CIs do not overlap between the PDR-low and PDR-high situation types, which indicates that there is a significant difference in the average amount of response latency time spent by this participant group in preparation for a PDR-high task versus a PDR-low task.

Similarly, Table 4.4 and Figure 4.3 present the descriptive statistics and a comparative means plot for speech rate across the four Chinese groups. The mean speech rates in the table and the line graph indicate that the EFL low proficiency group does not register much of a decrease (only five syllables per minute) in speech rate in association with a change in pragmatic task type from PDR-low to PDR-high. Neither does the ESL low proficiency group, which has an average decrease of 14 syllables per minute. The two high proficiency groups, however, show a decrease of 25 and 26 syllables in response to a task change from PDR-low to PDR-high.

RM-ANOVA tests

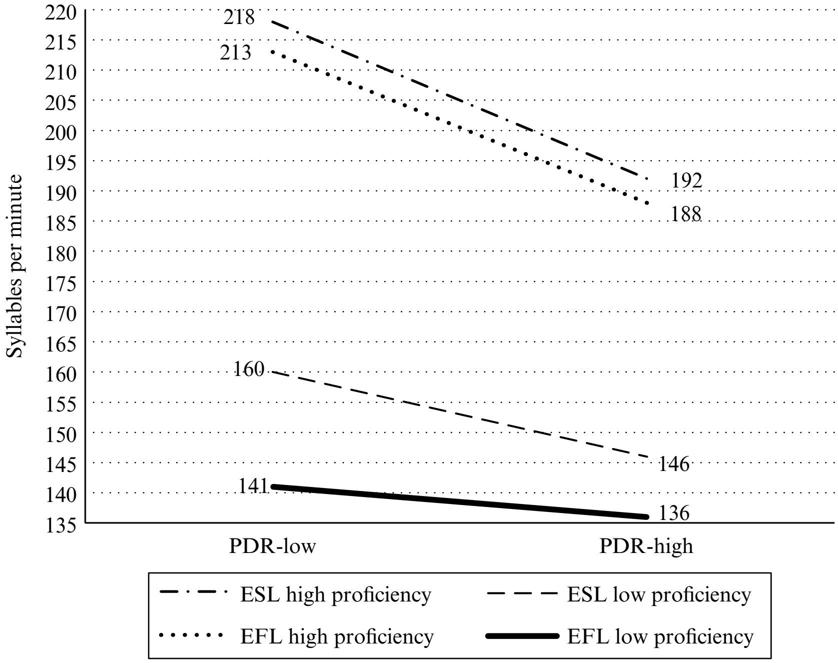
RM-ANOVAs hold the same model assumptions as ordinary ANOVAs: normality, homogeneity of variance and covariance matrices, and independence

Table 4.4 Descriptive statistics for speech rate by task type

Variable	Participant group	N	Task type	Mean	SD	Min	Max	Lower 99% CI	Higher 99% CI
Speech rate (Number of syllables per minute) (Rounded to the nearest integer)	EFL low proficiency	20	PDR-low	141	31	87	192	121	161
		20	PDR-high	136	25	89	189	120	152
	EFL high proficiency	20	PDR-low	213	32	155	294	192	234
		20	PDR-high	188	20	157	228	175	200
	ESL low proficiency	20	PDR-low	160	38	100	248	135	184
		20	PDR-high	146	33	81	214	125	167
	ESL high proficiency	20	PDR-low	218	38	153	315	193	242
		20	PDR-high	192	31	126	250	172	212

Note: SD = Standard deviation; CI = Confidence interval

Figure 4.3 Comparative means plot of speech rate for the four participant groups across pragmatic task types



of observations (Weinfurt 2000). The last two assumptions were valid, whereas there were minor deviations from normality in the response latency distributions. Fortunately, RM-ANOVA is generally considered robust to violations of the normality assumption (Maxwell and Delaney 1990, Stevens 1996).

A three-way RM-ANOVA was conducted to compare the effects of pragmatic task features (i.e. PDR), English proficiency, and learning setting on response latency and the speech rate of participants' spoken production of requests. The statistical results that appear in Tables 4.5 and 4.6 lead to the following findings in response to the research questions.

First, pragmatic task features (PDR) have significant, large effects on the response latency and speech rate of spoken request production: for response latency: $F\,(1, 76) = 71.48$, $p = 0.00$, $\eta_p{}^2 = 0.49$, for speech rate: $F\,(1, 76) = 27.54$, $p = 0.00$, $\eta_p{}^2 = 0.27$. PDR-high tasks tend to elicit performances formulated after longer response latency (i.e. pre-task planning time) and characteristic of slower speech rate. The effect sizes (partial eta squared) indicate that 49% of the within-subjects variance in response latency is accounted for by PDR and 27% of the within-subjects variance is accounted for by PDR as far as speech rate is concerned.

Table 4.5 Repeated measures Analysis of Variance on response latency

Source	Num. SS	Num. df	Num. MS	Den. MS	Den. df	F	*p*	η^2_p	Power
PDR (within-subjects)	23961.03	1	23961.03	335.23	76	71.48	0.00**	0.49	1.00
PDR × learning setting	283.56	1	283.56	335.23	76	0.85	0.36	0.01	0.09
PDR × proficiency	1664.10	1	1664.10	335.23	76	4.96	0.03	0.06	0.48
PDR × learning setting × proficiency	916.81	1	916.81	335.23	76	2.74	0.10	0.04	0.27
Learning setting (between-subjects)	525.63	1	525.63	1698.79	76	0.31	0.58	0.00	0.05
Proficiency (between-subjects)	9594.51	1	9594.51	1698.79	76	5.65	0.02**	0.07	0.54
Learning setting × proficiency	5.63	1	5.63	1698.79	76	0.00	0.95	0.00	0.03

*Note: Num. SS = Numerator Sum of Squares; Num. df = Numerator degrees of freedom; Num. MS = Numerator Mean Squares; Den. MS = Denominator Mean Squares; Den. df = Denominator degrees of freedom; **Significant at a Bonferroni-adjusted significance level (alpha = 0.05/2 = 0.025)*

Table 4.6 Repeated measures Analysis of Variance on speech rate

Source	Num. SS	Num. df	Num. MS	Den. MS	Den. df	F	*p*	η^2_p	Power
PDR (within-subjects)	11840.87	1	11840.87	429.96	76	27.54	0.00**	0.27	1.00
PDR × learning setting	196.49	1	196.49	429.96	76	0.46	0.50	0.01	0.06
PDR × proficiency	2488.67	1	2488.67	429.96	76	5.79	0.02**	0.07	0.55
PDR × learning setting × proficiency	176.41	1	176.41	429.96	76	0.41	0.52	0.01	0.06
Learning setting (between-subjects)	3593.43	1	3593.43	1576.26	76	2.28	0.14	0.03	0.23
Proficiency (between-subjects)	129259.83	1	129259.83	1576.26	76	82.00	0.00**	0.52	1.00
Learning setting × proficiency	943.20	1	943.20	1576.26	76	0.60	0.44	0.01	0.07

*Note: Num. SS = Numerator Sum of Squares; Num. df = Numerator degrees of freedom; Num. MS = Numerator Mean Squares; Den. MS = Denominator Mean Squares; Den. df = Denominator degrees of freedom; **Significant at a Bonferroni-adjusted significance level (alpha = 0.05/2 = 0.025)*

Second, there are differences in the mean response latency and mean speech rate between low and high proficiency learners: for response latency: $F\,(1, 76) = 5.65$, $p = 0.02$, $\eta_p^2 = 0.07$, and for speech rate: $F\,(1, 76) = 82.00$, $p = 0.00$, $\eta_p^2 = 0.52$. Compared with low proficiency participants, high proficiency learners need less response latency time and can produce requests with a faster speech rate. In addition, proficiency has differing magnitudes of effects on response latency and speech rate: it has a significant, medium effect on response latency and is able to explain 7% of the variance in the main effect of proficiency and its associated error as far as response latency is concerned; however, in regard to speech rate, proficiency has a very strong effect on speech rate and is able to explain 52% of the variance in the main effect of proficiency and its associated error.

Third, learning setting does not have a significant effect on either response latency or speech rate. The influence of PDR on response latency or speech rate does not depend on learning setting, nor does the influence of English proficiency depend on learning setting.

Finally, the influence of pragmatic task features (PDR) on speech rate depends on English proficiency, and vice versa: $F\,(1, 76) = 5.79$, $p = 0.02$, $\eta_p^2 = 0.07$. This significant, medium interaction effect between PDR and proficiency suggests that high proficiency learners demonstrate a larger decrease in speech rate as compared to low proficiency participants in association with a task change from PDR-low to PDR-high.

Discussion

Based on the results from this study, a pragmatic perspective seems to hold great promise for the ongoing effort to develop an effective working model for task effects in second language assessment. This study supports previous research (e.g. Fulcher and Márquez Reiter 2003, Taguchi 2007) in that pragmatic features can be manipulated to design speaking tasks with varying levels of cognitive demand for L2 or foreign language (FL) learners in a testing situation. The three pragmatic variables (i.e. Power (P), Distance (D), and Rank of imposition (R)) could be combined into eight different types of pragmatic tasks (e.g. P+D+R+, P+D+R−, P+D−R+, P+D−R−, P−/=D+R+, P−/=D+R−, P−/=D−R+, P−/=D−R−). The significant strong effect of PDR (with the two 'extreme' conditions being P+D+R+ and P−/=D−R−) on the temporal measures of speech formulation and production suggests that these pragmatic task features can set off a psycholinguistic trigger and, thus, affect information processing and speech act production.

The selection of response latency and speech rate as the dependent variables in this study helped to unfold a relatively full picture about the possible impact of task characteristics on the automaticity of speech act production. Response latency is a measure of task information processing and offline planning for task performance, whereas speech rate reflects online processing because its two closely related traits, speaking speed and pausing, are both connected with planning for the utterances to come. In addition, response latency and speech rate are worth including in empirical studies of productive pragmatic abilities because of their simplicity and facility for data collection.

However, learning setting was found not to have a significant main effect or an interaction effect with PDR or proficiency, despite the originality of collecting data from ESL learners in the USA and EFL learners in China with the belief that the development of pragmatic abilities requires situated learning in a context where English is spoken. The insensitivity of response latency and speech rate to learning setting differences suggests a need for closer examination of the production of requests from a pragmatic point of view, for example, through human ratings of pragmatic appropriateness. This subject is one that I hope to investigate in the next research study.

Last, but not least, the interaction between PDR and learner proficiency indicates that PDR-high tasks elicit request production with a greater decrease in speech rate from high proficiency L1 Chinese learners of English than low proficiency learners from the same L1 group. This suggests that PDR-high tasks would distinguish among high proficiency learners well and that including pragmatic items in L2 speaking assessments might be a promising direction for extending the high end of measurement of L2 speaking abilities. Constructing difficult but fair items at the high end of a measurement scale has always been a challenge for test developers. Reliable assessment of pragmatic ability would seem to be an area that can be expected to contribute to the validation of L2 speaking abilities at the high end of measurement and allow test developers to extend the current L2 speaking test scales.

References

Austin, J L (1962) *How to Do Things with Words*, Oxford: Oxford University Press.

Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bachman, L F (2002) Some reflections on task-based language performance assessment, *Language Testing* 19 (4), 453–476.

Bachman, L F and Palmer, A (1996) *Language Testing in Practice*, Oxford: Oxford University Press.

Bardovi-Harlig, K (1999) Exploring the interlanguage of interlanguage pragmatics, *Language Learning* 99, 677–713.

Bardovi-Harlig, K and Hartford, B S (1990) Congruence in native and nonnative conversations: Status balance in the academic advising session, *Language Learning* 40, 467–501.

Bardovi-Harlig, K and Hartford, B S (1993) Learning the rules of academic talk: A longitudinal study of pragmatic development, *Studies in Second Language Acquisition* 15, 279–304.

Brown, J D, Hudson, T D, Norris, J M and Bonk, W (2002) *An Investigation of*

Second Language Task-Based Performance Assessments, Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Edwards, M (no date) *Spot the Problem!* available online: www.indiana.edu/~dsls/publications/Edwardsedit.pdf

Elder, C, Iwashita, N and McNamara, T (2002) Estimating the difficulty of oral proficiency tasks, *Language Testing* 19 (4), 347–368.

Faul, F, Erdfelder, E, Lang, A-G and Buchner, A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, *Behavior Research Methods* 39, 175–191.

Foster, P and Skehan, P (1996) The influence of planning and task type on second language performance, *Studies in Second Language Acquisition* 18 (3), 299–323.

Fulcher, G and Márquez Reiter, R (2003) Task difficulty in speaking tests, *Language Testing* 20 (3), 321–344.

Ginther, A, Dimova, S and Yang, R (2010) Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring, *Language Testing* 27 (3), 379–399.

Iwashita, N, McNamara, T and 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.

Keselman, H J, Huberty, C J, Lix, L M, Olejnik, S, Cribbie, R A, Donahue, B, Kowalchuk, R K, Lowman, L L, Petoskey, M D, Keselman, J C and Levin, J R (1998) Social practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analysis, *Review of Educational Research* 68 (3), 350–386.

Kitao, K (1990) A study of Japanese and American perceptions of politeness in requests, *Doshisha Studies in English* 50, 178–210.

Lennon, P (1990) Investigating fluency in EFL: A quantitative approach, *Language Learning* 40, 387–412.

Levelt, W J M (1989) *Speaking: From Intention to Articulation*, Cambridge: MIT Press.

Loerts, H (2008) *Multivariate ANOVA & Repeated Measures*, available online: www.let.rug.nl/~nerbonne/teach/rema-stats-meth-seminar/presentations/Loerts-2008-MANOVA-Repeated-Measures.pdf

Maxwell, S E and Delaney, H D (1990) *Designing Experiments and Analyzing Data*, Belmont: Wadsworth Publishing Company.

Norris, J M, Brown, J D, Hudson, T D and Yoshioka, J (1998) *Designing Second Language Performance Assessments*, Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Purdue University Oral English Proficiency Program (2013) *OEPT technical manual*, available online: www.purdue.edu/oepp/documents/OEPT%20Technical%20Manual.pdf

Robinson, P (1995) Task complexity and second language narrative discourse, *Language Learning* 45, 141–175.

Robinson, P (2001) Task complexity, task difficulty, and task production: Exploring interactions in a componential framework, *Applied Linguistics* 22 (1), 27–57.

Robinson, P (2011) (Ed) *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*, Amsterdam: John Benjamins Publishing Company.

Roever, C (2001) *A Web-Based Test of Interlanguage Pragmalinguistic*

Knowledge: Speech Acts, Routines, Implicatures, unpublished PhD thesis, University of Hawaii at Manoa.

Roever, C (2011) Testing of second language pragmatics: Past and future, *Language Testing* 28 (4), 463–481.

Skehan, P (1998) *A Cognitive Approach to Language Learning*, Oxford: Oxford University Press.

Stevens, J (1996) *Applied Multivariate Statistics for the Social Sciences* (3rd edition), Hillsdale: Erlbaum.

Taguchi, N (2007) Task difficulty in oral speech act production, *Applied Linguistics* 28 (1), 113–135.

Takahashi, T and Beebe, L M (1993) Cross-linguistic influence in the speech act of correction, in Blum-Kulka, S and Kasper, G (Eds) *Interlanguage Pragmatics*, New York: Oxford University Press, 138–157.

Tavakoli, P and Foster, P (2008) Performance: The effect of narrative type on learner output, *Language Learning* 58 (2), 439–473.

Thomas, J (1983) Cross-cultural pragmatic failure, *Applied Linguistics* 4, 91–112.

Towell, R, Hawkins, R and Bazergui, N (1996) The development of fluency in advanced learners of French, *Applied Linguistics* 17, 84–119.

Weinfurt, K P (2000) Repeated measures analyses: ANOVA, MANOVA, and HLM, in Grimm, L G and Yarnold, P R (Eds) *Reading and Understanding More Multivariate Statistics*, Washington, DC: American Psychological Association, 317–361.

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.

Appendix

The oral DCT administered to all Chinese ESL/EFL participants

Note: The computer-based oral English DCT for Chinese participants consisted of eight situations including four situations that were not directly related to pragmatic politeness. All situational descriptions were translated into Chinese as well.

Welcome to the oral completion test!

Sound check

To begin, we would like you to do a sound check of the equipment. Wear the headset so that the microphone is on the left-hand side. Please put the microphone close to the side of your mouth – about one inch away will be fine. Whenever you are ready, press RECORD and say something like ‘I am participating in a research study. I hope this doesn’t take too long.’

After you stop recording, press PLAY to check if you were recorded. You will NOT be able to play back your recordings at any other time during the test.

If there is no problem with the equipment, you may continue with the test. However, if there was no recording or the quality of your recording was seriously reduced by static, try adjusting the position of the microphone or the level of your voice, and test again. You may also adjust the sound volume of the test, but be careful not to mute the test.

If you have any problems as you record and listen to the test, raise your hand and wait for the test administrator to assist you.

Directions

In this test, you will see and hear descriptions of eight different situations before you record what you would say if you were in the situations described. The recording of each situation description will be played only once. After the recording is finished, you will have 2 minutes or 120 seconds to formulate your response. However, you are highly encouraged to start recording your response whenever you feel you are ready by clicking RECORD. If you use all the 2-minute preparation time, the recording will begin automatically. Whenever a recording begins, you will be alerted with a prompt that states, 'RECORDING NOW'.

The response time is limited to two minutes or 120 seconds. The time you have to record your response will be presented at the top right-hand corner of the screen. YOU DO NOT HAVE TO SPEAK FOR THE ENTIRE 2 MINUTES. The response to some situations can be less than a couple of sentences. When you feel you have provided a sufficient response, click STOP.

Then click CONTINUE to move on to the next item.

Situation 1 of 8

You have an English exam tomorrow. You are studying in the school library to prepare for the exam. A good friend of yours is also studying in the library, sitting just a few seats away from you. Your pen just ran out of ink, so you go over to your friend and want to ask her to lend you a pen.

You say to your friend:

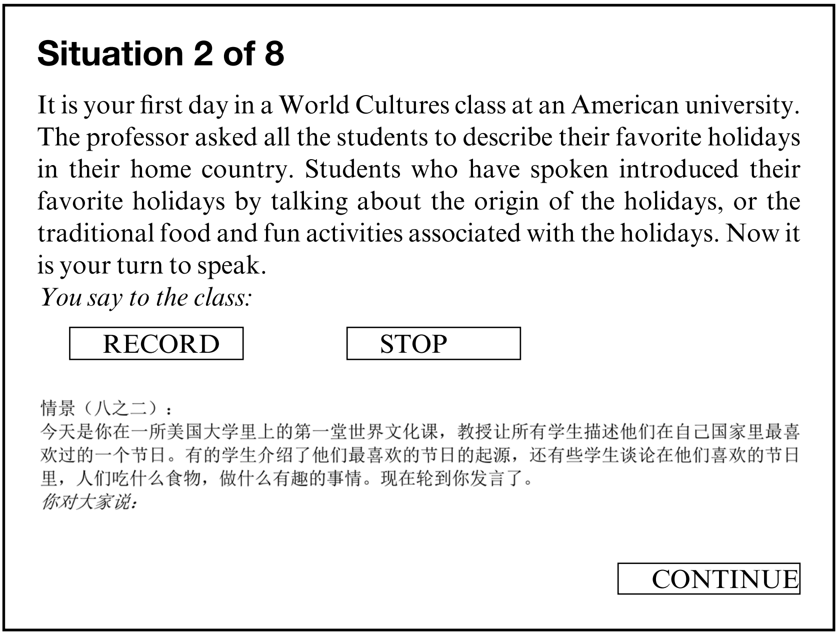
RECORD　　STOP

情景（八之一）：

你明天有英语考试。你在学校图书馆学习，准备明天的考试。你看到你的好朋友也在图书馆学习，就坐在离你几个座位远的地方。你的笔刚刚没有墨水了，你向那位朋友走过去，想跟她借支笔。

你对朋友说：

CONTINUE



Situation 2 of 8

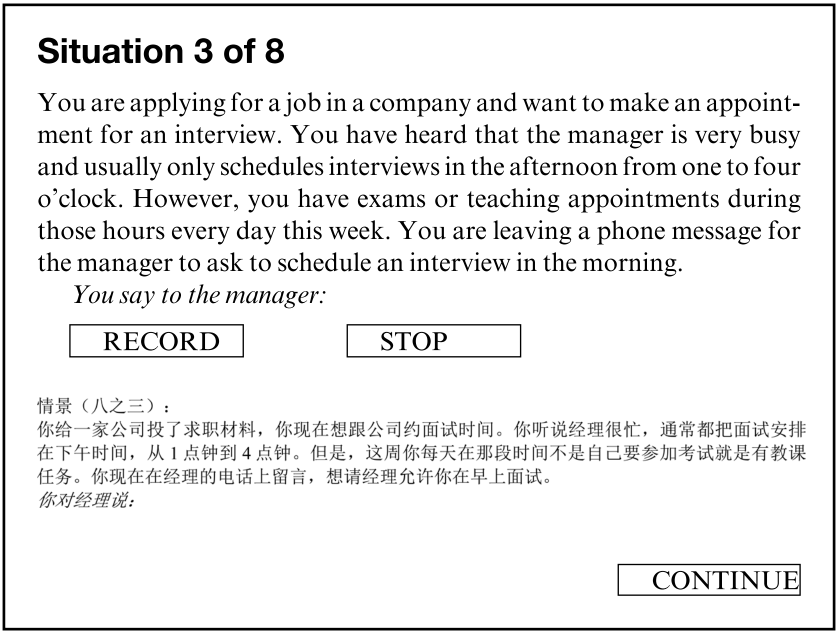
It is your first day in a World Cultures class at an American university. The professor asked all the students to describe their favorite holidays in their home country. Students who have spoken introduced their favorite holidays by talking about the origin of the holidays, or the traditional food and fun activities associated with the holidays. Now it is your turn to speak.

You say to the class:

情景（八之二）：

今天是你在一所美国大学里上的第一堂世界文化课，教授让所有学生描述他们在自己国家里最喜欢过的一个节日。有的学生介绍了他们最喜欢的节日的起源，还有些学生谈论在他们喜欢的节日里，人们吃什么食物，做什么有趣的事情。现在轮到你发言了。

你对大家说：



Situation 3 of 8

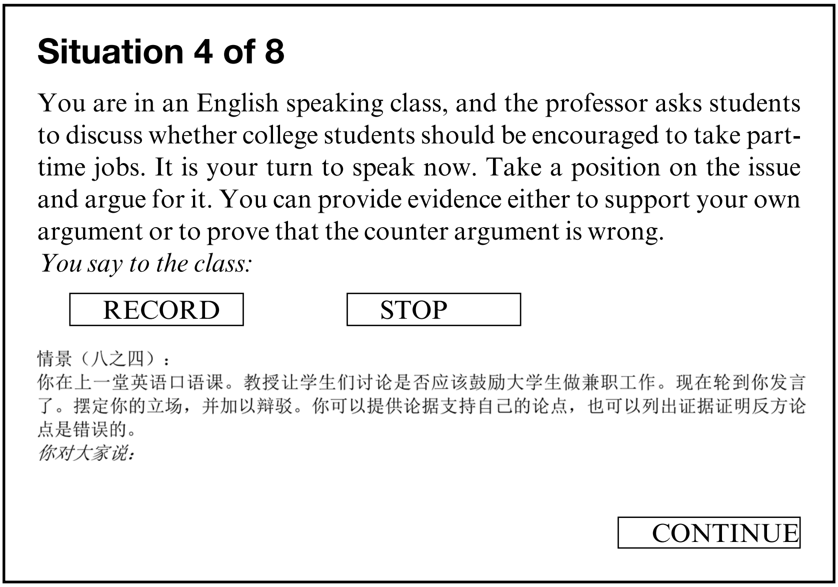
You are applying for a job in a company and want to make an appointment for an interview. You have heard that the manager is very busy and usually only schedules interviews in the afternoon from one to four o’clock. However, you have exams or teaching appointments during those hours every day this week. You are leaving a phone message for the manager to ask to schedule an interview in the morning.

You say to the manager:

情景（八之三）：

你给一家公司投了求职材料，你现在想跟公司约面试时间。你听说经理很忙，通常都把面试安排在下午时间，从 1 点钟到 4 点钟。但是，这周你每天在那段时间不是自己要参加考试就是有教课任务。你现在在经理的电话上留言，想请经理允许你在早上面试。

你对经理说：



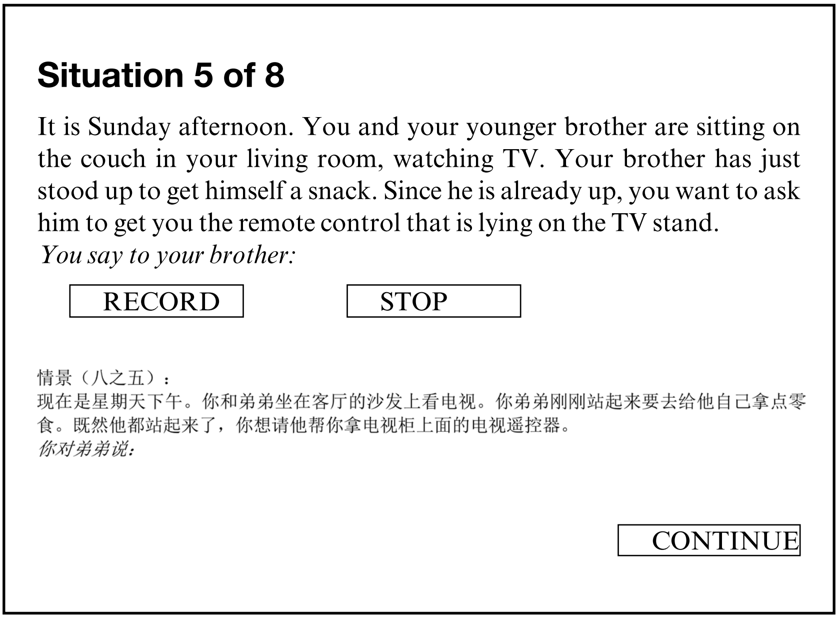
Situation 4 of 8

You are in an English speaking class, and the professor asks students to discuss whether college students should be encouraged to take part-time jobs. It is your turn to speak now. Take a position on the issue and argue for it. You can provide evidence either to support your own argument or to prove that the counter argument is wrong.
You say to the class:

RECORD STOP

情景（八之四）：
你在上一堂英语口语课。教授让学生们讨论是否应该鼓励大学生做兼职工作。现在轮到你发言了。摆定你的立场，并加以辩驳。你可以提供论据支持自己的论点，也可以列出证据证明反方论点是错误的。
你对大家说：

CONTINUE



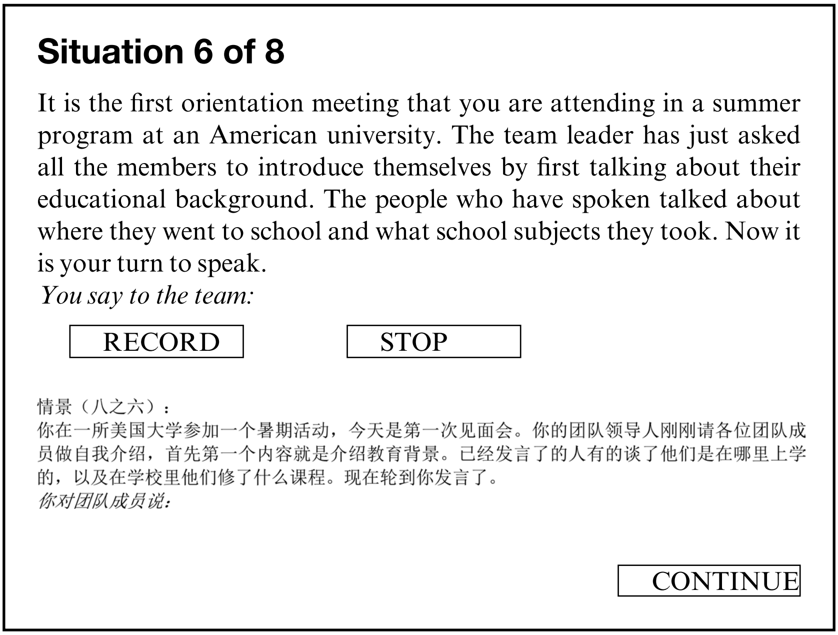
Situation 5 of 8

It is Sunday afternoon. You and your younger brother are sitting on the couch in your living room, watching TV. Your brother has just stood up to get himself a snack. Since he is already up, you want to ask him to get you the remote control that is lying on the TV stand.
You say to your brother:

RECORD STOP

情景（八之五）：
现在是星期天下午。你和弟弟坐在客厅的沙发上看电视。你弟弟刚刚站起来要去给他自己拿点零食。既然他都站起来了，你想请他帮你拿电视柜上面的电视遥控器。
你对弟弟说：

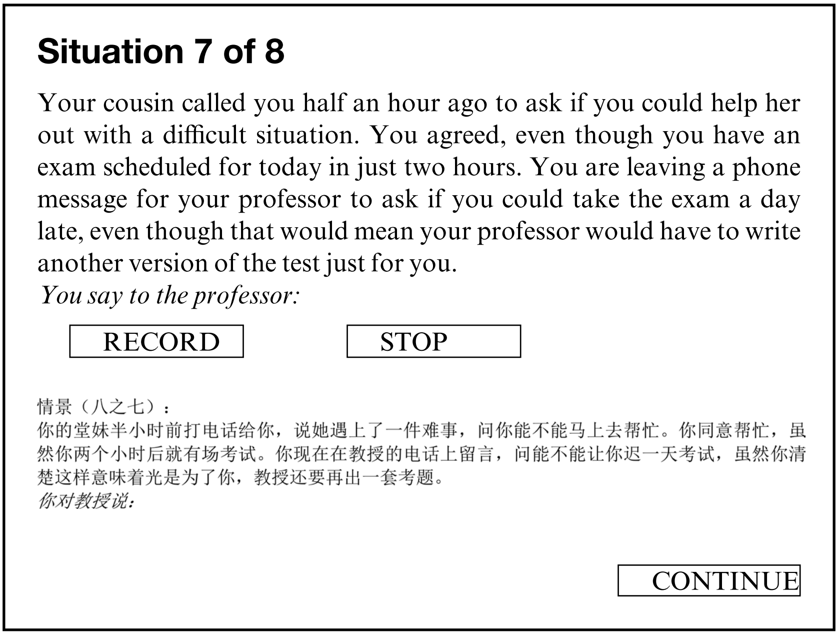
CONTINUE



Situation 6 of 8

It is the first orientation meeting that you are attending in a summer program at an American university. The team leader has just asked all the members to introduce themselves by first talking about their educational background. The people who have spoken talked about where they went to school and what school subjects they took. Now it is your turn to speak.
You say to the team:

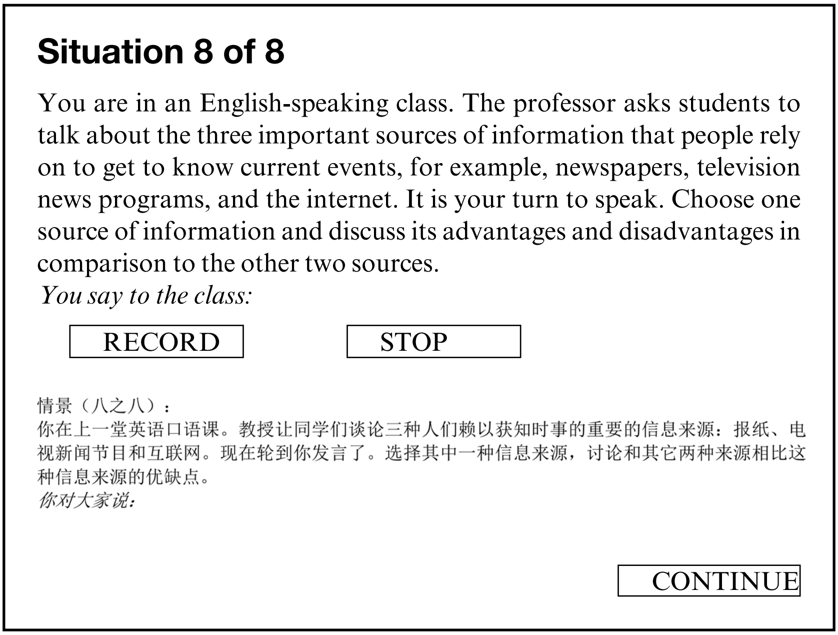
情景（八之六）：
你在一所美国大学参加一个暑期活动，今天是第一次见面会。你的团队领导人刚刚请各位团队成员做自我介绍，首先第一个内容就是介绍教育背景。已经发言了的人有的谈了他们是在哪里上学的，以及在学校里他们修了什么课程。现在轮到你发言了。
你对团队成员说：



Situation 7 of 8

Your cousin called you half an hour ago to ask if you could help her out with a difficult situation. You agreed, even though you have an exam scheduled for today in just two hours. You are leaving a phone message for your professor to ask if you could take the exam a day late, even though that would mean your professor would have to write another version of the test just for you.
You say to the professor:

情景（八之七）：
你的堂妹半小时前打电话给你，说她遇上了一件难事，问你能不能马上去帮忙。你同意帮忙，虽然你两个小时后就有场考试。你现在在教授的电话上留言，问能不能让你迟一天考试，虽然你清楚这样意味着光是为了你，教授还要再出一套考题。
你对教授说：



Situation 8 of 8

You are in an English-speaking class. The professor asks students to talk about the three important sources of information that people rely on to get to know current events, for example, newspapers, television news programs, and the internet. It is your turn to speak. Choose one source of information and discuss its advantages and disadvantages in comparison to the other two sources.

You say to the class:

情景（八之八）：

你在上一堂英语口语课。教授让同学们谈论三种人们赖以获知时事的重要的信息来源：报纸、电视新闻节目和互联网。现在轮到你发言了。选择其中一种信息来源，讨论和其它两种来源相比这种信息来源的优缺点。

你对大家说：

5 ESL teachers’ versus American undergraduates’ judgments of international teaching assistants’ accentedness, comprehensibility, and oral proficiency

Ching-Ni Hsieh

Educational Testing Service, New Jersey, US

Motivation for the research

The research on which I report in this chapter is situated within a larger study on *rater variability*. The term rater variability refers to variations in scores that raters give that are associated with rater characteristics but not with examinees’ actual performance or abilities (Engelhard and Myford 2003, McNamara 1996, Myford and Wolfe 2000). Rater variability in second language (L2) oral performance assessment is, among other factors, a function of rater experience and expectations within the context of a specific assessment. Variability due to rater characteristics, or rater effects, may adversely affect decision-making processes, particularly in high-stakes testing situations (Barrett 2001, Engelhard and Myford 2003, Schaefer 2008). Crucially, these rater effects introduce construct-irrelevant variance in the assessment process, and potentially obscure the construct being measured (Congdon and McQueen 2000). Construct-irrelevant variance refers to any factors that can affect test performance and have nothing to do with the construct being measured (Ferrier, Lovett and Jordan 2011). Rater effects, therefore, can compromise the validity and fairness of performance assessments (Kunnan 2005).

This study examined rater effects related to the evaluations of English oral proficiency, degree of foreign accent, and perceived comprehensibility of potential international teaching assistants (ITAs) at a large North American university (University M will be used as a pseudonym in this report). For the purposes of this study, *oral proficiency* is defined as an ITA candidate’s global communicative competence to function at an instructional setting in USA higher educational institutions (Douglas and Smith 1997). The construct definitions of accentedness and comprehensibility adopted in this study follow Munro and Derwing's work on L2 speech perception and production (Derwing and Munro 1997, Munro and Derwing 1995, 2001). *Accentedness*, as defined in Derwing and Munro (2009:478) refers to 'how different a pattern of speech sounds compared to the local variety'. *Comprehensibility* is defined as the listeners' estimation of how easy or difficult it is to understand a given speaker (Derwing and Munro 1997, Munro and Derwing 1995). The screening of ITAs, henceforth referred to as *ITA testing*, should be considered when examining rater effects involved in the rating process because the screening of qualified ITAs whose English proficiency is sufficient for instruction and whose pronunciation is comprehensible to linguistically naïve undergraduates is important for the purposes of supporting undergraduate learning. Rater effects for two groups of raters were of particular interest: 1) trained English as a Second Language (ESL) teachers, whose opinions are important for hiring or screening ITAs; and 2) novice American undergraduate raters, whose educational outcomes may be affected by their abilities to understand ITAs.

Of critical importance for ITAs and undergraduates are the ITA test raters, whose judgments of ITA language abilities are crucial for ITA employment decisions. Intrinsic to ITA testing is the assumption that the official raters are acting on behalf of the undergraduate student population at their institution, in other words, the pool of students who might constitute any of the classes taught by ITAs. The idea is that if raters judge potential ITAs' speech as sufficient in terms of overall proficiency or comprehensibility, then international examinees may assume that their linguistic abilities are appropriate for the role they plan to assume. From the viewpoint of the stakeholders (the university and the undergraduates), the idea is that if the assessment and rating criteria are valid, raters' accurate judgments may facilitate harmonious educational experiences for ITAs and undergraduates. Inaccurate judgments may result in frustrating and potentially detrimental situations for all involved; therefore, due to the high-stakes nature of ITA oral language assessment, universities with ITA testing programs should periodically check that their official raters judge potential ITAs' speech on a par with how undergraduates would rate them. Discrepancies in how the two groups (i.e. the official ITA raters and the undergraduates) respond to potential ITAs' oral language abilities should be investigated because significant differences in responses may impact the validity of ITA testing programs.

Review of the literature

Differences between professional raters and undergraduates in judging ITA speaking abilities constitute a source of rater effects. Rater effects, such as rater severity or leniency, are often viewed as sources of systematic variance in ratings that are associated with raters and not with the examinees (Eckes 2005, Hoyt 2000, Myford and Wolfe 2003). Research that compares ESL teaching professionals' and undergraduates' evaluations of ITA speech has found conflicting results (Bejar 1985, Clarke and Swinton 1980, Powers, Shedl, Wilson-Leung and Butler 1999). Orth (1983) found a weak correlation between ratings of oral proficiency awarded by undergraduates and ESL teachers ($r = 0.12$). Similarly, Oppenheim (1998) found that the relationship between ESL teachers' ratings of ITAs' English speaking abilities and undergraduate raters' ratings of ITAs' linguistic skills was moderately low ($r = 0.33$). On the contrary, Saif (2002) found that linguistically naïve, untrained undergraduate raters were able to rate ITAs' oral performances as consistently as ESL teaching experts did. Taken together, previous studies have shown that high levels of rater variability are present in the judgments of ITA speech among raters of different linguistic backgrounds.

In a number of studies that examine the concepts of accent and comprehensibility (Derwing and Munro 1997, 2005, 2009, Munro and Derwing 1995, 2001), native listeners were asked to make scalar judgments on 9-point, holistic rating scales for accentedness (i.e. 1 = no accent, 9 = heavily accented) and for comprehensibility (i.e. 1 = very easy to understand, 9 = extremely difficult or impossible to understand). Results of these studies have collectively established the reliability of accent and comprehensibility ratings awarded by both ESL teaching professionals and untrained undergraduate listeners (Derwing and Munro 1997, Munro and Derwing 1995).

To this day, the search for an ideal measure of accent continues to be fraught with problems due to both the complexity and vagueness of the construct. Some researchers approached the judgments of foreign accent by using computer-assisted instrumental analysis (e.g. Kang 2012, Pickering 1999). Given that the goal of this study was to understand differences in ESL teachers' and undergraduates' perceptions and judgments of ITA speech, Munro and Derwing's conceptualizations of accentedness and comprehensibility (1995) were thought to prove useful as working definitions for rater judgments because an exact level of accentedness was not central to answering the research questions. Rather, a measure that would afford comparisons between teacher and student ratings was needed. The measure proposed by Derwing and Munro (2009) seemed very appropriate for this purpose and the rating scales they employed were adopted for use in the current study.

A large body of research in applied linguistics has examined the characteristics of ITA speech, specifically, and foreign-accented speech, in general, to understand how different groups of listeners perceive such accented speech. These studies suggest that factors such as speech rate (Munro and Derwing 1998), discourse-level language use (Davies, Tyler and Koran 1989, Tyler 1992), intonation and tone (Kang 2010, Pickering 2001), and accent familiarity (Rubin and Smith 1990) all contribute to undergraduates' comprehension difficulties of ITA speech in different ways. A heavy foreign accent in particular, as many studies have shown (Bailey 1983, Fox and Gay 1994, Plakans 1997), has often been deemed as the cause of ITAs' poor communication skills and as the main source of undergraduates' comprehension difficulty.

Research questions

There has been little research on how raters are affected by different subcomponents of L2 speech within specific ITA testing contexts (Kang 2012). Issues involved in ITA oral proficiency and speech comprehensibility are important factors and considerations for all the stakeholders involved and more research is needed to investigate these issues. The study reported here aimed to fill this gap and addressed the following research questions:

1. Do ESL teachers and American undergraduate students evaluate ITA candidates' oral proficiency, accentedness, and comprehensibility differently? If so, to what extent?
2. What factors in ITA candidates' speech draw raters' attention when they evaluate ITA candidates' oral proficiency? Are different factors more or less salient to different rater groups?

Data collection procedures

Participants

Two rater groups participated in this study. The first group included 13 ESL teachers (five males and eight females). They were the official raters for the Speaking Proficiency English Assessment Kit (SPEAK), the test used to screen ITAs at University M. The teachers were between 29 and 56 years old, with a mean age of 39.9. All teachers had academic backgrounds in language education or linguistics and experience teaching ESL at a level similar to the SPEAK examinees. Their teaching experience ranged from six to 22 years, with a mean age of 12.5 years. They have been the SPEAK raters for one to 20 years, with a mean length of 4.5 years.

The second group consisted of 32 American undergraduate students who were native English speakers from a wide variety of academic programs (nine males and 23 females). The undergraduate raters were between 18 and 22 years old, with a mean age of 20.1 years. All undergraduates were born and grew up in the Upper Midwest, representing the typical makeup of the undergraduate population at University M. All undergraduates reported having limited exposure to nonnative speech during their upbringing and in their friend circles. They all had experiences taking courses taught by ITAs whose first languages were either Chinese or Korean. Few reported having ITAs who were Arabic, Japanese, Hindi, or Spanish speakers. None reported having prior experience rating nonnative English speakers' speech. All reported having normal hearing.

Examinees

Examinees were 28 international graduate students, seeking ITA positions at University M. There were 19 males and nine females (10 Chinese, 10 Korean, and eight Arabic native speakers). The examinees' oral responses to the SPEAK test during operational SPEAK test administrations were provided by the Testing Office at University M. Their official SPEAK scores ranged from 40 to 55, which represented the typical score distributions of SPEAK test takers at University M. The cut score for qualified ITAs is 50. A score of 40 means a clear fail and a score of 45 is a provisional pass, meaning that candidates can appeal and request for reconsideration of the assigned scores.

Rating materials

Three of the 12 tasks in the SPEAK test were chosen for rating: a picture description, a topic discussion, and a presentation on a revised schedule. The entire response time of these three tasks was approximately 4 minutes, which was deemed sufficient for reliable judgment of ITA speech based on the pilot study. All 45 raters rated all 84 of the speech samples (i.e. 28 examinees times three tasks) on three rating scales: accentedness, comprehensibility, and oral proficiency.

Rating scales

Raters judged examinee performances using three sets of rating scales. The first one was the 5-point holistic SPEAK rating scale and was used to assess examinees' oral proficiency. Raters utilized this scale, ranging from 20 to 60 (20 = no effective communication or no evidence of ability to perform the task; 60 = communication almost always effective, or task performed very competently) with a 10-point increment. The ratings indicated raters' evaluations of an examinee's overall task performance with respect to each task. The second and third rating scales were the 9-point holistic scales employed by Munro and Derwing (1995): for accentedness, 1= no accent while 9 = heavily accented; for comprehensibility, 1= very easy to understand while 9 = extremely difficult or impossible to understand.

Procedures

All raters participated in a pre-experimental session. They learned about the purpose of the research project, the research design, construct definitions of oral proficiency (i.e. accentedness and comprehensibility) and about the rating scales. Raters first completed a background questionnaire, followed by ratings of the speech samples. Because the ESL teachers were all trained raters, no rater training or norming session was undertaken. The undergraduate raters completed a one-on-one training session with the researcher. Each training session consisted of acquainting the raters with the rating tasks and the rating rubrics, and lasted approximately 1 hour. No calibration session was given to the undergraduate raters in order to capture the novice raters' rating behaviors.

Raters evaluated examinee performances online. The examinee order was randomized across tasks and raters. Raters were instructed to rate the recordings in a quiet room that had internet access. They were allowed to listen to each recording multiple times if they considered it necessary. They listened to each recording and assigned scores on oral proficiency first. Immediately after they assigned an oral proficiency rating to a response, they optionally provided written comments regarding their rating decisions. The SPEAK rating rubric was provided as a reference. They then assigned ratings on accentedness and comprehensibility. Rating the three tasks and providing written comments took between 4 and 6 hours. The researcher conducted retrospective interviews with each rater within a couple of days after the completion of the rating experience to explore the factors that raters took into consideration in making rating decisions. Each interview lasted between 30 minutes and 1 hour.

Data analysis

The rating data were analyzed using Many Facet Rasch Measurement (MFRM) analysis, using the computer program FACETS (Version 3.67, Linacre 2010). MFRM is particularly useful in the context of this investigation because it allows for comparisons of how groups differentially respond to the same sets of test data. The MFRM model implemented included four facets: examinees, raters, tasks, and rater status (ESL teachers versus undergraduate raters).

Three separate FACETS analyses were performed to determine whether the rater groups differed in severity when they rated the examinees' oral proficiency, accentedness, and comprehensibility. Rater measurement reports generated by FACETS were used to compare rating variability between rater groups.

The analysis of the written comments and interview data followed content analysis for qualitative data analysis (Miles and Huberman 1994). Brown, Iwashita and McNamara's (2005) empirically developed coding scheme for rater orientations in speaking tasks was consulted as an initial guide. After a few iterations of categorization and data segmentation, a coding scheme that encompassed all the factors commented on was developed for the current study. These categories included six main categories: linguistic resources, phonology, fluency, content, global assessment, and nonlinguistic factors, and 15 corresponding subcategories (see Table 5.1). To check the reliability of the coding, a second coder coded a random sample of 20% of the data. The overall percentage agreement achieved was 79.7%. After the second coder and the researcher discussed the 90 difficult cases one by one, a 100% agreement was reached. The researcher then coded the entire data set both at the main category and subcategory levels.

Table 5.1 Coding scheme

Main categories	Subcategories	Examples
Linguistic resources	Grammar	There were a few verb tense errors.
	Vocabulary	Very poor word choice.
	Expressions	There are some awkward expressions.
	Textualization	There is no strong use of cohesive devices.
Phonology	Pronunciation	The vowels seem to be lengthened.
	Intonation	The speech is full of intonation in odd places.
	Rhythm and stress	The stress inhibits complete comprehension.
	Accent	His accent was really heavy.
Fluency	Pauses	There were a lot of pauses in his speech.
	Repetition and repair	His repetitions of words affected the flow.
	Speech rate	She spoke too slowly.
	Global fluency	The speaker had some trouble with fluency.
Content	Task fulfillment	The task was not completed.
	Ideas	Hard to catch several ideas.
	Organization	Good organization to his response.
Global assessment	No subcategory	Well done; I could understand everything.
Non linguistic factors	No subcategory	Perhaps his anxiety was influencing his responses.

The analysis of the interview data included the constant comparison method of qualitative data analysis. Coherent and related comments in the interviews were grouped as one theme (McCracken 1988, Miles and Huberman 1994, Patton 1990). The major factors related to raters' decision-making processes were identified and will be discussed in conjunction with the quantitative results.

Results

Descriptive statistics

The mean rating data for the two rater groups were compared to determine whether there was an effect from rater background. The descriptive statistics (see Table 5.2) were derived by averaging the 28 examinee scores given by rater group; they were not the average ratings awarded by individual raters.

The descriptive statistics show that the ESL teachers as a group awarded a higher mean score on oral proficiency, but lower mean scores on accentedness and comprehensibility than the undergraduates did. The standard deviations indicate that the undergraduates displayed more variations in oral proficiency ratings than the ESL teachers.

Table 5.2 Descriptive statistics by rater group

Rater group	Measures	Max. possible score	Min.	Max.	M	SD
ESL teachers	Oral proficiency	60	35.6	52.6	43.1	3.4
	Accentedness	9	2.0	7.0	5.2	1.1
	Comprehensibility	9	1.9	5.4	3.6	0.9
Undergraduates	Oral proficiency	60	29.8	52.8	42.3	4.9
	Accentedness	9	3.3	7.6	6.0	1.0
	Comprehensibility	9	2.4	6.0	4.1	0.8

Note: M = Mean; SD = Standard deviation

FACETS analyses

The average severity measures of rater groups were compared to analyze any differences in the evaluations of ITA speech between groups. FACETS produces an estimate (in logit) of the degree of severity each rater exercised, the error associated with this estimate, and fit statistics for detecting model-data fit for each individual rater. To determine whether the raters differed in severity at the group level, the fixed (all same) chi-square tests were examined. The fixed (all same) chi-square analysis tested the null hypothesis that the rater groups could be thought of as equally lenient after allowing for measurement errors. Results of the comparison are presented in Tables 5.3, 5.4 and 5.5. Table 5.3 shows that the ESL teachers did not rate the examinees' oral proficiency more severely or leniently than the undergraduate raters, $x^2 = (1, N = 2) = 3.2, p = 0.07$.

Table 5.3 Rater group measurement report on oral proficiency

Rater group	Observed raw score	Observed count	Observed raw score average	Average severity measure (in logits)	Model SE
ESL teachers	47040	1092	43.1	−0.05	0.05
Undergraduates	113770	2688	42.3	0.05	0.03
M	80405.0	1890.0	42.7	0.00	0.04
SD	47185.2	1128.5	0.5	0.07	0.01

Note: Model SE = Model standard error; M = Mean; SD = Standard deviation; fixed (all same) chi-square = 3.2; df = 1, significance = 0.07

Table 5.4 shows that the ESL teachers rated more leniently than the undergraduate raters when they evaluated the examinees' accentedness. Results of the chi-square test indicate that the rater groups differed significantly in the average levels of severity they exercised when evaluating the examinees' accentedness, $x^2 = (1, N = 2) = 67.6, p < 0.001$.

Table 5.4 Rater group measurement report on accentedness

Rater group	Observed raw score	Observed count	Observed raw score average	Average severity measure (in logits)	Model SE
ESL teachers	5727	1092	5.2	−0.12	0.02
Undergraduates	16051	2688	6.0	0.12	0.02
M	10889.0	1890.0	5.6	0.00	0.02
SD	7300.2	1128.5	0.5	−0.16	0.01

Note: Model SE = Model standard error; M = Mean; SD = Standard deviation; fixed (all same) chi-square = 67.6; df = 1, significance = 0.00

Table 5.5 shows that the ESL teachers rated more leniently than the undergraduate raters when they evaluated the examinees' comprehensibility. Results of the chi-square test indicate that the rater groups differed significantly in the average levels of severity they exercised when evaluating the examinees' comprehensibility, $x^2 = (1, N = 2) = 75.4, p < 0.001$.

Table 5.5 Rater group measurement report on comprehensibility

Rater group	Observed raw score	Observed count	Observed raw score average	Average severity measure (in logits)	Model SE
ESL teachers	3933	1092	3.6	−0.13	0.02
Undergraduates	11090	2688	4.1	0.13	0.01
M	7511.5	1890.0	3.9	0.00	0.02
SD	5060.8	1128.5	0.4	0.18	0.01

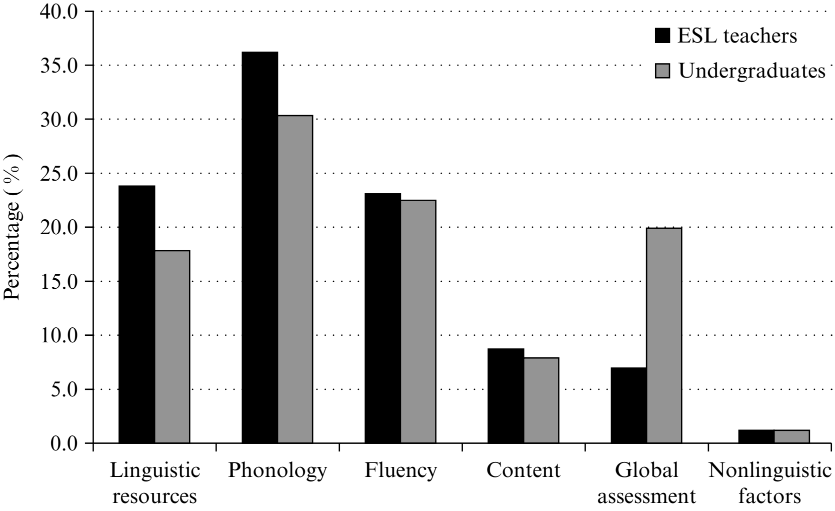
Note: Model SE = Model standard error; M = Mean; SD = Standard deviation; fixed (all same) chi-square = 75.4; df = 1, significance = 0.00

To summarize, the FACETS analyses revealed that the undergraduate raters were significantly more severe in their ratings of accentedness and comprehensibility than the ESL teachers. However, they did not differ in severity when judging oral proficiency.

Written comments and interviews

To answer the research question regarding factors that drew raters' attention when evaluating ITA speech, the written comments and interview data were analyzed. Figure 5.1 illustrates the percentages of the written comments coded for each main category. Phonology accounted for the largest group of comments for both rater groups. Neither group made many comments pertaining to the nonlinguistic factors, such as examinees' test-taking strategies, voice quality, and evidence of confidence or nervousness in the responses, suggesting that the primary criteria employed by the raters were related to the oral proficiency construct being measured.

Figure 5.1 Percentage distribution of comments coded for the main categories



To perform between-group comparisons regarding the rating-decision factors, the percentages of comments each rater made for each coded category, instead of the raw frequency, were used because of the imbalanced numbers of raters across rater groups. The tests of normality indicate that the distribution of the percentages calculated for each rater for each code did not meet the statistical assumptions of parametric tests. Thus, the nonparametric tests, Mann-Whitney *U* tests, were used to compare the coded data (see Table 5.6).

Results of the Mann-Whitney *U* tests show that the percentage of mentions was significantly different between the two groups only for the category of Global assessment, U (43) = 62.5, Z = −3.64, $p < 0.001$. The obtained

Table 5.6 Descriptive statistics and Mann-Whitney U tests for the coded categories

Main categories	ESL teachers (N = 13)		Undergraduates (N = 32)		Z-value	Effect size
	M	SD	M	SD		
Linguistic resources	20.6	10.9	17.1	10.7	−1.03	0.32
Phonology	40.6	17.8	31.5	15.7	−1.62	0.54
Fluency	23.3	9.5	20.7	11.7	−5.26	0.24
Content	8.6	7.7	6.5	6.1	−7.54	0.30
Global assessment	5.7	5.1	22.4	20.0	−3.64*	1.33
Nonlinguistic factors	1.2	1.4	1.8	3.1	−0.02	0.25

*Note: Effect size is based on Cohen's d. * = significance p<.001; M = Mean; SD = Standard deviation*

effect size of 1.33 based on Cohen's *d* shows a very strong effect (Cohen 1992), suggesting that the undergraduate raters used the global quality of the responses to justify their scores significantly more frequently than the ESL teachers did. Further differences between rater groups in terms of the way they commented on the various subcategories were also found in the written comments but will not be reported here (see Hsieh 2011 for a detailed report).

Interview data

The interview data further revealed the differences in raters' perceptions of factors that had affected their rating decisions. Three major factors emerged: 1) accent familiarity, 2) the importance of accent as a rating criterion, and 3) analytic versus holistic ratings.

Accent familiarity

Raters' varying background experiences with foreign accents surfaced as an important factor that influenced their ratings. Most teacher raters acknowledged that their accent familiarity had helped them better comprehend accented speech and, thus, influenced how they judged the ITA candidates. Three teacher raters stated candidly that they had become much more lenient raters over time due to their increasing familiarity with the ITA population.

Contrastively, most undergraduate raters felt that rating the ITA candidates was a very difficult task at first due to the heavy accents the candidates had. More than 80% of the undergraduate raters reported that they grew up in small, predominantly white communities and were unfamiliar with nonnative accents. 'I've never had any exposure to foreigners when I grew up. And actually, many of my friends don't either . . . so I might've been very harsh on my ratings 'coz I just didn't understand their accents,' one undergraduate rater said. Similar remarks were prevalent in other undergraduate raters' interview protocols, suggesting that there was a substantial discrepancy in the degree of accent familiarity between the rater groups, which had consequentially affected raters' rating behaviors.

Accent as a rating criterion

ESL teachers and undergraduates also differed in the way they employed foreign accents as part of the rating criteria. Ten of the teachers reported that they did not take accents into consideration in their judgments and thought that ITAs should not be judged based on their accents. One reason provided was that they did not deem it fair to the test takers. One commented that 'you can't judge whether one person can be a TA or not based on their accent. It's not fair and it's discriminating'. In addition to the fairness issue, other teachers suggested that successive exposures to accents would improve listeners' comprehension of ITA speech, and thus raters should not heavily base their judgments on how 'foreign' the ITA candidates' accents sounded. 'Eventually, the undergraduates would be able to pick up the accents a couple months into the semester if not earlier,' one said.

In sharp contrast, a general consensus among the undergraduates was that accent should be treated as a key criterion for the evaluation of ITAs. As they presumed, accent was the major factor interfering with their comprehension of ITA speech. 'I can't understand what they are saying because of their accents . . . if you can't figure out what the speaker was saying because of his accent, he shouldn't be a TA,' one reported. 'I think accent is very important and I think about their accents when I evaluate these speakers,' another commented. This fundamental difference in how the raters perceived accents as a rating criterion for screening ITAs partially explains the discrepancy in rater severity observed in the accent ratings.

Analytic versus holistic ratings

Another difference between the teachers' and the undergraduates' rating behaviors pertains to their approaches to ratings. The teachers commented frequently on specific linguistic features while the undergraduates tended to adopt a more global, impressionistic approach to making their rating decisions. Nine of the teachers listed specific speech features they attended to and explained how they used these features to derive their ratings. 'You think about their intonation, their word stress, and the flow, I mean, the overall fluency . . . and you also consider their grammar, their vocabulary, and the content too . . . my approach is more analytical, you know, just all the different features,' one stated.

The interview data showed, perhaps unsurprisingly, that few undergraduates used metalinguistic terms as frequently as the ESL teachers did to describe how they made the rating decisions. Twenty-eight of the undergraduates stated that they based their judgments primarily on whether they could understand a speaker or would like the person to be their TA. One commented: 'I just graded them based on if I want them to be my TA. Like for my math class, I'm not good at math, so I wouldn't want to have a foreign TA simply because I couldn't understand him'. Similar comments were common in the undergraduates' interview protocols, indicating that the undergraduate raters judged the ITA speaking proficiency more impressionistically, depending upon how they *felt* about the comprehensibility of the speakers or whether they would *want* the speakers to be their instructors.

In summary, the interview protocols revealed that the undergraduate raters' limited experience with accented speech had a substantial impact on their comprehension of ITA speech and, thus, on how they oriented their rating decisions. On the other hand, most teacher raters employed a variety of linguistic features to make their rating judgments and considered accent a peripheral aspect of ITA speech evaluation. Whereas the teacher raters disfavored the use of accents as part of the rating criteria, the great majority of the undergraduates appeared to use accent as a general evaluation guide. Finally, the ESL teachers and the undergraduates differed in their overall approaches to rating. While the teacher raters tended to rate analytically, the undergraduates were inclined to make their judgments based on whether they felt they could understand the speakers or whether they would like the speakers to be their TAs.

Discussion

Testing programs that administer high-stakes tests are responsible for delivering tests that are reliable, ethical, and valid. Testing programs that administer ITA screening exams are no exception to this rule. The tests they administer are ultimately used to decide who can obtain a teaching assistantship, which will impact not only the test takers themselves, but also the lives of the test takers' family members, the ITAs' students, and the universities that hire the ITAs.

Findings of this investigation suggest that ESL teachers' ratings of oral proficiency were similar to the ratings of novice undergraduate raters in many ways. Differences in rater severity between ESL teachers and American undergraduates were small and did not reach statistical significance. These results provide mixed support for previous research on undergraduate raters. Despite disparate rating experiences (expert versus inexperienced) and contrasting linguistic backgrounds (varied versus non-varied) between the two rater groups, the undergraduates assigned oral proficiency ratings that were comparable to those assigned by the ESL teachers. The overall similarity between ESL teacher and undergraduate raters' judgments on oral proficiency contradicts findings of Orth (1983) and Oppenheim (1998), and yet is consistent with the results of other research studies (e.g. Powers et al 1999, Saif 2002), corroborating previous findings that ratings of oral proficiency awarded by linguistically naïve undergraduate students and ESL professionals are similar and related.

Results are more complex and intriguing when the two groups' ratings on accent and comprehensibility are considered. There were significant, between-group differences in the two groups' ratings on accentedness and comprehensibility. Undergraduate raters were more severe when they judged the examinees' foreign accents. They also reported a significantly higher level of difficulty in comprehending the examinees' speech. But these results should not be surprising in light of previous research that suggests that American undergraduates tend to evaluate ITAs' foreign accented speech negatively (Bailey 1983, Fox and Gay 1994, Plakans 1997, Rubin 1992, Rubin and Smith 1990). The FACETS analyses reported support such a view and extend it with respect to ratings of accentedness and comprehensibility in between-group comparisons.

As a researcher, the question that interests me is why the ESL teachers were more lenient in their ratings of accentedness and comprehensibility. Findings from the interviews may provide some insight in helping us understand this outcome. One possible reason for the between-group difference pertains to the raters' amount of exposure to and experience with foreign-accented speech prior to the study. As the interview data suggest, all undergraduate raters had very limited contact with nonnative English speakers, whereas the ESL teachers had extensive exposure to and familiarity with nonnative accents. This extensive exposure to an array of diverse English pronunciations from learners of various L1 backgrounds enhanced the ESL teachers' abilities to decipher the meaning conveyed by accented, L2 speech. These results corroborate findings from a large body of previous work in speech perception and on the cognitive processing of accented speech – work that supports the general claim that the amount of exposure to World Englishes and/or interaction with nonnative speakers can enhance the listening comprehension of those English varieties (Derwing, Munro and Rossiter 2002, Elder and Davies 2006, Elder and Harding 2008, Major, Fitzmaurice, Bunta and Balasubramanian 2002, Munro and Derwing 1994).

The results also reveal that the ESL teachers and the undergraduates attended to several aspects of the linguistic dimensions in the examinees' speech differently, as predicted by past research (Chalhoub-Deville 1995, Elder 1993). Specifically, the teacher raters commented more frequently on a variety of linguistic features than did the undergraduates. The undergraduates, on the other hand, appeared to evaluate the examinees' oral performances more impressionistically. It seems logical to assume that the undergraduates had lumped many linguistic features under the concept of accent, features that the ESL teachers considered separately from an examinee's accent *per se*. For instance, many ESL teachers observed the examinees' narrow pitch ranges and unnatural stress patterns and commented on their impact on comprehensibility. The undergraduates, however, did not comment much on intonation or stress patterns, most likely because they are linguistically less sophisticated than the ESL teachers and were less able to describe such features metalinguistically. The majority of the undergraduates may have attributed their problems in deciphering problematic intonation and stress to the examinees' accents, and this practice may explain, in part, why the undergraduates awarded higher accent ratings (more accented) – their target for accent was larger than the ESL teachers' target for accent. Such an interpretation also implies that the differences in attention paid to various linguistic features may not reflect a difference in what features the raters actually attended to, but rather a difference in how they explained what features they attended to.

This study identified several nonlinguistic factors to which raters attended, including test-taking strategies, voice quality, and affective factors such as confidence or nervousness. None of these factors have been thoroughly discussed in previous studies (e.g. Brown et al 2005, Rubin 1992, Winke, Gass and Myford 2013). Previous research has demonstrated that nonlanguage factors, such as the speaker's ethnicity (Rubin 1992), could impact undergraduates' judgments of L2 speech. Nevertheless, the number of comments made by both groups on these nonlinguistic factors was small, suggesting that linguistic features of the speakers were the predominate constituents of the raters' orientations.

Results of the study demonstrated that the inclusion of accentedness and comprehensibility in the investigation of rater effects in ITA testing is important and illuminating. It can be argued that, within the context of ITA testing, oral proficiency alone may be too 'broad' a measure of ITA communication abilities, and thus not the ideal predictor of ITA success in using English for instructional purposes. Future research should continue to investigate and identify specific subcomponents of L2 speech that influence ESL teachers' and American undergraduates' judgments of ITA speech to better inform decision-making in ITA testing.

Implications for ITA testing

The results of this study suggest that it may be the case that ITA testing programs should take great care in having linguistically naïve undergraduates as official raters. As the results in this study revealed, undergraduates' personal experiences with foreign accents or ITAs may impact the way they judge ITA speech and, thus, introduce construct-irrelevant variance in the ratings. ITA testing programs should instead use undergraduates to check the threshold of what they may consider to be incomprehensible speech. To this end, it is also critical to raise the cut-off point for comprehensibility of the ITA speech to ensure that the undergraduate students are not disadvantaged by the result of the ITA test. On the other hand, we should not underestimate undergraduates' abilities to adapt and comprehend ITAs whose speech falls within that 'gray' zone (i.e. the zone between what undergraduate raters would call incomprehensible, but what expert ESL teachers would call comprehensible). The potential rating difference between the official ITA testing raters and the undergraduates should also be monitored, carefully evaluated and researched.

While results of the study indicate the presence of rater variability in the context examined, the differences in ratings should be interpreted with caution in several ways. First of all, this study employed untrained undergraduate raters whose ratings might be biased based on their personal experiences with foreign accents and ITAs, as shown in the interview data. While linguistically naïve undergraduates' ratings on comprehensibility might be better trusted as they represent the population who have ITAs as instructors in their university courses, trained raters would be better positioned to evaluate oral proficiency. Future research might consider ratings of oral proficiency awarded by trained undergraduate raters to derive better criterion references.

Secondly, the construct definition of accent is still evolving. Raters may have conceptualized it locally and employed the rating scales differently. More research is needed to examine specific features of accent. In addition, we need better measures of accent – measures that take into consideration the impact of accent on overall comprehensibility of speech.

Thirdly, the research context may have constrained the generalizability of the results. While the majority of the University M undergraduate students are in-state, native English speakers, it is unclear whether the findings reported here would hold for undergraduate raters from other geographical regions where the makeup of the student body and the wider communities are more linguistically diverse. Future research in ITA testing should investigate the impact of undergraduate raters' linguistic backgrounds on the evaluations of ITA speech.

References

Bailey, K M (1983) Foreign teaching assistants at US universities: Problems in interaction and communication, *TESOL Quarterly* 17, 308–310.

Barrett, S (2001) The impact of training on rater variability, *International Education Journal* 2, 49–58.

Bejar, I I (1985) *A Preliminary Study of Raters for the Test of Spoken English,* TOEFL Research Report RR-85-5, Princeton: Educational Testing Service.

Brown, A, Iwashita, N and McNamara, T F (2005) *An Examination of Rater Orientations and Test-taker Performance on English-for-Academic-Purpose*

Speaking Tasks, TOEFL Research Report RR-05-05, Princeton: Educational Testing Service.

Chalhoub-Deville, M (1995) Deriving oral assessment scales across different tests and rater groups, *Language Testing* 12, 16–35.

Clarke, J and Swinton, C (1980) *The Test of Spoken English as a Measure of Communicative Ability in English-medium Instructional Settings*, TOEFL Research Report RR 80-33, Educational Testing Service.

Cohen, J (1992) A power primer, *Psychological Bulletin* 112, 155–159.

Congdon, P and McQueen, J (2000) The stability of rater severity in large-scale assessment programs, *Journal of Educational Measurement* 37, 163–178.

Davies, C, Tyler, A and Koran, J (1989) Face-to-face with native speakers: an advanced training class for international teaching assisstants, *English for Specific Purposes* 8, 139–153.

Derwing, T M and Munro, M J (1997) Accent, intelligibility, and comprehensibility: evidence from four L1s, *Studies in Second Language Acquisition* 20, 1–16.

Derwing, T M and Munro, M J (2005) Second language accent and pronunciation teaching: A research-based approach, *TESOL Quarterly* 39, 379–398.

Derwing, T M and Munro, M J (2009) Putting accent in its place: Rethinking obstacles to communication, *Language Teaching* 42, 476–490.

Derwing, T M, Rossiter, M J and Munro, M J (2002) Teaching native speakers to listen to foreign-accented speech, *Journal of Multilingual and Multicultural Development* 23, 245–259.

Douglas, D and Smith, J (1997) *Theoretical Underpinnings of the Test of Spoken English Revision Project*, TOEFL Monograph Series RM-97-2, Princeton: Educational Testing Service.

Eckes, T (2005) Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis, *Language Assessment Quarterly* 2, 197–221.

Elder, C (1993) How do subject specialists construe classroom language proficiency? *Language Testing* 10, 235–254.

Elder, C and Davies, A (2006) Assessing English as a lingua franca, *Annual Review of Applied Linguistics* 26, 282–304.

Elder, C and Harding, L (2008) Language testing and English and an international language: Constraints and contributions, *Australian Review of Applied Linguistics* 31, 34.1–34.11.

Engelhard, G J and Myford, C M (2003) *Monitoring Faculty Consultant Performance in the Advanced Placement English Literature and Composition Program with a Many-faceted Rasch Model*, College Board Research Report No. 2003-1, Princeton: Educational Testing Service.

Ferrier, D E, Lovett, B J and Jordan, A H (2011) Construct-irrelevant variance in achievement test scores: A social cognitive perspective, in Madson, L E (Ed) *Achievement Tests: Types, Interpretations, and Uses*, New York: Nova, 89–108.

Fox, W S and Gay, G (1994) Functions and effects of international teaching assistants, *Review of Higher Education* 18, 1–24.

Hoyt, W T (2000) Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods* 5, 64–86.

Hsieh, C (2011) *ESL teachers' versus American undergraduates' judgments of oral proficiency, accentedness, and comprehensibility*, poster presented at the 33rd Language Testing Research Colloquium, Ann Arbor, Michigan.

Kang, O (2010) Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness, *System* 38, 301–315.

Kang, O (2012) Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance, *Language Assessment Quarterly* 9, 249–269.

Kunnan, A J (2005) Towards a model of test evaluation: Using the test fairness and the test context frameworks, in Taylor, L and Weir, C J (Eds) *Multilingualism and Assessment: Achieving Transparency, Assuring Quality, Sustaining Diversity – Proceedings of the ALTE Berlin Conference, May 2005*, Cambridge: UCLES/Cambridge University Press, 229–251.

Linacre, J M (2010) *FACETS Version 3.67*, Chicago: WINSTEPS.com.

Major, R C, Fitzmaurice, S F, Bunta, F and Balasubramanian, C (2002) The effects of nonnative accents on listening comprehension: Implications for ESL assessment, *TESOL Quarterly* 36, 173–190.

McCracken, G (1988) *The Long Interview*, London: Sage.

McNamara, T F (1996) *Measuring Second Language Performance*, Harlow: Longman.

Miles, M B and Huberman, A M (1994) *Qualitative Data Analysis: An Expanded Sourcebook*, Thousand Oaks: Sage.

Munro, M J and Derwing, T M (1994) Evaluations of foreign accent in extemporaneous and read material, *Language Testing* 11, 253–266.

Munro, M J and Derwing, T M (1995) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, *Language Learning* 41, 73–97.

Munro, M J and Derwing, T M (1998) The effects of speaking rate on listener evaluations of native and foreign-accented speech, *Language Learning* 48, 159–182.

Munro, M J and Derwing, T M (2001) Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate, *Studies in Second Language Acquisition* 23, 451–468.

Myford, C M and Wolfe, E W (2000) *Monitoring Sources of Variability within the Test of Spoken English Assessment System*, TOEFL Research Report RR-00-06, Princeton: Educational Testing Service.

Myford, C M and Wolfe, E W (2003) Detecting and measuring rater effects using many-faceted Rasch measurement: Part I, *Journal of Applied Measurement* 4, 386–422.

Oppenheim, N (1998) *Undergraduates' assessment of international teaching assistants' communicative competence*, paper presented at the annual meeting of the Teachers of English to Speakers of Other Languages, Seattle, 17–21 March 1998.

Orth, J L (1983) *University Undergraduate Evaluational Reactions to the Speech of Foreign Teaching Assistants*, unpublished PhD thesis, University of Texas at Austin.

Patton, M Q (1990) *Qualitative Evaluation and Research Methods*, Newbury Park: Sage.

Pickering, L (1999) *An analysis of prosodic systems in the classroom discourse of native speaker and nonnative speaker teaching assistants*, unpublished PhD thesis, University of Florida.

Pickering, L (2001) The role of tone choice in improving ITA communication in the classroom, *TESOL Quarterly* 35, 233–255.

Plakans, B S (1997) Undergraduates' experiences with and attitudes toward international teaching assistants, *TESOL Quarterly* 31, 95–118.

Powers, D E, Shedl, M A, Wilson-Leung, S and Butler, F A (1999) Validating the revised Test of Spoken English against a criterion of communicative success, *Language Testing* 16, 399–425.

Rubin, D (1992) Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants, *Research in Higher Education* 33, 511–531.

Rubin, D and Smith, K A (1990) Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants, *International Journal of Intercultural Relations* 14, 337–353.

Saif, S (2002) A needs-based approach to the evaluation of the spoken language ability of international teaching assistants, *The Canadian Journal of Applied Linguistics* 5, 145–167.

Schaefer, E (2008) Rater bias patterns in an EFL writing assessment, *Language Testing* 25, 465–493.

Tyler, A (1992) Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse, *TESOL Quarterly* 26, 713–729.

Winke, P, Gass, S and Myford, C (2013) Raters' L2 background as a potential source of bias in rating oral performance, *Language Testing* 30, 231–241.

Section 2
Assessing young learners in school contexts

6 Elementary English language learners and classroom content tests

Beth Clark-Gareca

State University of New York, New Paltz, US

Motivation for the research

Educational initiatives, such as Race to the Top and the Common Core Standards, are gaining momentum in the USA. These new mandates require that content knowledge for all children be frequently and rigorously assessed through standardized testing measures (see, for example, the Common Core State Standards Initiative 2013, Smarter Balanced Assessment Consortium 2013). Concurrent with the implementation of school-based accountability measures, the number of English language learners (ELLs) in the US is steadily increasing (Young, Cho, Ling, Cline, Steinberg and Stone 2008). In the 2011–2012 academic year, ELLs made up 9.1% of the total USA public school population and their numbers are expected to continue to rise in the coming years (National Center for Education Statistics 2015). As the number of English learners continues to rise, meeting their needs through English as a second language (ESL) instruction in schools becomes a serious consideration, especially in light of the ways that ESL instruction has been systematically reduced in recent years due, in part, to ever-shrinking budgets, and pressures from parents and other stakeholders for ELLs to participate fully in mainstream content classes and receive grade-level instruction to more adequately prepare for high-stakes examinations.

Literature review

As a result of these changing dynamics in USA schools, ELLs are spending more and more of their academic time in mainstream classrooms, and much of the responsibility for teaching and assessing ELLs in math and science falls to content teachers who may have little background in second language acquisition or second language methodology or assessment (Cizek 2007). In such cases, ELLs try to participate in routine classroom tasks and assessments to the best of their abilities, and teachers find themselves in the difficult position of evaluating ELL academic performance on tasks that are often well beyond the linguistic competences of the students.

Extending from this tension between language proficiency and content assessment, it stands to reason that ELLs tend to perform poorly on classroom content tests (Menken 2006, 2008). One reason for this poor performance is the complexity of the language of the tests themselves. Though content tests are not intended to measure language proficiency in the areas of math and science, they have been shown to correlate highly with measures of English reading comprehension (Abedi, Lord, Hofstetter and Baker 2000, Menken 2008, Shohamy 1997), meaning that content constructs and language constructs may be inextricably intertwined. Intuitively, it makes sense that if a child cannot understand the language of the test, it is highly unlikely that the test score will adequately reflect the child's content knowledge and that reasonable inferences and decisions can be made based on this performance.

Nonetheless, ELLs are often expected to take content tests and participate in math and science assessments alongside their L1 English-speaking classmates, practices that have led to the development of classroom-based test accommodations. Accommodations for ELLs, or 'changes to testing procedures, testing materials, or the testing situation in order to allow students' meaningful participation in the assessment' (Acosta, Rivera and Shafer-Willner 2008:vii) are intended to ease the testing process for ELLs, remove some of the linguistic burden that they face in understanding the language of the test, and allow them to participate more fully in classroom assessment. One concern associated with implementing test accommodations is that test constructs may be affected; in other words, the accommodation implemented may change the test in some fundamental way so that what is actually being measured will be affected. To date, very little is known about classroom-based accommodations in terms of *how* or *if* they are implemented on a regular basis (Cizek 2007). Even less is known about what the children for whom they are implemented think about their effectiveness in terms of their test taking experiences.

Tests and the implementation of test accommodations become important when the decisions that are made based on the results of classroom tests are considered. Though math and science tests constitute only one part of the evaluative systems in school-based achievement, they often play a substantial role in cumulative report card grades at the end of a marking period or the end of the year. Report card grades are, in turn, a standard criterion upon which scholastic decisions are made, including decisions about interventions, enrichment activities, exiting from ESL programs, tracking, retention, and overall access to academic pathways, such as advanced placement courses for college preparation. Additionally, report cards are an important way for parents to stay connected with their children's achievement in elementary school; thus, classroom assessment takes on a great deal of importance in the academic life of ELL children.

Interestingly, most classroom-based studies on assessment have focused primarily on *teacher* roles and practices during tests (Davison and Leung 2009, deJong and Harper 2005, McMillan 2003, McMillan, Myron and Workman 2002, Rea-Dickins 2004). This study intentionally addressed the need for greater exploration of *student* roles within testing, as well as the need for more research from the learners' perspective (Davison and Leung 2009, McNamara 2001, Rea-Dickins 2001, 2004). By investigating the perspectives of young ELLs whose academic lives are increasingly affected by tests, this provides a starting point from which to document ELL assessment practices in elementary contexts.

Research question

The data reported in this chapter were collected as part of a larger mixed methods study documenting ELLs' perspectives on classroom assessment practices, including accommodations and implementation on routine math and science tests. The research question that guided the portion of the study reported here was the following: How do ELLs perceive the classroom assessment practices in which they take part?

A particular focus of this study was to learn how well the children could reflect on their own learning processes and express their opinions on the assessment systems in which they participate.

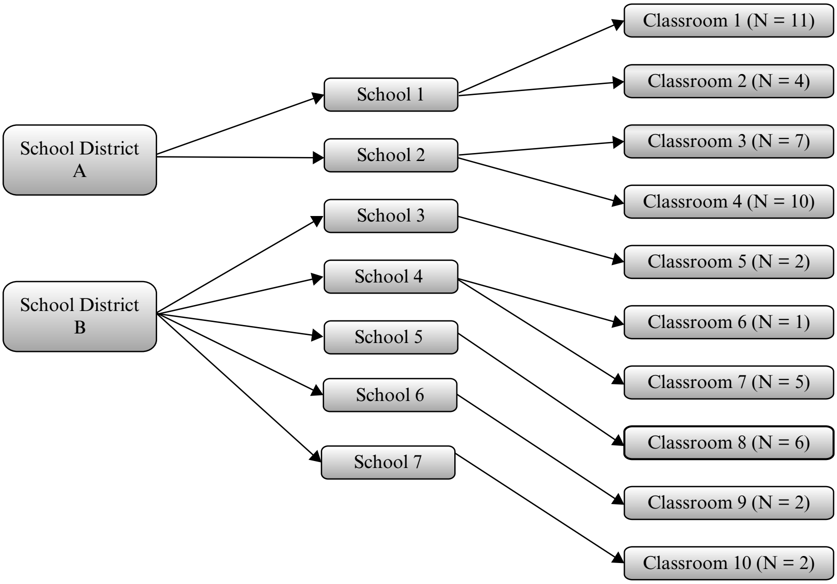
Data collection procedures

Data collection for this portion of the study consisted of three formal classroom observations in 10 Grade 4 classrooms in seven different schools. The initial observation in each classroom lasted 60 minutes and was intended to develop an understanding of the natural ecology of the classroom and to determine the positioning of ELLs in that space, as well as to establish a non-interfering presence in the classroom environment. The initial observation was followed by two focused observations during scheduled, routine classroom math and science test administrations that were conducted by the Grade 4 teachers. These tests lasted between 15 and 40 minutes. Itemized observation protocols were completed during each observation to capture ELL actions and interactions before, during, and after the observed content tests.

When all test observations were completed in each classroom, semi-structured, individual interviews were conducted with each Grade 4 participant. Interviews with the children lasted approximately 15–30 minutes, depending on students' comfort during the interview process and the length of their answers. During the interviews, participants were asked to speak about their perceptions of the tests that they had just taken and of testing in general. With the Spanish speakers, the language of the interview was settled upon through a gentle negotiation process at the beginning of each interview. The children who spoke other languages had high enough proficiency in English to allow the interviews to be conducted easily in English. If the children became uncomfortable or expressed any concerns at any point in the process, the interviews were swiftly concluded. Interview protocols were organized by questions relating to 1) perceptions of tests and accommodations in general, 2) perceptions of the specific math or science tests observed, and 3) perceptions and predictions of grades. Interviews were audio-recorded and transcribed for ease of subsequent analysis. Student academic records in the form of standardized test scores and report cards were also collected and analyzed insofar as they contextualized the participants and testing situations.

Data were collected in two school districts in the State of Pennsylvania. Figure 6.1 displays the contexts in which the data were collected.

Figure 6.1 Research contexts



As noted in Figure 6.1, District A was a small suburban school district with about 10,000 students in total; 9% of students were identified as ELLs. District B was a small city school district with about 15,000 students in total; 8% were identified as ELLs. In District A, data were collected in two different schools, in four different Grade 4 classrooms (Classrooms 1–4). In District B, ELLs were observed and interviewed in five different schools, and within those schools in six different Grade 4 classrooms (Classrooms 5–11). Participants in District B ranged from one to six students in each class. Overall, classrooms had varying numbers of participants, ranging from one ELL in Classroom 6 to 11 ELLs in Classroom 1. Neither school district had made adequate yearly progress (AYP) as determined by state standardized tests during the year of data collection.

Participants

Fifty Grade 4 ELLs, 23 boys and 27 girls, in general education math and science classrooms, participated in the study. All of the children were between the ages of 9 and 11 years. Most of the participants came from Spanish-speaking homes (n = 48); in addition, one child was a speaker of Kiswahili and one of Punjabi. Of the 50 participants, 21 children had completed all of their formal education within the USA, and 24 had completed some of their education in their country of origin. Five participants had been in school in the USA for less than a year. Seven children had been retained at some point during their schooling in the USA, and two others had been highly recommended for retention by their school; however, in light of their parents' opposition to retention, they had been promoted.

Other documentation related to student academic achievement and proficiency levels was collected as part of the study. These data are found in Table 6.1.

Table 6.1 Student achievement and English language proficiency

PSSA scores	Advanced	Proficient	Basic	Below Basic	Exempt	Other	Total
Reading	1	10	13	22	3	1	50
Math	3	16	6	21	3	1	50
Science	0	23	14	10	3	0	50
Report card grades	**A**	**B**	**C**	**D**	**F**	**Other**	**Total**
Reading	6	24	18	0	0	2	50
Math	5	25	18	0	0	2	50
Science	14	21	9	0	4	2	50
Participant proficiency		**Monitor**	**Advanced**	**Intermediate**	**Beginner**	**Other**	**Total**
		13	23	8	6	0	50

Note: PSSA = Pennsylvania State System of Assessment

Table 6.1 displays students' scores on the PSSA, a standardized test assessing Grade 4 reading, math, and science. Most of the participants in this sample scored in the lowest band (i.e. Below Basic) on the reading (n = 22) and math (n = 21) sections of the examination. In science, the majority of children scored in the Proficient band (n = 23). On their report cards, most participants were in the 'B' and 'C' grade range for reading and math. In science, participants tended to be in the 'A' to 'B' range, though four children had a final grade of 'F' in science for their Grade 4 year. Except for these four children, all ELLs in the study had passing marks on their report cards at the end of Grade 4. Finally, language proficiency levels, as measured by the World-class Instructional Design and Assessment (WIDA) Consortium's *Access for ELLs* (WIDA 2016), were identified through a test report analysis. The majority of the participants in this sample had reached an Advanced level of English language proficiency, though all proficiency levels were represented in the study.

Data analysis

Observation protocols and interview transcripts were analyzed to address the research question. The analysis was modeled after Lee's (2004) framework on science teacher belief systems, and student data were first analyzed along one of four dimensions related to assessment, and subsequently thematically coded, and tabulated using descriptive statistics. Specialized practices for ELLs were identified by using the Acosta et al (2008) definition and framework of accommodations, which was related to linguistic/non-linguistic and first language/second language (L1/L2) accommodations. Some accommodations implemented were linguistic in nature, such as reading test items aloud or using peer translation, and some involved changing non-linguistic aspects of the test in the form of allocating additional time or using small groups during the administration of tests. The accommodations observed during classroom tests provided a point of departure and guided some interview questions. Accommodations were also examined in light of students' perceptions of stigma being attached to these accommodations, including any feelings of discomfort related to special assessment practices, and perceptions of the fairness of the implementation process within the learning community. It should also be noted that numeric tabulations were meant to be indicative of rigorous analysis of the interview data and not to support the generalizability of the findings to other populations.

Student perceptions of assessment

Beginning with ELLs' perceptions related to assessment, a majority of ELLs reported that they liked to take tests. Thirty-nine ELLs (78%) reported that they liked math tests and 32 (64%) reported that they liked science tests. Most children reported that they enjoyed the process of being *assessed* in a specific content area more than they liked the content area in general. For example, Tomy (a pseudonym), a Monitor level ELL, reported that he did not like math as a subject, but he really liked taking math tests because, 'when I get hundreds and my dad see that, he gets happy'. Many children reported a positive attitude when approaching the test taking process, expressing that tests presented them with a challenge that they could rise to and meet, and provided an opportunity to receive concrete proof that they were doing well in school.

Of the 50 ELL participants, 22 (44%) reported explicitly that tests presented them with opportunities to learn. Vivian (a pseudonym), an Advanced learner, described how tests were a mechanism for learning, a sentiment echoed by many participants throughout the data: 'Science tests is like you are learning about something and you are reminding of that thing. If it is a science question and I remember the answer, it's like doing the test and learning even more about what it is'. By having the chance to revisit learned concepts on assessments, Vivian was able to expand and reinforce her learning through the test content.

Student perceptions of accommodations

The ELLs in the study were asked to comment on the accommodations that were used during the observed test administrations. Students were first asked to comment on their general perceptions of accommodations and, then, on the specific test accommodations observed in their classrooms, e.g. additional time, reading aloud, peer translation, teacher assists, and small group test administrations. Occasionally students also shared their perceptions about accommodations that were not seen in their classrooms (e.g. translated tests or bilingual dictionaries).

When the participants were asked to comment on the changes that were made for them on math and science tests, 46 students (92%) claimed to like taking tests with accommodations. Explaining their reasons why, 33 students (66%) reported that test accommodations helped them learn what they needed to know. Tifany, a Monitor level ELL, reported liking the changes her teacher made for her during assessment because, 'it could help me and learn more better'. Thirty-two ELLs (64%) said that they liked accommodations because using them during tests allowed them to get better grades. Rosa, also at the Monitor level, explained that accommodations provided opportunities in which '[We] can maybe get a better grade or we could get all the answer right'.

In the case of a small number of students (8%), concerns about special treatment during assessment were related. Two students referred explicitly to instances of being excluded from regular assessments because of their lower English proficiency levels. Rather than take the classroom test, these students were given an alternate task to complete, or they were simply asked to wait for their classmates to finish the test before moving on to the next activity with their classmates. Exclusion from tests resulted in feelings of discomfort for these ELLs because they wanted to participate fully in the assessments with their classmates. Exclusion did not meet the criteria of a test accommodation in the strictest sense, but it was an observed practice in two classrooms in the study.

Additional time taking exams was considered favorably by ELLs, particularly if they needed it. Adrian, an Advanced ELL, spoke of the value of being given extra time, saying that it made the test 'more easy to do'. Though extra time was reported to be a standard accommodation for ELLs in all 10 classrooms, it was reported not to be given on test tasks in three classrooms, where producing answers quickly and developing automaticity were included in the test constructs.

The practice of the teachers reading the test out loud and item by item was observed in eight classrooms, and the ELLs in these classrooms (n = 40) were asked to explain their understandings of this assessment practice. Of those children, 26 (65%) reported that they found it helpful when the teachers read their tests for them, largely because this practice supported their developing reading skills in English. Like many of her classmates, Lisa, an Advanced level ELL, was positive about this practice. She explained how hearing the content read aloud helped her to understand difficult words: 'So when we are all stuck on the words, like on "probability", she [the teacher] can read the question and we would know it was "probability", not just guessing the word'. By hearing the lexical item and matching it to the written word on the page, the children reported being better able to comprehend the test item.

Participants also reported that reading tests aloud helped them pace themselves during tests. Ana, an Intermediate level ELL, spoke of how this technique was used by her teacher in class: 'cause he reads it twice and then he give us some time like to circle the answer, which is the correct one or your best guess, and then he just goes to next and does the same'. When asked what she thought of the teacher reading the test aloud, Ana said it gave her more time to think before circling the answer than when she took the test on her own.

When asked about their perceptions of translated tests, 30 students (60%) said that they thought tests in their L1s would be difficult because they did not know how to read in their mother tongues. Magdalena, a Monitor level ELL, explained why she thought translated tests would be hard when she said: 'It would be like kinda harder to read because I'm like here, and now I kinda, sometimes I forget how to read in Spanish'. Other children reported that the linguistic demands of math assessments had an impact on their perceptions of translated tests. Pamela, a strong Spanish speaker, and a Monitor level ELL, contemplated taking math tests in Spanish, and finally decided it would be more difficult, saying, ‘Well, I don’t really know math in Spanish’. Fifteen ELLs (30%), all of whom were at a Beginner or Intermediate level of language proficiency reported thinking that a translated test would be easy. Candelaria, a Beginner, reported that taking tests in Spanish would be, ‘Normal, como que yo lo leo normal porque yo sé mucho en español. Como sé más en español . . . y en inglés yo sé un poco’. [Normal, like, I read it normal because I know a lot in Spanish. Like I know more in Spanish, and in English, I know a little.] Translated test forms were not observed in any classrooms studied, nor were they reported to be used regularly.

Students as a whole did not expect that their teachers would be able or willing to speak the native languages of the children in the class. In fact, none of the teachers spoke their students’ L1s proficiently, which supported a general perception by students that English was the language to be used at school. Only one Advanced level ELL, Fernanda, mentioned that she thought it would be useful for her teacher to know Spanish: ‘Well, she [the teacher], she should know like Spanish for the kids that know Spanish and not English. Like, if she would know Spanish she could read it for the people that don’t know’. Fernanda’s perspective was unique in the study; she was the only child who reported a belief that teacher practices could be changed to better serve the ELLs’ learning needs. The majority of the children accepted the assessment practices implemented by their teachers in their classrooms as a matter of course.

In the absence of teacher translators, 13 children (26%) spoke about the function of peer translation during assessments. An ad hoc system, peer translators were children who had high levels of language proficiency in both their L1 and in English, and could therefore assist the lower level ELLs with test translation on the spot during a test. Though most Beginners were forthcoming about relying on their classmates for help during math and science tests, students at higher proficiency levels expressed many ethical concerns as to whether talking during tests was cheating. Malena relied on her friend to help her translate her test: ‘Porque algunas palabras yo no me sé [sic] y ella me dice’. [Because some words I don’t know and she tells me]. Luisa, however, reported having some difficulty with her peer translator on a math test: ‘Yo le estaba diciendo si él sabe como se dibuja ese y él no me quería decir. El me dijo, “Es un examen. Tú lo tienes que hacer sola”. Pues yo me imaginé como era y lo puse’. [I was asking him if he knew how to draw that one, and he didn’t want to tell me. He told me, ‘It’s a test. You have to do it by yourself.’ So I guessed and I put down an answer.] The ELLs in need of the help of a translator most often reported that the cause of their difficulty was the language of the test and not their understanding of content concepts.

Teacher assistance was defined as any assistance provided during the test by the classroom or ESL teacher or by other specialized personnel. Teacher assistance was implemented in all classrooms, but not during all tests. A common perception of teacher assistance was that its purpose was to help children learn from their mistakes and achieve higher grades through item correction. Vivian related her understanding of teacher assistance: ‘She read the questions over for us a second time and then if you were having trouble with one of the tasks, she would come and tell you and she’s going help you and explain what you would do’. In Vivian’s example, the teacher helped her during the test by restating the item and offering her another chance to get it right. Jorge, an Intermediate level learner, explained how the ESL teacher helped during tests in his math class: ‘She’s going to . . . take us over there at ESL and help. And not help, like “Oh my God, that’s the answer! And that one’s the answer!” Not like that’. Jorge recognized that teachers could help students, but that there were constraints around what they could ethically do during a test administration.

During two classroom math tests, teachers were observed not to offer any assistance to ELLs. In the interviews, the children speculated on why their teachers did not help them. Derek thought that his teacher was not permitted to assist the ELLs, ‘cause he’s a teacher and the teacher can’t help the children’. Magdalena echoed the idea that help wasn’t allowed: ‘He doesn’t probably want to tell me the answers because he will get in trouble by the principals’. In spite of the fact that these statements were speculations and may not have been grounded in truth, the children accepted their teachers’ actions unequivocally.

The non-linguistic accommodation of small group test administration was also observed in several classrooms. During tests, 17 ELLs (34%) in five classrooms were observed taking tests in small groups. These special testing administrations were either conducted within the general classroom itself (often in the back or in a far corner) or in a separate classroom. All students reported that they appreciated this accommodation. Carmen, an Advanced level ELL, spoke of its benefits: ‘Sometimes when we’re in the back . . . we write our answer and then she tells us if it makes sense or not’. For Carmen, an opportunity to revise incorrect answers ensured a good grade on the test.

Student perceptions of test fairness

The topic of fairness during tests was often discussed extensively in the interviews. All students expressed an understanding that different children had different ability levels, and, thus, accommodations should be variably implemented according to students’ individual needs. Rosa, a Monitor level student, spoke about the practice of other ELLs being pulled into a small group or getting help from the ESL teacher: ‘I think that some people, you know, need help and some don’t. Because I can do it on my own . . . but some people need help so, yeah, I think that’s fair. Katarina asserted her belief that it was fair for Juliana, an ELL with special education needs, to receive more help on tests than she did: ‘I’m like the one that gets them all right. Not Juliana. Because she has a hard time understanding stuff, she doesn’t usually get very many things right’. The children perceived accommodations as special help that should be administered to whoever could benefit from them.

Students were also very accepting of their own need for accommodations on math and science tests. Mariela, an Advanced level student, felt that because she needed extra help, she should receive it, even if her classmates did not: ‘Because some people already know the tests or they can go by themselves. Some people don’t, so we need more help so they teach us in a small group’. This sense of justice manifested itself in one classroom in which only ELLs and students in special education were allowed the accommodation of additional time to complete the test. In their interviews, three ELLs remarked that this implementation was not fair because the other children in the class may also have benefited from this accommodation. Margarita spoke to this perceived lack of equity: ‘Cause maybe some people don’t know, and then they don’t get extra time! And they might get a grade zero for if they didn’t do the answer’. She felt that all children in the class should be given the accommodation of additional time if they needed it.

Student perceptions of scoring practices

Students reported welcoming any opportunity to earn better grades on content tests and spoke positively about the accommodated scoring practices in place in their classrooms. Accommodated scoring practices were defined as any special practice related to test evaluation that was explicitly enacted to give students the opportunity to earn a higher grade. Two primary accommodated scoring practices were implemented and identified in the observed data for ELLs, item correction and retesting.

Item correction was an ongoing practice in which teachers, upon examining students’ tests in progress, directed them to correct items that they had answered incorrectly. Item correction was mentioned as an assessment practice in interviews with 28 students (56%). Often referred to by the students as their teacher ‘giving hints’, these corrections were most often individually administered to students through teacher assistance. Luis, a Monitor level student, referred to his teacher’s tendency to implement item correction: ‘So, like, if we got it wrong, she could help us out so we could understand it and get a good answer on every test’. Similarly, Mariela appreciated item corrections made by her teacher during content assessments ‘because he could see that I did good or if I made a mistake to see that he said to check that. To go back and erase it to the right answer’.

Although ELLs were not happy if their performance required them to take tests again, all the ELLs in the study who had the option of retaking tests (n = 23) reacted favorably to this practice. Typically ELLs who failed a test were obligated to take it again, often because of school policies that ELLs should not be allowed to fail classroom tests or courses in general. Lisa recounted the value of retesting, explaining that she could 're-study and . . . be more focused'. Marisa was pleased to be offered a second chance on her math test, 'cause sometimes I don't do well on a test and that [retesting] give me an extra chance to do better'.

In other instances, students were given a choice as to whether to retake a test or not. Santiago enthusiastically explained the process of retaking tests in his classroom: 'Well, if we get a little like a part wrong, we can do it over if we want and even get even a better grade! But if we get it right, we don't have to. It's our choice if we want to or if we want to leave our score like that'. Retesting helped students build confidence in their ability to do well in school and helped them center their test preparation for the second time on their weaker content concepts.

Student perceptions of grading

ELLs had many insights into the assigning of grades on their math and science tests. Forty-three students spoke in their interviews about their perceptions of grades, with 30 students (70%) indicating that they did not understand the grading systems under which their work was evaluated. Competing rating systems used to grade students' work were reported to be the cause of significant confusion. Taking the example of the letter grade 'B', some students reported thinking that a grade of 'B' was not a good grade, referring to a 'B' score as equivalent to Basic on the PSSA, while others reported that a 'B' was a good grade, referring to the standard grading scale from 'A' to 'F'. Still there were other students who relied on familiar rating scales from their home countries, interpreting a letter grade of 'B' to mean 'Bueno' or 'Bien' which was typically ascribed an average value. In the following excerpt from an interview, Luis explained his understanding of the school's grading scale:

> Interviewer: What does an A or B mean?
> Luis: Advanced and Basic . . . Advanced means you did a great job and Basic means you did kind of a good job.

Confusing the rating scales from the classroom grading and the PSSA, Luis was unlikely to correctly interpret his performance or progress on his math test.

Students also expressed confusion related to the standard A to F grading scale used predominantly in the participant schools. When asked to elaborate on his understanding of good grades, Gabriel illustrated this confusion:

> Interviewer: What's a good grade?
> Gabriel: Um, A+ and just A.
> Interviewer: Okay, and what's a bad grade?
> Gabriel: An F. And sometimes a F+ is good. F-, not that good.

Gabriel's invention of the non-existent grades, F+ and F-, highlighted his limited understanding of the standard grading scale.

Grades given in percentages were reported to be another source of puzzlement for these young test takers, which was likely related to the fact that the mathematical concept of percentages had not yet been addressed in the Grade 4 math curriculum. The following excerpt exemplifies one student's lack of conceptual knowledge as she tried to explain the grading scale used on a 100-item math test.

> Lisa: Since there are a hundred facts, probably that's an A. And A minus would be a 99. A 98 would be a B. And then a B minus would be 98. (pause) No, an A would be 100, a B would be an 80. An A would be 100 to 90, like an A minus. A B would be 80 to a 70. And then below 60 is an F.
> Interviewer: So you think on this test you would get a what?
> Lisa: An F.

Lisa's understanding of percentages, as evidenced in this excerpt, was severely limited, which was the case for many ELLs in the study. Without explicit knowledge or understanding of the grading systems in place, they constructed meanings in unique but misguided ways. In light of this lack of understanding, students frequently spoke about their grades in the simple terms of extremes. The most common grade references made by participants were '100' and '0', and 'A' and 'F' for letter grades.

Discussion

Tapping into young ELLs' perspectives about testing yielded many interesting insights into their understandings of the process of assessment. Though there is a common notion that children are averse to taking tests (Jones, Jones, Hardin, Chapman, Yarbrough and Davis 1999, Mulvenon, Stegman and Ritter 2005), the findings from this study suggested that the large majority of ELLs liked taking tests in their content classes. Perhaps, the fact that these children were so young and had not yet had many negative school experiences informed these positive and optimistic perceptions; nonetheless, taking classroom tests was not reported to be a cause of dread or even concern for most of the participants interviewed.

Overall, the children reported little to no resistance to test taking because they believed assessment presented them with the opportunity to excel and learn, thus, emphasizing an understanding of the role of assessment for learning purposes (Brookhart 2007, Stiggins 2007a). In the observed tests, the students were likely justified in their expectations that they would do well, because many accommodated practices were designed and implemented explicitly with the purpose of raising their scores. Retesting and item correction always resulted in higher scores, thereby further encouraging the ELLs to approach assessment positively. Even though their final scores may not have been reflective of their true level of achievement in the strictest sense, accommodations and accommodated scoring practices allowed the children to participate in the regular classroom assessment activities to a greater extent than they could without the inclusion of these practices.

In terms of specific accommodations, the children reacted positively overall to all of the accommodations implemented for them largely because they believed that accommodations allowed them to do better. Interestingly, the students did not expect or even hope that their teachers would speak their L1s; thus, there was little exploration within this study of whether students' L1s could be a real resource to them during assessment. In some cases, though, the children expressed an understanding that their conceptual knowledge in a content area, for example math, was cognitively stored in English, and that translated tests would not be helpful. This finding supports earlier work in high-stakes environments that translated tests that do not match the language of instruction are typically not effective in leveling the linguistic playing field for ELLs (Stansfield 2003).

The children in the study believed that differentiating assessment practices for all students based on their individual needs was essential and fair. This perception was consistent with their understandings of tests as instruments of learning and that all learners need to be taught at their appropriate academic levels. The participants overwhelmingly expressed a natural empathy for their fellow classmates that were in the process of learning English or had special needs, as was evidenced by their willingness to translate for them.

The empathy that children felt for one another could be connected to the lack of competition fostered in the classrooms; that is, few of the participants in the study spoke of measuring their performance against anyone other than themselves. Perhaps the composition of these classrooms discouraged competition: all of the classes observed were designated lower content tracks based on students' PSSA scores the year prior. It may be that this fact encouraged an atmosphere in which everyone could do well if they tried their best, thus emphasizing effort over achievement.

Stiggins (2007b) has noted that children have the potential to be powerful educational decision-makers, but in this study, the children were not given significant decision-making power in terms of their own assessment practices and decisions. The vast majority of assessment practices were implemented by the teacher and without input from ELLs. Not involving the children in these decisions seems like a missed opportunity. Both teachers and students could have benefitted from directed conversations with the ELLs about their own understandings of their language acquisition processes and content comprehension. Such conversations could have helped the children further develop their metacognitive skills, without which students will move more slowly toward reaching necessary levels of learner autonomy that are characteristic of successful academic learners.

A striking finding of this study was that most ELLs did not understand the grading practices put in place for them. This finding may be representative of the fact that they have had exposure to many different systems and scales of evaluation, as well as a lack of understanding of the mathematical concepts used to express scores in math, i.e. percentages. It may also suggest that systems of measurement need to be consistent to be meaningful for test takers (Bachman and Palmer 1996), and without teachers explicitly teaching the meaning behind the A to F system, the 0–100 percent system, or the Below Basic to Advanced PSSA system, ELL students may not be able to interpret them correctly. Another necessary condition for assessment to be useful is for test takers to recognize and understand the criteria under which their knowledge is being evaluated (Bachman and Palmer 1996). The children in this study were found to have few effective ways of truly evaluating their academic or linguistic gains or losses, and were left without resources as to how to improve their performance.

In an effort to answer the call for more learner-centered research in the field of assessment (Davison and Leung 2009, McNamara 2001, Rea-Dickins 2001, 2004), this study's findings are particularly relevant to school environments where more and more classroom time is being dedicated to assessment. This investigation represents a starting point in ongoing work on what actually happens when ELLs take tests in elementary content classrooms. Considering the lack of research relating to young ELLs and classroom content assessment, the exploratory nature of this study was to determine the extent to which young ELLs could express their opinions about school-related topics. Perhaps the most noteworthy finding of this research was that 9 and 10-year-olds proved to be wonderfully capable of talking about their assessment experiences clearly, honestly, and eloquently.

References

Abedi, J, Lord, C, Hofstetter, C and Baker, E (2000) Impact of accommodation strategies on English Language Learners' test performance, *Educational Measurement: Issues and Practice* 19 (3) 16–26.

Acosta, B, Rivera, C and Shafer-Willner, L (2008) *Best practices in state* *assessment policies for accommodating English language learners: A Delphi study*, available online: files.eric.ed.gov/fulltext/ED539759.pdf

Bachman, L and Palmer, A (1996) *Language Testing in Practice*, Oxford: Oxford University Press.

Brookhart, S (2007) Expanding views about formative classroom assessment: A review of the literature, in McMillan, J (Ed) *Formative Classroom Assessment: Theory into Practice*, New York: Teachers College Press, 43–62.

Cizek, G (2007) Formative classroom and large-scale assessment: Implications for future research and development, in McMillan, J (Ed) *Formative Classroom Assessment: Theory into Practice*, New York: Teachers College Press, 99–115.

Common Core State Standards Initiative (2013) *Implementing the common core state standards*, available online: www.corestandards.org

Davison, C and Leung, C (2009) Current issues in English language teacher-based assessment, *TESOL Quarterly* 43 (3), 393–415.

deJong, E and Harper, C (2005) Preparing mainstream teachers for English language learners: Is being a good teacher good enough? *Teacher Education Quarterly* 32 (2), 101–124.

Jones, M G, Jones, B, Hardin, B, Chapman, L, Yarbrough, T and Davis, M (1999) The impact of high-stakes testing on teachers and students in North Carolina, *Phi Delta Kappan*, 81 (3), 199–203.

Lee, O (2004) Teacher change in belief and practices in science and literacy instruction with English language learners, *Journal of Research in Science Teaching* 41 (1), 65–93.

McMillan, J (2003) Understanding and improving teachers' classroom assessment decision making implications for theory and practice, *Educational Measurement: Issues and Practice* 22 (4), 34–43.

McMillan, J, Myron, S and Workman, D (2002) Elementary teachers' classroom assessment and grading practices, *Journal of Educational Research* 95 (4), 203–213.

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

Menken, K (2006) Teaching to the test: How No Child Left Behind impacts language policy, curriculum, and instruction of English language learners, *Bilingual Research Journal* 30 (2), 521–546.

Menken, K (2008) *English Language Learners Left Behind*, New York: Multilingual Matters.

Mulvenon, S, Stegman, C and Ritter, G (2005) Test anxiety: A multifaceted study on the perceptions of teachers, principals, counselors, students, and parents, *International Journal of Testing* 5 (1), 37–61.

National Center for Education Statistics (2015) *Fast facts: English language learners*, available online: nces.ed.gov/fastfacts/display.asp?id=96

Rea-Dickins, P (2001) Mirror, mirror on the wall: Identifying processes of classroom assessment, *Language Testing* 18 (4), 429–462.

Rea-Dickins, P (2004) Understanding teachers as agents of assessment, *Language Testing* 21 (3), 249–258.

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

Smarter Balanced Assessment Consortium (2013) *School years*, available online: www.smarterbalanced.org

Stansfield, C (2003) Test translation and adaptation in public education in the USA, *Language Testing* 20, 189–207.

Stiggins, R J (2007a) Conquering the formative assessment frontier, in McMillan, J (Ed) *Formative Classroom Assessment: Theory into Practice*, New York: Teachers College Press, 8–28.

Stiggins, R J (2007b) Assessment through the students' eyes, *Educational Leadership* 64 (8), 22–26.

WIDA (2016) *Access for ELLs*, available online: www.wida.us/assessment/access20.aspx

Young, J, Cho, Y, Ling, G, Cline, F, Steinberg, J and Stone, E (2008) Validity and fairness of state standards based assessments for English language learners, *Educational Assessment* 13, 170–192.

7 Exploring relationships between multi-word vocabulary, transparency, and literacy development

Sara A Smith

California State University East Bay, US

Motivation for the research

Vocabulary is commonly recognized as an essential element of reading ability, contributing to reading itself and influencing many early reading-related skills (e.g. word identification, phonological awareness, orthographic knowledge) (Chiappe, Chiappe and Gottardo 2004, Nation and Snowling 1998). One area of vocabulary research of increasing interest is multi-word phrases (MWPs), ranging from non-transparent, idiomatic items like 'break the ice' to more transparent set phrases like 'catch a cold'. Researchers have argued that these word combinations are not processed, stored or recalled as separate words but can instead be conceptualized as one 'giant lexical unit' and should be included in measures of vocabulary size (Nippold 1998, Wray 2002). Corpus studies have shown that these types of items are extremely common in English (Erman and Warren 2000, Sinclair 1991) and the predictability of such fixed expressions is hypothesized to facilitate fluency and processing speed (Wray 2002), as well as encoding and decoding (Poulsen 2005).

A variety of terms have been presented in the literature, such as collocations, idioms, and routines, to categorize formulaic language. For the purposes of the current study, the term *multi-word phrases* (MWP) will be used to mean specific combinations of words that co-occur together more than would be predicted by chance. The current study does not seek to resolve typological conflicts but to explore the relationship between these items and the development of reading skills. Some or all of the MWPs included on the task used in the current study would be considered either collocations and/or idioms, depending on the definition, and all of the MWPs fall on the spectrum of formulaic language.

Previous research has explored other elements of formulaic language in children, such as the developmental timeline for figurative language comprehension and how learners decipher novel idioms (Cain, Oakhill and Lemmon

2005, Cain and Towse 2008, Levorato and Cacciari 1992, 1995, 1999). Fewer attempts have been made to discretely measure productive elements of MWP knowledge among children; at present only two studies (Crutchley 2007, Smith and Murphy 2014) have done so. Additionally, despite evidence that transparency can have a strong impact on text comprehension for second language (L2) adult learners, previous research with children has not attempted to isolate transparency as a variable or examine its role in productive MWP knowledge. At present, little is known about the relationship between MWP knowledge and reading among young learners, possibly due to the challenges presented both in measuring the construct and testing children.

Review of the literature

MWP vocabulary development

Vocabulary knowledge is relevant at every age of schooling; however, research studies on monolingual children suggest that vocabulary at age 6 is a strong predictor of later reading comprehension and vocabulary size in first grade, and it can account for more than 30% of the variance in reading comprehension by eleventh grade (Cunningham and Stanovich 1997, Scarborough 2001). The larger a child's vocabulary, the faster he or she will learn subsequent words (Nation 2001). Biemiller (2005) reported that children in the lowest quartile for vocabulary knowledge added an average of 570 new root words per academic year, while children with an average-sized vocabulary added 840 or more. Vocabulary is generally a strong predictor of reading and listening comprehension (Freebody and Anderson 1981, Vermeer 1992) and reading in turn has been shown to increase knowledge of word meanings (Anderson and Freebody 1985, Stahl and Nagy 2005, Taguchi 1997).

Productive (or expressive) vocabulary in particular has a strong relationship with reading, possibly because productive vocabulary is hypothesized to have a greater impact on word recognition and require accessing more semantic knowledge than receptive vocabulary (Wise, Sevcik, Morris, Lovett and Wolf 2007). Previous research with young readers has found productive vocabulary to have a stronger correlation with a number of reading skills (phoneme deletion and blending, word identification and non-word identification) (Chiappe et al 2004). Productive vocabulary is thought to have a stronger relationship with reading and pre-reading skills and be a more phonologically representative task than receptive vocabulary (Chiappe et al 2004). Despite the clear link between single word vocabulary and reading, little research thus far has explored MWP language development in monolingual children or English Language Learners (ELLs), despite the fact that these phrases are frequently occurring in daily speech and school curricula (Nippold 1991) and consequently, will likely impact reading.

Transparency in vocabulary development

Non-transparent (opaque) MWPs, such as idioms, present a particular challenge for even advanced adult L2 learners, particularly in relation to literacy (Cooper 1999, Liontas 2002, Martinez and Murphy 2011). In the current study, a *non-transparent phrase* refers to a phrase in which the individual lexical items do not carry their literal meaning; instead the phrase has meaning as a whole. By contrast, a *transparent phrase* is a phrase in which each individual word can be literally interpreted. Both Cooper (1999) and Liontas (2002) examined adult L2 learners' ability to learn opaque idioms and found that the greater phrase non-transparency, the more of a challenge the item presented. Martinez and Murphy (2011) presented L2 adults with matched texts, one text containing MWPs of varying degrees of non-transparency and another with the same words, used individually and literally. Participants performed less accurately on reading comprehension questions for texts containing MWPs (Martinez and Murphy 2011). Notably, when later asked, learners significantly overestimated their own comprehension of passages containing MWPs, indicating they either did not notice the phrases or wrongly assumed they had correctly interpreted the phrases (Martinez and Murphy 2011). Similarly, Bishop (2004) found that adult L2 English learners looked up the meanings of unknown individual lexical items in reading passages significantly more than unknown MWPs and concluded that non-transparent MWPs go unrecognized by L2 learners.

These findings among L2 adults have direct implications for teaching and learning vocabulary in an L2. How might MWPs present challenges for children who are at the developmental age where they acquire the ability to understand that some words (e.g. 'cats', 'dogs') take on a different meaning in proximity to other words (e.g. '*raining* cats and dogs') (Levorato and Cacciari 1992)? Currently, there are no studies that empirically demonstrate if a non-transparent MWP presents more of a challenge for young learners than a transparent one. Similarly, no studies empirically demonstrate a quantifiable effect of phrase transparency on MWP knowledge. Additionally, if MWP transparency is shown to impact phrase knowledge, we may wish to explore not only how MWP knowledge relates to reading, but also the import of MWP transparency in relation to reading outcomes.

English language learners

MWPs may present a particular challenge for ELLs, who account for roughly 17.5% of the pupils enrolled in UK state schooling and come from a variety of language backgrounds; as many as 240 languages are represented in UK schools (National Association for Language Development in the Curriculum 2012). Attainment studies have shown that ELLs in the

UK generally have lower academic achievement than monolingual peers (Burgoyne, Kelly, Whiteley and Spooner 2009), even when controlling for other variables, such as socio-economic status (SES) (National Association for Language Development in the Curriculum 2012). Research shows ELLs often lag behind monolingual peers in measures of reading comprehension (Hutchinson, Whiteley, Smith and Connors 2003, Lesaux, Rupp and Siegel 2007) and vocabulary has been identified as a particular area of knowledge that constrains comprehension (Garcia 1991, Verhoeven 1990). Research has consistently found that vocabulary has a significant impact on ELL school outcomes (Hutchinson et al 2003), and ELL students generally have smaller English-receptive single word vocabularies than monolingual classmates (Bialystok 2010), although counts generally do not consider vocabulary in other languages.

MWPs are likely present in the language ELLs are expected to comprehend. Nippold (1991) found that 6%–10% of the written content in a series of educational reading texts intended for learners 8–12 years old contained idiomatic language, indicating children do need knowledge of these vocabulary items to comprehend age-appropriate school material. It is, therefore, essential to examine this aspect of vocabulary knowledge and its relationship with reading among both monolingual children and ELLs, who may be disproportionately negatively impacted by transparency as a variable.

In order to establish if transparency impacts MWP knowledge among children, knowledge of transparent MWPs must be measured in isolation, thereby minimizing confounding elements, such as phrase frequency and phrase component frequency. Even more challenging is measuring this aspect of vocabulary among ELLs, who may particularly struggle with language comprehension if task instructions are complex and in their non-dominant language. In addition, young ELLs may tire more easily than monolingual peers, due to increased cognitive demands of the task or their inability to demonstrate knowledge fully in a language elaboration task.

Measuring MWPs and measuring transparency

Although MWPs are a well-established area of learner difficulty, relatively few experimental studies have attempted to measure MWPs empirically or quantify the challenge presented by non-transparency, possibly due to the difficulty involved in operationalizing and measuring the construct. Many existing tests are unsuitable for the current study and have not been used with children due to one or more of the following: the unsystematic selection of test items (Bahns and Eldaw 1993, Biskup 1992, Bonk 2001, Farghal and Obiedat 1995, Jaen 2007, Keshavarz and Salimi 2007, Rinaldi 2000); too few test items (Bahns and Eldaw 1993, Farghal and Obiedat 1995); minimal or no measures of reliability for the test items (Biskup 1992, Bahns and Eldaw 1993, Farghal and Obiedat 1995, Rinaldi 2000); lack of attention to frequency of the whole MWP, component words, or both (Biskup 1992, Bahns and Eldaw 1993, Bonk 2001, Farghal and Obiedat 1995, Jaen 2007, Mochizuki 2002, Rinaldi 2000); a possible lack of measurement validity due to internal assumptions (Bonk 2001, Farghal and Obiedat 1995); or a reading-heavy test format requiring comprehension of written passages (Martinez and Murphy 2011).

The conceptualization of MWP knowledge in many existing test formats presents a potential issue; that is, if MWPs are conceptualized as a complete lexical unit, processed, and stored as a whole, then a productive test of MWP knowledge should measure the production of the whole phrase, not a component of the phrase (Revier 2009). However, most existing productive MWP assessments rely on presenting parts of an MWP to elicit the remaining component(s), providing little information about knowledge of the phrase as a whole (Bahns and Eldaw 1993, Bonk 2001, Farghal and Obiedat 1995, Marton 1977).

There are even more challenges involved in measuring and isolating phrase transparency as a variable in order to measure its impact on comprehension, knowledge, and reading for children, particularly children with language limitations. Few previous studies have attempted to discretely measure aspects of phrasal or formulaic language knowledge among children; the exceptions are Crutchley (2007) and Smith and Murphy (2014). Crutchley (2007) explored receptive knowledge of phrasal verbs among learners between ages 6 and 11 using the Assessment of Comprehension and Expression (ACE 6–11), which is a subtest for non-literal language. Verb phrases (i.e. verb + particle) were presented in the context of a sentence with four answer options, including three incorrect distractors containing literal interpretations of the verb. Crutchley (2007) found that as age increased, a greater number of children answered each test item correctly, and by age 11 most knew the target phrasal verbs. Error analysis showed participants treated the phrasal verbs holistically and guessed meanings based on context, rather than on analyzing separate parts (Crutchley 2007). However, this process does not isolate transparency as a variable, because all phrasal verbs presented are non-transparent. Additionally this measure is receptive, and does not measure MWPs.

The only existing measure of MWP knowledge suitable for children, both monolinguals and ELLs, is the task validated by Smith and Murphy (2014). Smith and Murphy (2014) validated the multi-word phrase task (MPT) among British monolingual English speakers and ELLs between the ages of 7 and 10. The MPT contains transparent, semi-transparent and non-transparent verb + object phrases, matched for overall frequency and component frequency, making it ideal for isolating the effect of transparency on item knowledge. A *transparent* (meaning both verb and object carried their literal, first dictionary entry meaning), a *semi-transparent* (meaning the verb no longer had a literal meaning), and a *non-transparent* target phrase (meaning neither verb nor object had a literal meaning and the phrase was entirely opaque) are presented for 10 verbs. This categorization framework is based on the collocation classification presented by Nesselhauf (2005) and adapted by Revier (2009). Overall task performance increased with age for both language groups, the task was found to be suitable for learners from a variety of language backgrounds, and the design was accessible and age appropriate (Smith and Murphy 2014).

Research questions

The current study seeks to address the gap in our understanding regarding the impact of MWP non-transparency, and the relationship between MWP knowledge and the development of reading skills among children (monolinguals and ELLs) by exploring the following research questions in two experiments: What is the impact of transparency on MWP knowledge among monolingual and ELL children between ages 7 and 10 in UK school years Grades 3, 4 and 5? What is the contribution of MWP knowledge and MWP transparency to variance in the performance on tests of reading (e.g. single word reading, reading accuracy, reading rate and reading comprehension) among these young learners?

The current study presents two experiments. The first experiment isolates transparency as a variable and quantifies its impact, and the second measures the relationship between MWP transparency and reading outcomes among young learners.

Data collection procedures

Experiment 1: The impact of MWP transparency on phrase knowledge

Participants

The current study included 108 children between ages 7 and 10 years. The children had no diagnosed special education needs and had been in English language schooling since the ages of 4 or 5 years with a minimum attendance record of 75% during the previous academic year. Although the minimum attendance record may seem low, it was the highest percentage of days attended that I could require and still successfully recruit participants. Children were recruited from three schooling year groups: Year 3 (n = 32), Year 4 (n = 47) and Year 5 (n = 29). There were 68 monolingual English-speaking children and 40 ELLs. The ELLs had diverse language backgrounds, reflecting the general ELL school population: Bengali (n = 14), Chinese (n = 1), Egyptian Arabic (n =1), German (n = 1), Hindi (n = 1), Italian (n =1), Kiswahili (n =1), Konkani (n = 4), Malayalam (n = 2), Portuguese (n = 1), Somali (n = 1), Tagalog (n = 2), Tamil (n = 2), Turkish (n = 2) and Urdu (n = 6). See Table 7.1 for participant information.

Table 7.1 Experiment 1 participant information

	Overall sample	Monolinguals	ELLs
YEAR 3: n (male, female)	32 (12, 20)	20 (8, 12)	12 (8, 4)
Mean age in years	7.54	7.43	7.71
YEAR 4: n (male, female)	47 (18, 29)	31 (13, 18)	16 (5, 11)
Mean age in years	8.33	8.31	8.38
YEAR 5: n (male, female)	29 (10,19)	17 (7, 10)	12 (3, 9)
Mean age in years	9.72	9.88	9.5

Materials

The task used in the present study, the MPT, was specifically designed for and validated with this population (Smith and Murphy 2014). The task presents a sentence prompt, and the test taker must complete the sentence ending by forming a two or three word *verb* + an *object* phrase using words from a 3 × 3 or 3 × 2 puzzle box. The test taker selects a word from each column of the box and these words together make the end phrase. For example, a sentence prompt would read: 'Sam talks to his friends during lessons and doesn't *pay attention*'. The puzzle box contains the components for the target phrase and two distractor verbs and two distractor objects. In the above example question, the puzzle box would contain the word *pay* in one column and *attention* in the other. No other 'real' phrases can be formed with words in the puzzle box. Test takers receive one point for a correct answer and no partial credit is given, as the goal of the test is to measure knowledge of the whole phrases. (See the Appendix for test instructions and example test questions.)

The MPT was developed for use with young children and was validated with British children between ages 7 and 10 (Smith and Murphy 2014). The task was also specifically designed to explore the effect of transparency because it contains matched, comparably frequent phrases made up of comparably frequent components and allows for isolation of transparency as a variable. Target test items were verb + object MWPs containing 10 high-frequency verbs taken from the 1,000 most frequently used words in English according to the British National Corpus (BNC) and objects from the 5,000 most frequently used nouns in English, according to the BNC.

Delineating transparency categories (i.e. *transparent*, *semi-transparent*, and *non-transparent*) is recognized as contentious. Cowie (1998) and Wood (1986) have argued that there is a spectrum of analyzability for formulaic language, and no clear line can be drawn between transparent and non-transparent language. However, the current study does set boundaries on the continuum of transparency. Given that little previous research has addressed the role of MWPs in reading comprehension, separating phrases into categories by degree of transparency enables greater exploration of possible factors that influence how children perform on this task and how multiword phrases contribute to reading comprehension. Test item frequency was controlled for to ensure comparable item familiarity across the transparency bands. A one-way ANOVA was used to compare frequency, as measured by the number of BNC occurrences among the three transparency categories. The results showed no significant differences, (F(2, 30) = 0.275, p = 0.761). See Table 7.2 for all target test items, presented by verb and transparency.

Table 7.2 Target phrases by verb and transparency

Verbs	Transparent	Semi-transparent	Non-transparent
break	break a bone	break the silence	break the ice
carry	carry an umbrella	carry a risk	carry the day
catch	catch mice	catch fire	catch a cold
change	change direction	change trains	change hands
cut	cut a hole	cut jobs	cut corners
hold	hold hands	hold a conversation	hold your tongue
get	get a message	get a taxi	get the sack
pay	pay a bill	pay attention pay a visit	pay the price
run	run a race	run a business run tests	run the show
take	take the money	take your pick take the lead	take sides

Learners were also given the Wechsler Abbreviated Scale of Intelligence for Children (WASI) non-verbal matrices (Wechsler 1992) and a language background questionnaire adapted from Beech and Keys (1997). The primary purpose of administering the WASI was to ensure that there were no significant differences in overall cognitive skills between the two language groups and confirm that all participants were within the ‘typically developing’ range.

Procedures

Assessments were administered individually to each learner during one 20 to 40-minute session. MPT written instructions and example test questions were presented on the front of the testing booklet; the tester read instructions aloud to the test taker and administered example test questions with feedback (Smith and Murphy 2014). All test items were presented in written form and read aloud to test takers, along with each lexical item in the puzzle box. After the test taker selected an answer, the complete sentence with puzzle box answer was read aloud to the test taker. Though the test is presented in written form, the test taker was not technically required to read during the MPT (Smith and Murphy 2014).

Experiment 2: The contribution of MWP knowledge and transparency to reading

Participants

Forty children with a mean age of 8.36 years took part in the current study. Twenty were monolingual English-speaking children and 20 were Bengali-speaking ELLs, all enrolled in two state-run primary schools in England with comparable percentages of pupils receiving free school lunches (a proxy for SES). Participant inclusion criteria were the same as Experiment 1. ELL participants exclusively spoke Bengali at home with all parents or guardians. No participants were receiving out of school non-English additional language or reading instruction, including Arabic reading instruction.

Materials

Along with the MPT and WASI, additional language measures were administered to participants: the Test of Word Knowledge (TOWK) standardized test of receptive and expressive vocabulary (Wiig and Secord 1992), and the York Assessment of Reading Comprehension (YARC) (Snowling, Stothard, Clarke, Bowyer-Crane, Harrington, Truelove, Nation and Hulme 2009). All learners were given a language background questionnaire adapted from Beech and Keys (1997).

Four different reading scores from the YARC were collected: Single Word Reading, Reading Accuracy, Reading Comprehension and Reading Rate. Learners were presented first with a Single Word Reading Test to determine the appropriate starting level for passage reading, and then were given progressively more complex passages to read aloud. The Single Word Reading Test also provided a Single Word Reading score. After each passage the test taker was asked a series of comprehension questions. Scores were given for the total time taken to read the passage (i.e. rate), the number of errors (i.e. accuracy) and the responses to the comprehension questions (i.e. comprehension) (Snowling et al 2009).

Procedures

Assessments were administered individually to each learner during one 40- minute session. The MPT was administered as described in Experiment 1. All assessments were administered in the same order to all participants.

Results

Experiment 1

Table 7.3 presents participant performance on the MPT. There was a strong effect for transparency (F(2, 101) = 256.38, p < 0.001, h2 = 0.811). Weighted mean contrasts showed that transparent items were answered accurately and significantly more often than semi-transparent items (F(1, 102) = 438.32, p < 0.001, h2 = .811) and semi-transparent items were answered accurately and significantly more than non-transparent items (F(1,102) = 110.78, p < 0.001, h2 = 0.521). This provides evidence of a substantive difference between transparent, semi-transparent and non-transparent MWPs. Transparency did not interact with participant age (F(4, 204) = 2.14, p = 0.078, h2 = 0.04) or language background (F(2, 101) = 0.53, p = 0.59, h2 = 0.01). Across all three groups and for both language groups transparent items remained more accurately answered than semi-transparent items, which in turn were more accurately answered than non-transparent items. This strong effect for transparency indicates a substantial difference in the challenge posed by transparent, semi-transparent, and non-transparent MWPs and provides further evidence that non-transparent MWPs present a disproportionate challenge to learners.

Experiment 2

Table 7.4 shows participant performance on the administered measures. No significant difference in WASI performance was found between the two language background groups. Fixed-order multiple regression models were used to explore the unique contribution of the MPT to variance in YARC outcome scores while controlling for other variables. Separate models were run for each of the four YARC outcome measures (i.e. Single Word Reading, Reading Accuracy, Reading Rate, Reading Comprehension) and for the MPT overall and the individual section scores. This procedure allowed the unique contribution of each MPT section to be explored and prevented collinearity. Adjusted R^2 was used rather than R^2 to account for a smaller sample size. Language background, WASI score, and TOWK scores were entered as separate steps in the regression models before the predictor measures. A negative beta co-efficient for language background indicates ELLs performed more poorly on the outcome variable.

Table 7.5 shows the percentage of variance (the increase in adjusted R^2 with each additonal step) in Single Word Reading scores that can be explained by MPT section scores when controlling for other variables (language background, WASI scores, and TOWK). The overall MPT score accounts for 34% of variance in scores, a significant contribution to the model. Transparent (6%) and non-transparent (6%) section scores also made uniquely significant contributions to the model. Table 7.6 shows the fixed order of regression models for Reading Accuracy. While the MPT overall score accounted for a significant amount of variance (9%), no individual transparency section made a significant contribution to the model. Table 7.7 presents models for Reading Rate scores (bold entries indicate the final step for each of the transparency categories). MPT performance accounted for a significant amount of variance (38%), and both transparent and non-transparent section scores each contributed a significant amount of unique variance (i.e. 15% and 7% respectively). Among Reading Comprehension scores (Table 7.8), total adjusted R^2 showed the entire model accounts for 46% of the variance in Reading Comprehension scores, the most of any model thus far. A significant amount of variance (25%) was explained by the MPT overall score. However, no individual transparency section made a significant, unique contribution.

Table 7.3 Experiment 1 performance means

Assessments		**Mean ± SD**	**Mean ± SD**	**Mean ± SD**
All	Overall (out of 33)	20.34 ± 5.74	22.44 ± 4.99	16.78 ± 5.18
	Transparent (% score)	85.09 ± 16.60	90.59 ± 12.44	75.75 ± 18.66
	Semi-transparent (% score)	61.92 ± 22.38	69.38 ± 19.55	49.23 ± 21.32
	Non-transparent (% score)	37.78 ± 20.06	44.26 ± 19.87	26.75 ± 15.09
Year 3	Overall (out of 33)	17.97 ± 5.31	19.30 ± 5.96	15.75 ± 3.08
	Transparent (% score)	79.69 ± 17.13	82.50 ± 18.03	75.00 ± 15.08
	Semi-transparent (% score)	53.06 ± 20.79	57.65 ± 23.70	45.42 ± 12.06
	Non-transparent (% score)	30.94 ± 18.38	35.50 ± 20.89	23.33 ± 9.85
Year 4	Overall (out of 33)	20.06 ± 5.71	22.77 ± 3.73	14.81 ± 5.26
	Transparent (% score)	85.53 ± 17.67	92.26 ± 7.17	72.50 ± 24.08
	Semi-transparent (% score)	61.23 ± 22.85	72.52 ± 15.47	39.38 ± 18.70
	Non-transparent (% score)	34.68 ± 17.55	41.61 ± 16.75	21.25 ± 9.57
Year 5	Overall (out of 33)	23.41 ± 4.96	25.53 ± 3.53	20.42 ± 5.16
	Transparent (% score)	90.34 ± 12.39	97.06 ± 5.88	80.83 ± 13.11
	Semi-transparent (% score)	72.79 ± 19.08	77.47 ± 14.98	66.17 ± 22.78
	Non-transparent (% score)	50.34 ± 20.61	59.41 ± 16.00	37.50 ± 20.06

Note: SD = Standard deviation

Table 7.4 Experiment 2 participant information and performance means

		Overall sample Mean ± SD	Monolinguals Mean ± SD	ELLs Mean ± SD
N (male, female)		40 (23, 17)	20 (10, 10)	20 (13, 7)
Age in years		8.36 ± .12	8.31 ± .14	8.38 ± .15
Assessments				
MPT	Overall	17.40 ± 5.96	20.2 ± 4.95	14.6 ± 5.65
	Transparent (% score)	75.00 ± 21.96	85.00 ± 16.70	102.45 ± 13.46
	Semi-transparent (% score)	51.1 ± 22.05	60.75 ± 20.71	103.91 ± 10.83
	Non-transparent (% score)	33.25 ± 16.39	38.00 ± 14.36	103.68 ± 13.21
TOWK	Expressive	15.23 ± 3.2	17.3 ± 1.98	13.15 ± 2.83
	Receptive	20.28 ± 4.33	21.75 ± 3.27	18.8 ± 4.82
YARC	Reading Accuracy	43.95 ± 9.32	45.6 ± 9.73	42.3 ± 8.83
	Reading Comprehension	53.95 ± 10.23	58.45 ± 7.07	49.45 ± 11.05
	Reading Rate	59.83 ± 18.15	65.3 ± 12.59	54.35 ± 21.32
	Single Word Reading	35.05 ± 10.40	35.4 ± 8.76	34.7 ± 12.04
WASI	Non-verbal reasoning ability	14.2 ± 5.96	14.85 ± 5.43	13.55 ± 6.52

Table 7.5 Fixed order regression analysis for MPT contribution to YARC Single Word Reading

MPT		Standardized beta	Adjusted R^2 change	F change	Sig. F change
Variables entered at each step					
Step 1: Language background		−0.03	−0.03	0.04	0.840
Step 2: WASI		0.29	0.04	3.35	0.080
Step 3: TOWK Expressive		0.35	0.08	2.86	0.100
Step 4: TOWK Receptive		0.05	0.06	0.09	0.760
Step 5: MPT		0.8	0.4	20.54	0.000***
By transparency category					
Assessment	**Variables entered at each step**	**Standardized beta**	**Adjusted R^2 change**	**F change**	**Sig. F change**
Transparent	**Step 1: Language background**	−0.03	−0.03	0.04	0.840
	Step 2: WASI	0.29	0.04	3.35	0.080
	Step 3: TOWK Expressive	0.35	0.08	2.86	0.100
	Step 4: TOWK Receptive	0.05	0.06	0.09	0.760
	Step 5: Semi-transparent	0.63	0.28	11.71	0.000***
	Step 6: Non-transparent	0.43	0.33	3.52	0.070
	Step 7: Transparent	0.43	0.39	4.64	0.040*
Semi-transparent	**Step 1: Language background**	−0.03	−0.03	0.04	0.840
	Step 2: WASI	0.29	0.04	3.35	0.080
	Step 3: TOWK Expressive	0.35	0.08	2.86	0.100
	Step 4: TOWK Receptive	0.05	0.06	0.09	0.760
	Step 5: Transparent	0.62	0.29	12.76	0.001***
	Step 6: Non-transparent	0.48	0.41	7.80	0.010**
	Step 7: Semi-transparent	0.06	0.39	0.60	0.810
Non-transparent	**Step 1: Language background**	−0.03	−0.03	0.04	0.840
	Step 2: WASI	0.29	0.04	3.35	0.080
	Step 3: TOWK Expressive	0.35	0.08	2.86	0.100
	Step 4: TOWK Receptive	0.05	0.06	0.09	0.760
	Step 5: Transparent	0.62	0.29	12.76	0.000***
	Step 6: Semi-transparent	0.38	0.33	3.06	0.090
	Step 7: Non-transparent	0.45	0.39	4.27	0.050

*Note: * p <0.05; ** p < 0.01; *** p < 0.001*

Table 7.6 Fixed order regression analysis for MPT contribution to YARC Reading Accuracy

MPT		Standardized beta	Adjusted R^2 change	F change	Sig. F change
Variables entered at each step					
Step 1: Language background		−0.18	0.01	1.26	0.270
Step 2: WASI		0.24	0.04	2.27	0.140
Step 3: TOWK Expressive		0.45	0.13	4.86	0.030
Step 4: TOWK Receptive		−0.09	0.11	0.29	0.590
Step 5: MPT		0.45	0.2	4.94	0.000***
By transparency category					
	Variables entered at each step	**Standardized beta**	**Adjusted R^2 change**	**F change**	**Sig. F change**
Transparent	**Step 1: Language background**	−0.18	0.01	1.26	0.270
	Step 2: WASI	0.24	0.04	2.27	0.140
	Step 3: TOWK Expressive	0.45	0.13	4.86	0.030
	Step 4: TOWK Receptive	−0.09	0.11	0.29	0.590
	Step 5: Semi-transparent	0.00	0.12	4.36	0.040*
	Step 6: Non-transparent	0.45	0.24	3.35	0.080
	Step 7: Transparent	0.16	0.23	0.49	0.490
Semi-transparent	**Step 1: Language background**	−0.18	0.01	1.26	0.270
	Step 2: WASI	0.24	0.04	2.27	0.140
	Step 3: TOWK Expressive	0.45	0.13	4.86	0.030
	Step 4: TOWK Receptive	−0.09	0.11	0.29	0.590
	Step 5: Transparent	0.16	0.16	2.80	0.104
	Step 6: Non-transparent	0.45	0.17	5.47	0.030*
	Step 7: Semi-transparent	0.00	0.23	0.00	0.990
Non-transparent	**Step 1: Language background**	−0.18	0.01	1.26	0.270
	Step 2: WASI	0.24	0.04	2.27	0.140
	Step 3: TOWK Expressive	0.45	0.13	4.86	0.030
	Step 4: TOWK Receptive	−0.09	0.11	0.29	0.590
	Step 5: Transparent	0.32	0.16	2.80	0.100
	Step 6: Semi-transparent	0.33	0.17	1.78	0.190
	Step 7: Non-transparent	0.45	0.23	3.40	0.070

*Note: * p <0 .05; *** p < 0.001*

Table 7.7 Fixed order regression analysis for MPT contribution to YARC Reading Rate

MPT	Standardized beta	Adjusted R^2 change	F change	Sig. F change
Variables entered at each step				
Step 1: Language background	−0.31	0.07	3.91	0.060
Step 2: WASI	0.01	0.04	0.00	0.860
Step 3: TOWK Expressive	0.18	0.04	0.69	0.410
Step 4: TOWK Receptive	0.15	0.03	0.77	0.390
Step 5: MPT	0.84	0.41	23.56	0.000***

By transparency category

	Variables entered at each step	Standardized beta	Adjusted R^2 change	F change	Sig. F change
Transparent	**Step 1: Language background**	−0.31	0.07	3.91	0.060
	Step 2: WASI	0.01	0.04	0.00	0.860
	Step 3: TOWK Expressive	0.18	0.04	0.69	0.410
	Step 4: TOWK Receptive	0.15	0.03	0.77	0.390
	Step 5: Semi-transparent	−0.01	0.25	11.43	0.000
	Step 6: Non-transparent	0.46	0.31	3.64	0.070
	Step 7: Transparent	**0.52**	**0.46**	**7.21**	**0.010****
Semi-transparent	**Step 1: Language background**	−0.31	0.07	3.91	0.060
	Step 2: WASI	0.01	0.04	0.00	0.860
	Step 3: TOWK Expressive	0.18	0.04	0.69	0.410
	Step 4: TOWK Receptive	0.15	0.03	0.77	0.390
	Step 5: Transparent	0.52	0.32	16.24	0.000***
	Step 6: Non-transparent	0.46	0.43	7.60	0.010**
	Step 7: Semi-transparent	**−0.01**	**0.42**	**0.00**	**0.980**
Non-transparent	**Step 1: Language background**	−0.31	0.07	3.91	0.060
	Step 2: WASI	0.01	0.04	0.00	0.860
	Step 3: TOWK Expressive	0.18	0.04	0.69	0.410
	Step 4: TOWK Receptive	0.15	0.03	0.77	0.390
	Step 5: Transparent	0.68	0.32	16.24	0.000***
	Step 6: Semi-transparent	0.33	0.35	2.30	0.139
	Step 7: Non-transparent	**0.46**	**0.42**	**4.81**	**0.036***

Note: * $p < .05$; ** $p < .01$; *** $p < .001$

Table 7.8 Fixed order regression analysis for MPT contribution to YARC Reading Comprehension

MPT		**Standardized beta**	**Adjusted R^2 change**	**F change**	**Sig. F change**
Variables entered at each step					
Step 1: Language background		−0.45	0.18	9.42	0.000
Step 2: WASI		0.14	0.18	0.96	0.330
Step 3: TOWK Expressive		0.37	0.22	3.70	0.060
Step 4: TOWK Receptive		−0.03	0.21	0.04	0.850
Step 5: MPT		0.69	0.46	17.13	0.000***
By transparency category					
	Variables entered at each step	**Standardized beta**	**Adjusted R^2 change**	**F change**	**Sig. F change**
Transparent	**Step 1: Language background**	−0.45	0.18	9.42	0.000
	Step 2: WASI	0.14	0.18	0.96	0.330
	Step 3: TOWK Expressive	0.37	0.22	3.7	0.060
	Step 4: TOWK Receptive	−0.03	0.21	0.04	0.850
	Step 5: Semi-transparent	0.25	0.41	12.71	0.001***
	Step 6: Non-transparent	0.36	0.44	2.93	0.100
	Step 7: Transparent	0.16	0.44	0.68	0.420
Semi-transparent	**Step 1: Language background**	−0.45	0.18	9.42	0.000
	Step 2: WASI	0.14	0.18	0.96	0.330
	Step 3: TOWK Expressive	0.37	0.22	3.70	0.060
	Step 4: TOWK Receptive	−0.03	0.21	0.04	0.850
	Step 5: Transparent	0.16	0.31	6.14	0.020*
	Step 6: Non-transparent	0.36	0.44	8.43	0.010**
	Step 7: Semi-transparent	0.25	0.44	0.99	0.330
Non-transparent	**Step 1: Language background**	−0.45	0.18	9.42	0.000
	Step 2: WASI	0.14	0.18	0.96	0.330
	Step 3: TOWK Expressive	0.37	0.22	3.70	0.060
	Step 4: TOWK Receptive	−0.03	0.21	0.04	0.850
	Step 5: Transparent	0.42	0.32	6.14	0.020*
	Step 6: Semi-transparent	0.51	0.40	6.02	0.020*
	Step 7: Non-transparent	0.36	0.44	3.02	0.090

*Note: * $p < .05$; ** $p < .01$; *** $p < .001$*

Discussion

The current study quantifies transparency as a variable and explores the relationship between productive knowledge of verb + object MWPs and reading performance among young learners. Findings revealed a strong effect for transparency on item difficulty, as was predicted. Transparency seems to be a powerful variable in determining how children respond to MWPs. Non-transparent MWPs present a disproportionate challenge when compared to their transparent counterparts, which is significant considering how commonly they occur in the lexicon. These findings support generative theories, such as those proposed by Nesselhauf (2005) and Erman (2009) that posit a distinct difference between non-transparent, idiomatic MWPs and transparent MWPs that simply co-occur frequently. This result also provides support for the assertion made by Cain and Towse (2008) that the two possible meanings (one literal and one figurative) are what make non-transparent MWPs challenging.

Additionally, MWP knowledge clearly has a relationship with reading, distinct from single word vocabulary. MPT performance accounted for a significant amount of unique variance in Single Word Reading (34%), Accuracy (9%), Reading Rate (38%) and Reading Comprehension (25%), when controlling for other variables. Performance on the MPT predicted more variability than expressive or receptive single word vocabulary, neither of which made a significant contribution when controlling for MWP knowledge. Transparent item scores accounted for a significant amount of variance in Single Word Reading (6%) and Reading Rate (15%); non-transparent item scores made a significant contribution to the Single Word Reading model (6%) and the Reading Rate model (7%), when controlling for other variables.

It is not surprising that MWP knowledge made a significant contribution to YARC Reading Rate; previous research among adults has shown greater MWP knowledge correlates with faster language processing speed in monolinguals and L2 English learners (Arnon and Snider 2010, Wray 2002). The presence of known formulaic language speeds up reading; formulaic sequences have been shown to be read more quickly than non-formulaic equivalents by adults (Siyanova-Chanturia, Conklin and Schmitt 2011) and the presence of formulaic language is thought to facilitate processing by easing the burden on memory (Martinez and Schmitt 2012). This supports similar findings that greater MWP knowledge correlates with language processing speed in adult monolingual English speakers and L2 English learners (Arnon 2009, Wray 2002, 2009). However, it is not yet clear if knowing more formulaic language speeds up reading for young learners or if both faster reading and knowledge of phrase items like those on the MPT might both be the result of another element, such as amount of time spent in the practice of reading. It is interesting, however, that the semi-transparent item scores alone did not account for a significant amount of unique variance in any of the reading tests. This finding could perhaps indicate a difference in the skills used to answer these types of items. For example, perhaps semi-transparent items were similar enough to either transparent items or non-transparent items and, therefore, did not make a unique contribution.

While single word vocabulary is already well acknowledged as significant for reading, the current study contributes new findings that MWP vocabulary knowledge also has a relationship with reading, distinct from single word vocabulary. It has been argued that MWP vocabulary is different from single word vocabulary from a neurological and grammatical perspective and this uniqueness necessitates separate investigation. Assumptions regarding knowledge of individual lexical items cannot be applied to MWP knowledge (Wray 2002). The results from this research support this assertion: MPT scores accounted for more variability in reading outcomes than single word vocabulary scores.

The relationship between MWP knowledge and reading could reflect a number of things; general vocabulary knowledge and general exposure to written text might increase both reading skill and MWP knowledge. It is possible that more frequent readers will encounter more language in general and consequently more MWPs and, therefore, would perform better when given a measure of MWP knowledge. Ideally, there should be no task overlap between the MPT and the YARC. Because a test taker can complete the MPT without reading, learners likely rely on aural and visual processing. It is possible that the tasks are more easily completed by children who are stronger readers and able to read the items in the puzzle box while formulating a response. The regression analyses only controlled for vocabulary and language background; there are many other aspects of language and many other skills that still might contribute to both reading outcomes and MPT outcomes, such as semantic knowledge, higher order text integration skills, and the ability to incorporate new language information and reconcile it with previous information. A number of elements in this relationship remain unknown; consequently, the results should be interpreted cautiously.

These findings provide evidence of the import of MWP vocabulary and underscore the need for further investigation. There are clear limitations to the current exploratory study. The second experiment contained a limited sample size and included ELLs from only one language background; therefore, the findings cannot be generalized to the broader population. Additionally, the MPT is not a standardized measure, although it was developed for UK learners and has been through a process of validation (Smith and Murphy 2014). Despite these constraints, findings provide increased insight into MWP vocabulary development among young learners and emphasize a need for greater investigation in this area of vocabulary development, including its relation to literacy skill.

Implications and further research

It is hoped that the current findings may provide a springboard for future research. Identifying the predictive role of MWP knowledge in reading comprehension suggests that more attention to this aspect of vocabulary development may improve reading comprehension in both ELL and monolingual children. The emphasis for future research should be placed on designing research studies that can help us better understand how MWP knowledge develops, more precisely identifying impact on reading outcomes and options for focusing on MWPs for vocabulary instruction and evaluation. It is likely that there is a reciprocal relationship between reading and MWP knowledge; learners who read more might encounter greater numbers of MWPs. Future research may consider examining this relationship further. Idioms and MWPs are often explicitly taught in L2 English classes. Given the apparent evidence of the relationship of MWPs with reading comprehension shown in the results from this study, it may be worth considering more explicit teaching of MWPs, particularly non-transparent, as vocabulary items in elementary and ELL classrooms. Additionally, it may be worthwhile to review curricula to evaluate the presence of non-transparent MWPs in materials intended for young learners and consider appropriateness of the materials and the difficulty level.

There are many potential areas on which to focus further research into developing measures of MWPs. One productive research direction would be to consider phrase frequency in child-directed speech or child-produced speech. A future measure could be to compile a corpus of the MWPs in child-directed speech or create target test items using phrases found in age-appropriate curricula or reading materials and use phrase prevalence in these materials to assess the frequency. When moving forward with creating measurement tools for this age group and this element of language, an important element seems to be ensuring that the measure elicits holistic knowledge of the phrases without imposing too much on the child. If the ability to understand non-literal language is still developing until as late as the age of 10, as hypotheses like those of Levorato and Cacciari (1992) suggest, then an overly complex assessment could place a huge demand on what is already a confusing area for the learner. This should be kept at the forefront of testing goals and the present study has shown this to be possible.

References

Anderson, R C and Freebody, P (1985) Vocabulary knowledge, in Singer, H and Ruddell, R B (Eds) *Theoretical Models and Processes of Reading*, Newark: International Reading Association, 343–371.

Arnon, I (2009) *Starting Big: The Role of Multi-word Phrases in Language Learning and Use*, PhD thesis, Stanford University, available online: purl.stanford.edu/zs652gt8462

Arnon, I and Snider, N (2010) More than words: Frequency effects for multi-word phrases, *Journal of Memory and Language* 62, 67–82.

Bahns, J and Eldaw, M (1993) Should we teach EFL students collocations? *System* 8, 101–114.

Beech, J R and Keys, A (1997) Reading, vocabulary and language preference in 7- to 8-year old bilingual Asian children, *British Journal of Educational Psychology* 67, 405–414.

Bialystok, E (2010) Global-local and trail-making tasks by monolingual and bilingual children: Beyond inhibition, *Developmental Psychology* 46, 93–105.

Biemiller, A (2005) Size and sequence in vocabulary development: Implications for choosing words for primary grade instruction, in Hiebert, A and Kamil, A (Eds) *Teaching and Learning Vocabulary: Bringing Research to Practice*, Mahwah: Lawrence Erlbaum, 223–242.

Bishop, H (2004) Noticing formulaic sequences: A problem of measuring the subjective, *LSO Working Papers in Linguistics* 4, 15–19.

Biskup, D (1992) L1 influence on learners' renderings of English collocations: A Polish/German empirical study, in Arnaud, P J L and Béjoint, H (Eds) *Vocabulary and Applied Linguistics*, Basingstoke: Macmillan, 85–93.

Bonk, W J (2001) Testing ESL learners' knowledge of collocations, in Hudson, T and Brown, J D (Eds) *A Focus on Language Test Development: Expanding the Language Proficiency Construct Across a Variety of Tests*, Technical Report number 21, Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center, 113–142.

Burgoyne, K, Kelly, J, Whiteley, H and Spooner, A (2009) The comprehension skills of children learning English as an additional language, *British Journal of Educational Psychology* 79, 735–74.

Cain, K and Towse, A (2008) To get hold of the wrong end of the stick: Reasons for poor idiom understanding in children with reading comprehension difficulties, *Journal of Speech, Language and Hearing Research* 51, 1,538–1,549.

Cain, K, Oakhill, J and Lemmon, K (2005) The relation between children's reading comprehension level and their comprehension of idioms, *Journal of Experimental Child Psychology* 90, 65–87.

Chiappe, P, Chiappe, D L and Gottardo, A (2004) Vocabulary, context and speech perception among good and poor readers, *Educational Psychology* 24, 825–843.

Cooper, T C (1999) Processing of idioms by L2 learners of English, *TESOL Quarterly* 33, 233–262.

Cowie, A P (1998) Phraseological dictionaries: Some East-West comparisons, in Cowie, A P (Ed) *Phraseology*, Oxford: Clarendon Press, 209–228.

Crutchley, A (2007) Comprehension of idiomatic verb + particle constructions in 6- to 11-year-old children, *First Language* 27 (3), 203–226.

Cunningham, A E and Stanovich, K E (1997) Early reading acquisition and its relation to reading experience 10 years later, *Developmental Psychology* 33, 934–945.

Erman, B (2009) Formulaic language from a learner perspective: What the learner needs to know, in Corrigan, R, Moravcisk, E, Ouali, H and Wheatley, K (Eds) *Formulaic Language: Acquisition, Loss, Psychological Reality, and Functional Explanations Volume 2*, Philadelphia: John Benjamins Publishing Company, 324–346.

Erman, B and Warren, B (2000) The idiom principle and the open choice principle, *Text* 20, 29–62.

Farghal, M and Obeidat, H (1995) Collocations: A neglected variable in EFL, *IRAL* 33 (4), 315–331.

Freebody, P and Anderson, R C (1981) *Effects of Vocabulary Difficulty, Text Cohesion, and Schema Availability on Reading Comprehension*, Center for the Study of Reading Technical Report 225, Champaign: University of Illinois.

García, G E (1991) Factors influencing the English reading test performance of Spanish-speaking Hispanic children, *Reading Research Quarterly* 26, 371–392.

Hutchinson, J, Whiteley, H, Smith, C and Connors, L (2003) The developmental progression of comprehension-related skills in children learning, *EAL Journal of Research in Reading* 26, 19–32.

Jaen, M (2007) A corpus-driven design of a test for assessing the ESL collocational competence of university students, *International Journal of English Studies* 7 (2), 127–147.

Keshavarz, M H and Salimi, H (2007) Collocational competence and cloze test performance: A study of Iranian EFL learners, *International Journal of Applied Linguistics* 17, 81–92.

Lesaux, N K, Rupp, A A and Siegel, L S (2007) Growth in reading skills of children from diverse linguistic backgrounds: Findings from a 5-Year longitudinal study, *Journal of Educational Psychology* 99, 821–834.

Levorato, M and Cacciari, C (1992) Children's comprehension and production of idioms: The role of context and familiarity, *Journal of Child Language* 19, 415–433.

Levorato, M and Cacciari, C (1995) The effects of different tasks on the comprehension and production of idioms in children, *Journal of Experimental Child Psychology* 60, 261–283.

Levorato, M and Cacciari, C (1999) Idiom comprehension in children: Are the effects of semantic analyzability and context separable? *The European Journal of Cognitive Psychology* 11, 51–66.

Liontas, J (2002) Context and idiom understanding in second languages, *EUROSLA Yearbook* 2, 155–185.

Martinez, R and Murphy, V (2011) The effect of frequency and idiomaticity on second language reading comprehension, *TESOL Quarterly* 45, 267–290.

Martinez, R and Schmitt, N (2012) A phrasal expressions list, *Applied Linguistics* 33 (3), 299–320.

Marton, W (1977) Foreign vocabulary learning as problem No. 1 of language teaching at the advanced level, *Interlanguage Studies Bulletin* 2, 33–57.

Mochizuki, M (2002) Explorations of two aspects of vocabulary knowledge: Paradigmatic and collocational, *Annual Review of English Language Education in Japan* 13, 121–148.

Nation, I S P (2001) *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press.

Nation, K and Snowling, M J (1998) Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties, *Journal of Memory and Language* 39, 85–101.

National Association for Language Development in the Curriculum (2012) *EAL Pupils: The Latest Statistics about EAL Learners in Our Schools*, available online: www.naldic.org.uk/research-and-information/eal-statistics/eal-pupils

Nesselhauf, N (2005) *Collocations in a Learner Corpus*, Amsterdam: John Benjamins Publishing Company.

Nippold, M A (1991) Evaluating and enhancing idiom comprehension in language-disordered students, *Language, Speech and Hearing Services in Schools* 22, 100–106.

Nippold, M A (1998) *Later Language Development: The School-aged and Adolescent Years*, Austin: Pro-Ed.

Poulsen, S (2005) *Collocations as a language resource (a functional and cognitive study in English phraseology)*, unpublished PhD thesis, University of Southern Denmark.

Revier, R L (2009) Evaluating a new test of whole English collocations, in Barfield, A and Gyllstad, H (Eds) *Researching Collocations in Another Language*, Basingstoke: Palgrave Macmillan, 125–138.

Rinaldi, M (2000) Pragmatic comprehension in secondary school-aged students with specific developmental language disorder, *International Journal of Language and Communication Disorders* 35 (1), 1–29.

Scarborough, H S (2001) Connecting early language and literacy to later reading (dis)abilities: Evidence, theory and practice, in Neuman, S B and Dickinson D K (Eds) *Handbook of Early Literacy Research: Volume 1*, New York: The Guilford Press, 97–110.

Sinclair, J (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Siyanova-Chanturia, A, Conklin, K and Schmitt, N (2011) Adding more fuel to the fire: An eye-tracking study of idiom processing by native and nonnative speakers, *Second Language Research* 27, 251–272.

Smith, S and Murphy, V A (2014) Measuring productive elements of multi-word phrase vocabulary knowledge among children with English as an additional or only language, *Reading and Writing* 28 (3), 1–23.

Snowling, M J, Stothard, S E, Clarke, P, Bowyer-Crane, C, Harrington, A, Truelove, E, Nation, K and Hulme, C (2009) *York Assessment of Reading for Comprehension*, London: GL Assessment.

Stahl, S and Nagy, W (2005) *Teaching Word Meaning*, Mahwah: Lawrence Erlbaum.

Taguchi, E (1997) The effects of repeated readings on the development of lower identification skills of FL readers, *Reading in a Foreign Language* 11, 97–119.

Verhoeven, L (1990) Acquisition of reading in a second language, *Reading Research Quarterly* 25, 90–114.

Vermeer, A (1992) Exploring the second language learner lexicon, in Verhoeven, L and DeJong, J (Eds) *The Construct of Language Proficiency: Applications of Psychological Models to Language Assessment*, Amsterdam/Philadelphia: John Benjamins Publishing Company, 147–162.

Wechsler, D (1992) *Wechsler Individual Achievement Test*, San Antonio: The Psychological Corporation.

Wiig, E H and Secord, W (1992) *Test of Word Knowledge*, San Antonio: The Psychological Corporation.

Wise, J C, Sevcik, R A, Morris, R D, Lovett, M W and Wolf, M (2007) The relationship among receptive and expressive vocabulary, listening comprehension, pre-reading skills, word identification skills, and reading comprehension by children with reading disabilities, *Journal of Speech, Language, & Hearing Research* 50, 1,093–1,109.

Wood, M M (1986) *A Definition of Idiom*, Bloomington: Indiana University Linguistics Club.

Wray, A (2002) *Formulaic Language and the Lexicon*, Cambridge: Cambridge University Press.

Wray, A (2009) Formulaic sequences and language disorder, in Ball, M J, Perkins, M R, Müller, N and Howard, S (Eds) *The Handbook of Clinical Linguistics*, Oxford: Blackwell Publishing, 184–197.

Appendix

MPT instructions

FINISH THE SENTENCE
Name ____________________________________

I'm going to show you some sentences missing their endings. Your job is to make the end of the sentence using the words in the puzzle box below. Circle one word from each colored set of words. Those words need to go together to make the end of the sentence. I'll read everything out loud to you, including the words in the boxes. Don't worry if you think you can't find a perfect answer, you can just choose the words you think go together the best or have a guess.

Let's do an example together.

Sam talks to his friends during lessons and doesn't __________________.	
break	studies
catch	[✓ attention]
[✓ pay]	work

Now, let's have a look at the other answers we can make. They don't seem very good, do they?

8 Teacher perspectives on social language assessment

Kimberly K Woo

Teachers College, Columbia University, New York City, US

Motivation for the research

The school setting presents a unique set of linguistic demands for both social (interpersonal) and academic functioning, which can be challenging for native English speakers and second language learners alike, especially when entering school for the first time (Schleppegrell 2001). A dividing line is frequently drawn between 'social' and 'academic' language use, a distinction especially evident in the instruction and assessment of second and foreign languages. Cummins' (1980) proposal is that language proficiency falls along two dimensions, one academically oriented (Cognitive/Academic Language Proficiency (CALP)) and one socially oriented (Basic Interpersonal Communication Skills (BICS)). This conception of language proficiency has had an enormous influence on how the sequence and nature of second and foreign language development are understood (Gu 2014).

Review of the literature

In the decades since BICS and CALP were initially proposed, the conceptual distinction between the dimensions has been emphasized (Cummins 1999). Social and academic language are often discussed as a binary, with each defined or characterized in relation to the other. For instance, where CALP is described as more difficult due to its greater relative cognitive load and contextually embedded nature, BICS is primarily characterized by being less so (e.g. Cummins 1991, Fang, Schleppegrell and Cox 2006, Hawkins 2005).

In recent years, growing interest in academic language in instruction and assessment has led to a considerable body of literature exploring the nature of academic language and seeking to establish definitive characteristics, frameworks, or schema to describe the construct (Bailey (Ed) 2007, Bailey and Huang 2011, Schleppegrell 2004). This work has considered academic language across the K-12 range, including early grade settings (Lucero 2012). Bunch (2009) and Shiel, Cregan, McGough and Archer (2012) highlight the varied ways in which academic language has been conceptualized (i.e. focusing on aspects of language use, such as the level of contextualization and cognitive demand; vocabulary and grammatical structures; the degree to which information is conveyed; the level of familiarity between participants; and the level of formality). Despite this work, the definition and features of academic language remain under dispute and continue to vary considerably among scholars, researchers, and education professionals (Bunch 2009, Lucero 2012, Maxwell 2013). Contributing to the debate over what defines academic language is the interplay between the academic and social domains of language. In a summary of Leung's (2014) review essay, Haneda (2014:90) writes:

> Classroom interaction does not always singularly focus on the academics (i.e. the content of an academic subject) but is intermixed with many interpersonal asides and other remarks that can only make sense in the context of a particular classroom community. It is through the-back-and-forth movement, between the interpersonal and the academic as well as between the formal and the informal that the teaching and learning of an academic subject is enacted in actual classrooms.

It has been argued that, in practice, the relationship between social and academic language can be much more complex than a binary relationship would suggest. Leung (2014:143) notes that 'the complex and dynamic interactions and communication between teachers and students in the classroom do not always fit with the neat BICS-CALP divide'. Likewise, Bailey and Heritage (2008:15) write:

> The distinction that Cummins (1981) made between everyday, social uses of language, Basic Interpersonal Skills (BICS), and the language used for learning in school, Cognitive Academic Language Proficiency (CALP), has been criticized for equating BICS with simplicity and CALP with complexity (e.g., Bailey, 2007). Social uses of language can be cognitively demanding and take place outside the immediate context of the "here and now" as well . . . we have found the contextual and cognitive demand distinction to inadequately distinguish between academic and social language.

Issues with considering academic and social language as a discrete binary are especially apparent in the instruction and assessment of English language learners (ELLs) in the early grades (e.g. Aukerman 2007, Jeynes 2006). For young ELLs, academic development and social development are recognized as being closely intertwined, and children's social competence is said to influence and/or predict their academic achievement and functioning in school (Arnold, Kupersmidt, Voegler-Lee and Marshall 2012, Halle, Hair, Wandner, McNamara and Chien 2012, McClelland and Morrison 2003, Raver and Zigler 1997). Social interaction and the language it requires are frequently part of curricula and state and professional standards for early grade language instruction, which creates an overlap between the academic and social dimensions of language proficiency and further blurs the dichotomy (Jeynes 2006).

Ambiguity in defining social and academic language is problematic given the role of these constructs in the high-stakes language tests used with young ELLs. In the public school system in the USA, ELLs make up a major portion of the student body, and it is projected to expand to nearly 40% of the student population by 2030 (Borsato and Padilla 2007). Once identified as ELLs, students in grades as early as kindergarten are mandated to participate in annual English language proficiency (ELP) assessments, which typically take the form of large-scale standardized tests. These measures are high stakes in nature, not only because scores are used for federal accountability, but also because they frequently inform decisions regarding the nature and continuation of language support services, which can have an appreciable impact on the instruction students receive and the trajectory of their academic careers (Schappe 2005).

Since the No Child Left Behind Act was passed in 2001, ELP testing in the USA has been based on the presuppositions that there is an academic/social divide for language proficiency and that the primary objective of ELP tests is to target academic language proficiency (Abedi 2007, Albers, Kenyon and Boals 2009, Bailey and Huang 2011, Chalhoub-Deville and Deville 2008). However, some ELP tests, such as the New York State English as a Social Language Achievement Test (NYSLAT) and the World-class Instructional Design and Assessment (WIDA[1]) Consortium's ACCESS for ELLs, include items targeting students' proficiency in spoken social interaction.

Relative to the research exploring the nature of academic language, there have been few examinations of social language as a construct in language instruction and assessment. Given traditional assumptions that underpin the academic and social language proficiency dichotomy and the extensive work on academic language, it would be tempting to argue that social language can be defined in terms of what academic language is not. However, the interactivity that poses a challenge to defining academic language suggests that this assumption cannot be relied upon, and it remains unclear not only what social language test items are measuring, but also what is meant by social language. Given the high-stakes nature of ELP tests, and recognizing that young ELLs are a population especially vulnerable to the consequences of testing (Espinosa 2005, McKay 2006, Pinter 2006), greater clarity surrounding the construct is necessary to help ensure that tests provide meaningful data and so that appropriate inferences can be made.

Research questions

As part of a larger validation study exploring the construct of social language in ELP tests that are used with young ELLs, a small-scale empirical examination was conducted to explore the following research question: how do teachers define and assess social language? Teachers' perspectives and practices were of interest as content-related evidence of validity. Content validity speaks to the extent to which items on a measure are sufficiently and appropriately relevant to and representative of the intended construct, that is, the extent to which items capture all facets of the targeted construct. Teachers' perspectives serve as evidence of the extent to which social language test tasks relate to or represent the construct of social language. Also embedded in this inquiry was an exploration of the role social language plays in early grade classroom instruction and assessment, particularly with ELLs.

Data collection procedures

Data that targeted how teachers defined and assessed social language were collected over an academic year. First, a survey was administered to teachers across five public elementary schools in New York City's Chinatown, a neighborhood selected for its historical significance as a cultural and linguistic immigrant enclave community, as well as its high proportion of early childhood ELLs. The five schools were all part of the same school district and shared similar demographics and curricula. Teachers were eligible for participation if they were the homeroom teacher for a kindergarten or first grade class or if they provided 'push-in' or 'pull-out' English as a second language (ESL) instruction for lower-grade students. With push-in instruction teachers work with ELLs in their regular classrooms; in pull-out instruction ELLs are pulled out of regular classrooms and receive separate instruction. Eligibility was not made contingent on a specific classroom proportion of ELLs as the survey aimed to elicit teacher perspectives across a range of K-1 contexts.

Of the 69 eligible teachers across the five schools, 30 (44%) participated in the survey. Respondents were primarily female (93.3%) and were from an Asian or Chinese/Chinese-American background (70%). Fourteen teachers (47%) were kindergarten homeroom teachers, 11 (37%) were homeroom teachers for first grade classes, and five (17%) taught lower grade push-in/pull-out ESL. With respect to class composition, 64% of the homeroom teachers (16 out of 25) reported that more than half of their students had been identified as ELLs or that they taught self-contained ESL or bilingual classrooms. Teachers participating in the study were highly experienced, with 60% having taught for 10 years or more, and another 13% having between five and 10 years of experience. Over half of the teachers indicated that they had been teaching in their current position for five years or more.

The survey consisted of three parts. The first part collected demographic information about the teachers (e.g. certification information, teaching experience, language background). The second focused on teachers' perceptions and descriptions of social language. Participants responded to open-ended items asking for commentary on how they would describe social language and whether they believed there to be a relationship between social and academic language, elaborating on why they did or did not think there was a relationship. In this section, teachers also completed a series of Likert-like scales to rate the extent to which they believed descriptors, features, abilities, and interactions help identify or are characteristic of social language. The characteristics, descriptors, abilities, and interactions included were informed by traditional discussions of academic and social language (e.g. Bailey (Ed) 2007, Cummins 1980, Schleppegrell 2004), functions of general communicative language proficiency (Bachman 1990), descriptions of social language and interaction on state standards and local report cards, and common oral interactions discussed by Russell and Grizzle (2008) and Canale and Swain (1980). In the third part of the survey, teachers reported on their instruction and assessment practices, indicating the frequency with which they employed common classroom assessments of oral language. Additionally, through a series of agreement and rating scales, teachers were asked to reflect on the kinds of oral language assessments used in their classes, the social or academic nature of classroom activities, and the aspects of language they viewed as most important in evaluating students' social language ability.

In follow-up to the survey, teachers were invited to participate in a semi-structured interview revisiting survey themes. These interviews were intended to supplement survey data with the nuance and detail of individual teachers' experiences, providing confirmation and/or contrast with survey findings. Each interview asked teachers to describe social language and the elements they felt were important in recognizing it in use. Teachers were also prompted to reflect on students' opportunities to use social language in a typical school day, their expectations for students' language proficiency, and their instruction and assessment practices, in general and specific to oral and social language development.

Of the 30 survey respondents, four teachers (13.3%) agreed to participate in the interviews. Three were homeroom teachers of kindergarten classes (Kate, Ma'am, and Amy), and one was the homeroom teacher of a first grade class (Bernadette). More than half of the students in each teacher's class were ELLs. All four teachers were English-dominant with knowledge of conversational Chinese. The interviewed teachers varied in their total teaching experience (from two to 15 years) and in their experience in teaching ELLs in the lower grades (from one to 15 years). All but one interviewed teacher was certified in ESL or bilingual education in addition to holding a childhood common branch license. In the State of New York, teachers who have certification for pre-kindergarten to Grade 6 will have an early childhood license or a common branch license.

Results

Teachers' definitions of social language were examined by considering their responses to survey and interview prompts that explicitly asked for a definition for social language. To encourage more descriptive responses, prompts asked teachers to describe the aspects of children's language use that would signal engagement in social interaction or speaking for social purposes. Of the 30 survey participants, 25 responded to the direct prompt as an open-ended survey item (83.3%) and all four teachers responded to this prompt in interviews.

Open-ended items from the survey and interview transcripts were thematically coded and analyzed using directed content analysis (Hsieh and Shannon 2005) to determine how teachers described social language in terms of six categories: 1) the audience and participants involved, 2) the setting or context, 3) the content, 4) functions or purposes, 5) the distinct features or characteristics of language, and 6) the relationship between social and academic language. These categories were informed by the elements and structure used by Bailey and Heritage (2008) to describe social language, as well as Gee's (2011) discussion of social language(s). These qualitative data were supplemented by quantitative data using descriptive statistics (i.e. frequencies and means) calculated for survey items, which used rating, agreement, and Likert-like scales to determine the characteristics that teachers most frequently associated with social language.

To determine the role of social language in classroom instruction and how teachers approached assessing the construct, the frequencies and means were calculated on survey items in terms of how teachers categorized common classroom activities and interactions and how frequently they used common classroom assessments of oral language to assess social language.

Teacher interviews were also coded for instances during the school day in which students had opportunities to engage in oral language in response to teachers' actions, social language instruction, or assessment.

Teachers' definitions of social language

Participants

In data from both the survey and interviews, teachers commonly expressed the belief that social language use was primarily marked by speech between peers. Of the nine teachers who responded to the open-ended items on the survey, 36% made explicit references to the individuals with whom students spoke when engaging in social language. All nine respondents pointed to peer-to-peer speech; seven did so exclusively, while two indicated that students could also engage in social language with adults. When characterizing classroom interactions (e.g. whole-class, child–child, and teacher–child) as more or less social on a 4-point scale (1 = not social at all, 2 = somewhat social, 3 = very social, and 4 = entirely social), teachers indicated that children's interactions with both their peers and their teachers could be seen as social, though to varying degrees. Consistent with responses to the open-ended item, of the three interactions, child-to-child speech was characterized as the most social, with the majority of respondents (70%) rating these as 'very social' or 'entirely social', compared to the others, which were rated by more than 65% as only 'somewhat social' or 'not social at all'.

Although communication between peers was highly emphasized in teachers' definitions, data also indicated that young children's social language use is not exclusive to this type of interaction. As suggested by both the qualitative and quantitative parts of the survey, teachers believed that students do occasionally engage in social language with adults. This was corroborated in the interviews, although two major points emerged in considering children's interactions with adults as part of social language.

The first point to emerge was whether and to what extent children engage adults in social language use. Child–adult social interaction is dependent on certain conditions, such as the child's relationship with the adult. For example, the engagement depends on how the adult communicates with them, in other words, the connection they have, and whether the adult is perceived as an authority figure. Additionally, due to children's developing maturity, if engagement with adults is considered part of social language, its presence may be indicative of differing social language proficiency at different grade levels. While interviewed teachers believed that the ability to recognize and appropriately speak to different audiences is a skill marking social language proficiency, it is also a skill that younger children struggled with more so than older children.

Amy, one of the kindergarten teachers, observed that her students spoke socially with adults and peers alike 'because they don't know how to distinguish between adults and other children'. She explained further: 'My kids, they're very immature . . . they really don't know the difference between talking to an adult – I mean – most kids when they talk to adults, they talk to them the same way as having a friend'. It was noted that teaching students to differentiate their speech for different audiences was a specific teaching point during the kindergarten year. By first grade, teachers said that students engaged adults less frequently in social interaction, and it was assumed that they would not speak to adults and peers in the same way.

Context or speech setting

Given that context is commonly used to delineate between social and academic use (Bailey (Ed) 2007, Shiel et al 2012), it was of interest to see if and how teachers' definitions were framed by the setting in which speech occurs (e.g. location, time of the day, the activities in which students were engaged). In the binary view of social and academic language, academic language is traditionally deemed 'the language of schooling' (Schleppegrell 2004), while social language is that used anywhere else. To some degree, this was echoed in teachers' survey responses.

Respondents who commented on context in the open-ended survey question often used broad descriptors, including 'non-academic settings'. Students were said to be engaging in social language use if they were speaking during 'play or playtime', in the schoolyard or playground, during choice time, 'outside of school', or at parties. Of these settings, 'play/playtime' was the most frequently indicated, noted by six teachers (55% of those who spoke about the context). In interviews, the context of language emerged more frequently in the discussion of assessment practice than in how teachers characterized or identified social language use.

When directly asked to define social language, setting-related comments echoed traditional notions, thereby reserving academic language for the school environment and placing social language use anywhere else. Interviewed teachers described social language as 'how they would speak outside the classroom and outside of school' and 'what you do at home'.

However, there were data to suggest that teachers did not entirely view social language as exclusively external to the school setting. On the open-ended item, two teachers explicitly stated that social language is language that is used in the classroom or school, and one teacher highlighted use during choice time, a school-day activity. Corroborating these comments, when survey items asked teachers to rate the 'socialness' of common class activities on a 4-point scale from 'not social at all' to 'entirely social', the majority of teachers deemed all classroom activities as at least 'somewhat social' in nature, with the exception of tests. Recognizing that 'socialness' is relative, the classroom activities that were viewed as being most social in nature included morning meetings, think–pair–share (i.e. brainstorm individually, pair up with another student, and share your ideas), pair work, free/choice time, and show and tell. Classroom activities that were seen as less social than other activities included conferencing, reporting to the class, and testing.

Content

On the open-ended item and interviews, teachers identified a range of topics, which they believed to signify social language use. These included talking about wants and needs, likes and dislikes, personal experiences, interests, feelings, friends and friendship, afterschool activities, or home life. Teachers also believed students to be engaging in social language if they were sharing imaginary stories or personal information, if students were playing, or if they used expressions of manners (e.g. language such as ‘please’, ‘excuse me’). Of these topics, the content most frequently associated with social language was students’ expression of ‘needs and wants’.

It was of interest to note that teachers independently identified topics that paralleled those topics featured in a separate section of the survey as representative of common themes in English proficiency tests and early childhood classroom activities. These topics include home and family, school and school activities, friends, likes/dislikes/favorites, and recent events. When asked to rate these themes as either social, academic, neither, or both, the topics that were most frequently identified as social in nature were ‘friends’ and ‘likes/dislikes/favorites’; however, all topics were rated to be, to some degree, both social and academic in nature.

Language functions

In responses to the open-ended survey item, the purpose or intended function of speech was the most frequently occurring indicator of social language use. Sixty percent of responding teachers noted speech objectives in characterizing social language use. Teachers identified a variety of social goals, such as self-expression, communication or interaction, engagement in play, survival, negotiation, sharing information or experiences, making requests, building friendships, greetings and polite speech, giving instructions, and shared storytelling. Of these, the most frequently mentioned were self-expression (53%), communication (47%), interaction (40%), and play (27%).

The commonality among nearly all speech functions identified was an emphasis on speaking for a communicative purpose and the presupposition of interactive engagement with another person. This purpose was an overt theme, given that ‘to communicate’ and ‘to interact’ were among the goals mentioned most frequently by teachers, but also implicit were less-frequently mentioned functions, such as ‘to negotiate’, ‘to build friendships’, or ‘to engage in play’.

Interviewed teachers’ definitions echoed the communicative and interactive nature of social language with statements such as: ‘To me, social language seems about the way they communicate with their friends’ and ‘social language, I would actually consider that more of a way students would communicate . . . how students speak to each other’.

Descriptive characteristics and features

Teachers characterized social language in several ways; most commonly, teachers described the performance of social language using broad terms, such as ‘informal’, ‘natural’, ‘casual’, ‘everyday’, and ‘conversational’. Several respondents pointed to general linguistic features that they felt would signify social language use, such as the use of short sentences, the presence of informal or inconsistent grammar, and students' choice of vocabulary, including use of slang. Others characterized social language in terms of the features they believed indicated effective or proficient usage, such as the ability speak in an 'age-appropriate' manner and 'express themselves without getting frustrated'. Additional skills noted as important in social language use included the ability to maintain a topic throughout conversation, follow social norms or rules, and express their thoughts 'in an organized way'. Teachers' descriptions not only focused on the language they expected students to produce, but also affective and non-verbal signals; several teachers made mention of the role of body language, facial expressions, gestures, and eye contact in social language use.

The survey also presented teachers with 12 descriptors typically used to characterize language under the social/academic dichotomy and asked teachers to rate the descriptors on a 4-point scale (1 = not well, 2 = somewhat well, 3 = well, 4 = very well) on the degree to which they felt each descriptor characterized language used in social situations. The average rating for all but three characteristics was 2.0 (somewhat well) or higher, indicating that teachers felt that nearly all of the listed features described social language to some degree. The two descriptors that teachers identified as best characterizing social language were 'conversational' (mean (M) = 3.21, standard deviation (SD) = 0.62) and 'interactive' (M = 3.07, SD = 0.62). Of the descriptors for social language, these were the only characteristics that had an average rating of 3.0 or higher ('well' or 'very well'). Features least frequently identified as descriptive of social language were rated by less than a quarter of respondents as 'well' or 'very well'. The descriptors for language functions were 'formal', 'abstract', and 'scripted'.

Like many survey respondents, interviewed teachers spoke about social language in broad terms, but explained that it was difficult to identify specific features when many of their students were such low-level ELLs that getting students to speak at all was a major instructional objective. To this end, and again echoing survey responses, interviewed teachers spoke to the non-verbal aspects of social language use.

Bernadette highlighted an affective component – student's comfort level – as a key feature in identifying social language use. She described how, in informal observations, students were more comfortable speaking when engaged in social language, were 'more outgoing and able to communicate', and more 'willing to express their ideas'. On this note, both she and another interviewed teacher expressed that social language was marked by greater fluidity in students' speech, with 'no long pauses or silence'.

Amy also made mention of an affective component, stating: 'Social language proficiency, it's being able to talk with others without all that friction'. She also defined social language in terms of 'whether or not [students] are able to use social cues'. When asked to elaborate, she pointed to students' abilities to interpret and respond to contexts in an appropriately empathetic or co-operative manner, in other words: 'Whether they know how to say the right things to each other if they're upset or if someone fell or, like, being considerate to one another. That kind of thing'.

'Socially appropriate' language also appeared in Kate's interview, though her explanation focused more on linguistic elements, such as word choice, sentence structure, and conciseness: 'Don't say, "Give me", [Say] "please", using words like that . . . It's like, "Please give me that piece of paper", versus, "Hey, hey, hey". . . It's telling them exactly what it is that you want in a very clear concise manner, but you know, politely'.

Of the interviewed teachers, Kate was the only one to specify linguistic elements that she felt characterized social language. These were rarely mentioned by the others and any mentions made tended to be descriptions of what would *not* be part of how they defined social language, such as Amy's statement, 'I don't expect them to be grammatically correct'.

The relative informality of social language was an additional distinguishing feature highlighted in both surveys and interviews. On the survey, this was indicated directly and indirectly in teachers' open-ended responses (e.g. use of the word 'informal', noting use of slang), as well as on the descriptor rating scales, where 'formal' was among the three descriptors rated as describing social language the least well. In interviews, teachers pointed to informality in terms of speaking to different audiences, illustrated by their personal experiences speaking to colleagues as opposed to administration, and in discussing whether interactions between adults and children constitutes social language.

Kate noted that while students attempted to engage her in social interaction on a fairly regular basis, the informality of the exchanges was at times at odds with the level of respect appropriate for a teacher, both for herself and within the culture of the community at large:

> I do draw a line, like, "I'm not your friend – I am here to be your teacher". Maybe after school, there'll be a little less formality, but there's already a formality when you address me as "Miss Kate", right? There's already, like, a line drawn . . . and they acknowledge it, I think, for the most part, even the ELLs. Even if they don't acknowledge it, their parents will be like, "This is your teacher. Don't even go – This is your *laoshi* [teacher]".

The relationship between social and academic language

Given that social language has traditionally been defined in terms of a dichotomy with academic language, it was worthwhile to see the extent to which teachers' definitions followed this paradigm. On both the survey and in the interviews, teachers' initial response to being explicitly asked to define social language was to rely on a distinction between academic and social language. As noted, on the open-ended survey item, social language was specifically characterized using terms such as 'non-academic' or 'outside the school setting'.

A similar pattern was found in the follow-up interviews, where teachers' immediate responses were to produce statements such as, 'When I think "social", I definitely don't think "academic"', or 'For me, social language is all about nonacademic work . . . I see social language completely as non-academic'. Given this dichotomous framework, it was not unusual for teachers to attempt to define social language through an absence of academic language. Teachers frequently relied on contrasting examples to highlight characteristics of social language, such as Kate's characterization of social language as having a relative lack of jargon by pointing to the prevalence of technical vocabulary in academic language.

Although teachers frequently relied upon an academic/social language distinction to inform their definitions, data also suggested ambivalence in the extent to which teachers believed these to be discrete constructs. On true–false survey items asking teachers to rate the degree to which they believed social and academic language to be related or distinct, responses were divided in asserting that 'social and academic languages are different' for ELLs and general language use, and a clear majority (83%) of teachers indicated belief that there was considerable overlap between social and academic language for students in the early grades (K-2). While many (69%) respondents pointed to connection between the ability to interact socially and English language ability, 26 of 29 respondents (90%) felt that it is more important for students to learn to use language for academic purposes than for social interaction. Further, in an open-ended item, teachers were asked to describe their perceptions of the relationship between academic and social language. On this item, teachers frequently indicated that they believed there was a relationship between the two, although the nature of this relationship varied (i.e. the constructs were viewed as complementary, integrated, or supportive). Taken together, these findings indicate that although teachers may not necessarily believe social and academic language are different things, there is a clear value of one over the other.

Teachers' assessment of social language

On both the survey and interviews, teachers were asked to comment on their assessment practices, speaking to the kinds of measures they used to assess social English and the frequency with which these were used. In interviews, teachers were encouraged to speak about any and all assessments used with students, to provide a sense of the assessment experiences of early grade

ELLs, and, when possible, to describe or provide examples of what such assessments look like. The survey focused on the frequency with which teachers used four common classroom assessments to evaluate students' oral social language. The four assessment methods were the following: observations, one-on-one conferencing, class presentations or sharing, and classroom tests. These methods were common to the elementary classroom settings and reflected a range of formality. Teachers were instructed to indicate how frequently they used each assessment using a 7-point scale ranging from 'never used' to 'every day'. Teachers were also instructed to indicate and describe any additional assessment methods or measures they may regularly use.

Although one-on-one conferencing and class presentations were common, observations were the most frequently used assessments, with the majority of teachers stating that that they were conducted daily (64%), and an additional five teachers (18%) indicating that they observed students' social language use several times a week. Classroom tests were the least frequently used, with 48% of teachers noting that they were never used. When asked if they employed other or additional assessments of social oral language, most teachers (70%) indicated that they did not.

The teacher interviews presented a considerable contrast to the survey findings. Although survey responses reported regular assessment of oral academic and social language use, in interviews teachers indicated that assessment of any oral language, much less social oral language, was infrequent, if occurring at all. One teacher claimed to never assess students' oral language, though her later statements indicated that she monitored students' social language use through informal observations. Informal observations were a recurring theme in interviewed teachers' discussions of how they assessed students' social language use and oral language at large.

Consistent with survey responses, data from the interviews showed that observations were the most frequently used all-purpose classroom assessment; however, when it came to assessing social language, these observations were rarely planned and were more typically incidental to events as they occurred over the course of the school day. Certain points in the day were noted as being more amenable for observing social language interactions (e.g. 'morning is where you can catch them'), though because of the unpredictable and 'impulsive' nature of the young students, events in which teachers used to evaluate social interaction were often incidental conflicts or disagreements, and these occurrences were often approached as teachable moments, as well as opportunities for assessment.

Of the interviewed teachers, Bernadette was the most methodical and premeditated in observing and documenting students' social interactions. However, in describing her process, students' language use was only one minor element in her definition of social functioning; her focus was on students' broader participation in the class community (e.g. focus on and attention to classroom activities, interpersonal relationships, and physicality). Furthermore, even she noted: 'Those [observations] are just incidental . . . sometimes, if it just hits me while I'm doing a lesson, when I know that, I'll jot that down'.

In speaking to the opportunities for students to use social language in the typical school day, certain periods were consistently identified as conducive to social language use. Consistent with survey data, these periods included lunchtime, snack time, and trips to the schools' playgrounds. Given such settings, it was unsurprising that assessment of social language was so limited; these periods provided little opportunity for teachers to see social language because these were typically times when teachers were not present. The primary classroom-based setting in which interviewed teachers felt they were able to observe students' social language in use was during choice or center time, a period offered weekly or daily depending on the teacher, in which students engaged in independent and small group activities ranging from desktop games to dramatic play.

Surprisingly, in addition to observations, interviewed teachers indicated that they occasionally relied on students' writing as an assessment of oral and social language proficiency. This practice appeared founded on belief that students 'write the way they speak'. As Ma'am commented, 'I feel like at this age they write what they would say. So by looking at their writing, I'm able to tell what they are able to say'.

Kate offered evidence in favor of this notion, pointing to how the grammatical errors students made in speech, such as missing prepositions, were equally present in their writing, noting, 'They wrote "I like play", and that's exactly how they speak. "I like play". Or, um, the other sentence was, "The dog is wet" rather than, "The dog wet". I realize, because in Chinese there's no prepositions'.

For Ma'am, the use of written language as a means to assess oral language ability may also have been linked to the fact that she had similar goals for her students in both language domains. In commenting on her instructional objectives with regard to students' speaking, she noted that her primary goals for students were to have them produce any language at all and to expand beyond one-word answers: 'It's all about stretching a sentence this year for me . . . I realized students just – "Bathroom". "Rainy". How's the weather? "Rainy". You know, let's expand on that a little bit more'.

In explaining why the instruction and assessment of oral and social language were so infrequent, interviewed teachers pointed to restrictions on available time due to the increased prioritization on academic language at the cost of opportunities to develop and assess students' oral and social interaction.

Ma'am described the situation as: 'It's just basically what's happening to schools. Like kindergarten, there's no play. There's no time for them to play with each other, not enough time to, like, socialize. It's all just academics'.

Similarly, Amy commented: 'Social language wise, I mean – you know what's sad is that I don't even get to focus on too much of the social aspect of it. It's mostly academic now'. She further added that students were deterred from attempting to engage in social language because of limited opportunities for unstructured talk. When asked if students engaged in social language with her, Amy responded: 'When there's time to talk, you know? But, because that time is so limited, they don't usually do that'.

Interviewed teachers were acutely aware of time restrictions and noted that while they often attempted to build time into students' schedules to provide opportunities for social language development, the periods most closely affiliated with social language, such as choice time or snack time, frequently became truncated by academic demands: 'I do make sure they get it every day. Sometimes it's shorter, sometimes it's longer; getting behind [on academic work], it eats into their choice time. So something that should be maybe 40 minutes really turns into like 25 minutes'.

In light of these time limitations and in terms of speaking about assessment of students' social language abilities, teachers were found ultimately to make holistic judgments rather than rely on specific linguistic criteria. For instance, Kate admitted to relying on general impressions and broad indicators, such as how recently students had arrived in the USA and 'if they can follow my directions without me having to cue them'.

Likewise Amy noted: 'I don't really have that much time to observe them . . . I can tell who's having trouble and who's not'. Amy, the teacher with the longest on-grade and ESL experience, highlighted how changing emphasis and demands have influenced her assessment practice:

> I've got to get things done. There really isn't that much time to address it [social interaction] . . . A long time ago I used to have something that goes home . . . I would check off, you know, what they were able to complete, what they were able to do on the day – personal growth, as well as behavior, and academics. But, I don't have time to do those checklists any more.

Discussion

As a small-scale exploration, this study found that, in some ways, teachers' definitions of social language mirrored what might be expected given a traditional binary view of academic and social language proficiency. Teachers' discussion of the construct was often initially couched in contrast to academic language, and it was common for teachers to attempt to define social language through the absence of academic language or through circular descriptions (e.g. 'social language is language used in social situations'). Using these data as starting points, certain features emerged to provide a preliminary working definition for social language (e.g. relative informality, frequent use in non-classroom settings, marked by speech between peers).

However, even given the relatively small sample, teachers expressed a range of views that provided nuance and complexity to this developing definition, which is summarized in Table 8.1.

Table 8.1 Teacher characterization of social language

Context of use (Participants)	Primarily peer-to-peer May occur between students and adults (e.g. parents, teachers), depending on relationship and setting
Context of use (Setting)	Primarily outside of school (home) In school, most likely to occur during transitions, lunch, playtime or recess, choice time, paired activities, and morning meetings
Content and topics	Personal information, interests, and feelings Needs and wants, likes, dislikes, and favorites Friends and friendship Out-of-school activities and home life
Function	Communication: Making requests, sharing information and experiences Self-expression: Expressing preferences, storytelling Interaction: Play, negotiation, making and building friendships, showing manners
Features	Informal and casual. Use of everyday vocabulary (not jargon). May be marked by use of slang, or short, ungrammatical, or incomplete sentences Conversational and interactive May rely on body language, facial expressions, gestures and eye-contact
Markers of effective or proficient use	Ideas are organized and clearly communicated Comfort and fluidity (expression without frustration or long pauses) Mindfulness of audience (e.g. appropriately adjusting register in deference to social roles) Topic maintenance Adherence to social norms and cues

Although frequently marked by speech between peers, social language may also occur between children and adults under certain circumstances. Social interactions are said to typically occur outside the classroom; however, there are a number of common classroom activities that are at least partially social in nature. Furthermore, teachers' responses suggested that they saw overlaps, if not similarities, between social and academic language in terms of content and context of use. Teachers saw a place for social language in the school setting and, on both the survey and in interviews, teachers pointed to social language as 'survival' language allowing for basic functioning in the school setting (e.g. expressing wants and needs and resolving conflict). The overlaps and caveats expressed were consistent with criticisms of the dichotomous view of social and academic language and echo the difficulties expressed in the literature about attempting to pinpoint a definition for academic language (Bunch 2009, Lucero 2012).

To this end, although initial steps have been taken to characterize social language, additional work is needed to determine whether and to what extent social language may be defined on its own terms. Such work would be especially relevant to the field of language testing given that tests continue to include social language tasks, and work to date has found social and academic language statistically indistinguishable in standardized language tests for young language learners (Gu 2014). Additionally, as this examination targeted a relatively small and specific sample of teachers to define social language, future work may also consider the construct in other contexts and larger contexts of study.

With respect to how teachers assess social language, findings were mixed. Although the general survey results pointed to regular classroom assessment of social language, teacher interviews consistently indicated that in practice, any assessment of oral language, much less that focusing on social language, was infrequent and incidental. Although most classroom situations were viewed as at least partially social in nature, teachers pointed to increasing academic demands as limiting the time available for instruction and assessment of oral and social language. The settings that teachers identified as most related to social language (e.g. lunch, playground) were also settings in which teachers were typically removed from the students; teachers felt that their opportunities for assessment were limited. In light of limitations on their time, teachers were found to assess students' social and oral language holistically, based on general impressions and/or incidents of a regulatory nature (e.g. classroom management, conflict resolution).

These findings were surprising given that early grade language classrooms are a context in which students' oral and social development are expected to be highly emphasized (McGroarty 1984). The findings from this study were more consistent with the practices observed by Oliver, Haig and Rochecouste (2005) in a secondary classroom of native speakers of English, where oral and communicative language are incorporated into instruction and assessment on a limited basis, despite teachers' open acknowledgement of their importance, with greater emphasis placed on written language. Relatedly, it was interesting to find in this study that some teachers considered students' writing as reflective of oral language ability. The choice that some teachers made to focus on writing may be interpreted as prioritization of academic language, as per Schleppegrell's (2004) notion of academic language as literate language, but it may also be indicative of an integrated language-based approach to ESL instruction that has been shown to work well in elementary ESL classrooms (Kim 2008).

A common thread in this investigation was teachers' framing of social language as necessary for negotiating the school setting, be it defining social language as that used for 'survival' or using incidents of conflict as teachable moments and opportunity for assessment. In this way, teachers' characterizations of social language are arguably better aligned with a category of language Bailey and Heritage (2008) refer to as 'School Navigational Language' (SNL). SNL is described as 'the language needed to communicate with teachers and peers in the school setting in a very broad sense' (Bailey and Heritage 2008:15). SNL is offered as a middle ground between social and academic language. Although SNL is an alternative option to the academic/social language proficiency binary, under Bailey and Heritage's framework it is categorized as a variant of 'academic' language, given that it is used to support in-school learning. Although early-grade teachers acknowledge a grey area in the social and academic dichotomy, it is unclear the extent to which they recognize SNL as a language category or that it has academic value. In efforts to work within a social/academic dichotomy, it seems teachers have placed SNL in the social domain and have downplayed it in light of what they perceive as more pressing academic demands.

In terms of implications for policy and practice, these findings point to a crucial mismatch between assessment and instruction. Originally motivated by the presence of tasks targeting social language on high-stakes ELP tests, this study found that although social language is a construct valued by teachers, it is not emphasized in classrooms due to teachers' perceptions that it is not academically relevant or prioritized by the larger education system. Beyond implications for the validity of including these tasks, de-emphasizing social language (or SNL, as the case may be) in instruction and assessment means that students may not be receiving the necessary language support. Extensive literature points to the importance of social language in students' linguistic and general academic development. For young students, the classroom represents a major social setting, and for ELLs, it may be one of the few settings in which they may be required to use English. Contrary to common assumptions that students naturally acquire social language, it has been argued that young language learners require instruction to aid in its development (Bailey and Heritage 2008, Gu 2014, Gresham, Elliott, Vance and Cook 2011). This work suggests that, to help bridge the disconnection, there is need for both tests and teachers to acknowledge and address areas of overlap in the academic/social binary, such as SNL.

Notes

[1] www.wida.us/aboutus/mission.aspx

References

Abedi, J (2007) English language proficiency assessment and accountability under NCLB Title III: An overview, in Abedi, J (Ed) *English Language*

Proficiency Assessment in the Nation: Current Status and Future Practice, Davis: University of California, Davis School of Education, 3–12.

Albers, C A, Kenyon, D M and Boals, T J (2009) Measures for determining English language proficiency and the resulting implications for instructional provision and intervention, *Assessment for Effective Intervention* 34 (2), 74–85.

Arnold, D H, Kupersmidt, J B, Voegler-Lee, M E and Marshall, N A (2012) The association between preschool children's social functioning and their emergent academic skills, *Early Childhood Research Quarterly* 27, 376–386.

Aukerman, M (2007) A culpable CALP: Rethinking the conversational/academic language proficiency distinction in early literacy instruction, *The Reading Teacher* 60 (7), 626–636.

Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bailey, A L (Ed) (2007) *The Language Demands of School: Putting Academic English to the Test*, New Haven: Yale University Press.

Bailey, A L and Heritage, M (2008) *Formative Assessment for Literacy Grades K-6: Building Reading and Academic Language Skills Across the Curriculum*, Thousand Oaks: Corwin Press.

Bailey, A L and Huang, B H (2011) Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing* 28, 343–365.

Borsato, G N and Padilla, A M (2007) Educational assessment of English language learners, in Suzuki, L A and Ponterotto, J G (Eds) *Handbook of Multicultural Assessment* (3rd edition), San Francisco: Jossey Bass Publishers, 471–489.

Bunch, G C (2009) 'Going up there': Challenges and opportunities for language minority students during a mainstream classroom speech event, *Linguistics and Education* 20 (2), 81–108.

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

Chalhoub-Deville, M and Deville, C (2008) Nationally mandated testing for accountability: English language learners in the US, in Spolsky, B and Hult, F M (Eds) *Handbook of Educational Linguistics*, Hoboken: Blackwell, 510–522.

Cummins, J (1980) The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue, *TESOL Quarterly* 14 (2), 175–187.

Cummins, J (1991) Conversational and academic language proficiency in bilingual contexts, *AILA Review* 8, 75–89.

Cummins, J (1999) *BICS and CALP: Clarifying the Distinction*, available online: eric.ed.gov/?id=ED438551

Espinosa, L M (2005) Curriculum and assessment considerations for young children from culturally, linguistically, and economically diverse backgrounds, *Psychology in the Schools* 42 (8), 837–853.

Fang, Z, Schleppegrell, M J and Cox, B E (2006) Understanding the language demands of schooling: Nouns in academic registers, *Journal of Literacy Research* 38 (3), 247–273.

Gee, J P (2011) *Social Linguistics and Literacies*, New York: Routledge.

Gresham, F M, Elliott, S N, Vance, M J and Cook, C R (2011) Comparability of the social skills rating system to the social skills improvement system: Content and psychometric comparisons across elementary and secondary age levels, *School Psychology Quarterly* 26 (1), 27–44.

Gu, L (2014) Language ability of young English language learners: Definition, configuration and implications, *Language Testing* 31 (1), 89–109.

Halle, T, Hair, E, Wandner, L, McNamara, M and Chien, N (2012) Predictors and outcomes of early versus later English language proficiency among English language learners, *Early Childhood Research Quarterly* 27, 1–20.

Haneda, M (2014) Why should we care about academic language? *Linguistics and Education* 26, 88–91.

Hawkins, M R (2005) ESL in elementary education, in Hinkel, E (Ed) *Handbook of Research in Second Language Teaching and Learning*, New York: Routledge, 25–43.

Hsieh, H and Shannon, S E (2005) Three approaches to qualitative content analysis, *Qualitative Health Research* 15 (9), 1,277–1,288.

Jeynes, W H (2006) Standardized tests and Froebel's original kindergarten model, *Teacher's College Record* 108 (10), 1,937–1,959.

Kim, Y (2008) The effects of integrated language-based instruction in elementary ESL learning, *The Modern Language Journal* 92 (3), 431–451.

Leung, C (2014) Researching language and communication in schooling, *Linguistics and Education* 26, 136–144.

Lucero, A (2012) Demands and opportunities: Analyzing academic language in a first grade dual language program, *Linguistics and Education* 23, 277–288.

Maxwell, L A (2013) Common core ratchets up language demands for language-learners, *Education Week* 33 (10), 14–16.

McClelland, M M and Morrison, F J (2003) The emergence of learning-related social skills in preschool children, *Early Childhood Research Quarterly* 18, 206–224.

McGroarty, M (1984) Some meanings of communicative competence for second language students, *TESOL Quarterly* 18 (2), 257–272.

McKay, P (2006) *Assessing Young Language Learners*, New York: Cambridge University Press.

Oliver, R, Haig, Y and Rochecouste, J (2005) Communicative competence in oral language assessment, *Language and Education* 19 (3), 212–222.

Pinter, A (2006) *Teaching Young Language Learners*, Oxford: Oxford University Press.

Raver, C C and Zigler, E F (1997) Social competence: An untapped dimension in evaluating Head Start's success, *Early Childhood Research Quarterly* 12, 363–385.

Russell, R L and Grizzle, K L (2008) Assessing child and adolescent pragmatic language competencies: Toward evidence-based assessments, *Clinical Child and Family Psychology Review* 11, 59–73.

Schappe, J F (2005) Early childhood assessment: A correlational study of the relationship among student performance, student feelings, and teacher perceptions, *Early Childhood Education Journal* 33 (3), 187–193.

Schleppegrell, M J (2001) Linguistic features of the language of schooling, *Linguistics and Education* 12 (4), 431–459.

Schleppegrell, M J (2004) *The Language of Schooling: A Functional Linguistics Perspective*, New York: Routledge.

Shiel, G, Cregan, Á, McGough, A and Archer, P (2012) *Oral Language in Early Childhood and Primary Education (3–8 Years)*, Research Report number 14, Dublin: National Council for Curriculum and Assessment.

Section 3
Language assessment concerns in local contexts

9 The implications of test taker perceptions for test validity in community college settings

Tasha Darbes

Pace University, New York, US

Motivation for research

Since Messick (1989) expanded the notion of validity to include social consequences, the field of language assessment has examined assessment as a social practice, which includes the influence of context, the decision-making process that helps us understand how tests are used, and the interpretations and experiences of the test takers in this process (Broadfoot and Black 2004, McNamara 2006). Once testing is posited to be a social process, the analysis of tests and their validity expands to include multiple actors or stakeholders (Moss, Girard and Haniford 2006), including the test takers themselves.

The experiences and perceptions of test takers have been identified as key, yet they are under-researched phenomena in educational assessment (Deil-Amen and Tevis 2010, McInerney, Brown and Liem 2009). Test takers are not passive recipients of the scores they receive; they interpret and use scores consciously in making important decisions. In addition, test scores may create unconscious negative and positive affect for test takers (Shohamy 2007). As such, the perceptions of test takers can play an important role in how tests are used, how testing policies are enacted in practice, and even how they are validated (Fox and Cheng 2008).

One way that the perceptions and experiences of test takers can threaten arguments related to validity is through their impact on construct irrelevant variance (CIV). CIV is concerned with the presence of factors that can affect test performance and yet have nothing to do with the construct being measured (Ferrier, Lovett and Jordan 2011, Haladyna and Downing 2004). For example, if a student who does not know how to type is given a timed essay test on a computer, the score may not be an accurate reflection of the construct of writing ability because the test is inadvertently also measuring typing ability, which is not part of the underlying construct. The construction of validation arguments for particular tests calls for the inclusion of test takers in the process, as threats due to construct-irrelevant variation are otherwise difficult to detect (Haladyna and Downing 2004).

In addition to the type of validity threats that result from CIV, Messick also included another type of validity – consequential validity, which refers to the negative and positive social consequences of testing. He takes the position that negative social consequences alone are not sufficient to invalidate a test; nevertheless, language testers need to ensure that such consequences are not 'attributable to any source of test invalidity' (Messick 1989:11). However, critical language testing theorists, such as McNamara (2006) and Shohamy (2001), have extended that supposition and view tests as embedded in and reproducing social practices as tools of power. As such, tests are not neutral, value-free projections of psychometric principles, but they are part of ideological struggles and produce social consequences for those who pass or fail. This critical view of language testing severs the connection between a 'psychometrically good' test and a 'socially good' test by arguing that tests that function as social gatekeepers, as social sorters, and as tests that lock out and misrepresent certain groups of test takers are no longer valid tests (McNamara 2006).

These concerns over the social impact of tests are especially pertinent to community college settings in the USA – nonresidential, junior colleges that offer courses to people living in the area. These colleges are experiencing a rise in the number of culturally and linguistically diverse students (Szelenyi and Chang 2002), including immigrants and the children of immigrants (Erisman and Looney 2007). Community colleges are important spaces for educational and economic advancement of diverse students (Suárez-Orozco 2004, Teranishi, Suárez-Orozco and Suárez-Orozco 2011).

Although all students are confronted with educational 'potholes' during their transition to college, the testing and placement processes of community colleges have been identified as being particularly problematic, while at the same time they are critical for the future engagement of diverse multilingual students (Bunch and Panayotova 2008, Perry, Bahr, Rosin and Woodward 2010). However, there is little research available that examines the assessment and placement processes from the perspectives of participants, in particular, the responses of immigrant-origin students who have complex perceptions of their own abilities and their ethnolinguistic identities (Bunch, Endris, Panayotova, Romero and Llosa 2011), thereby making this population an important addition to research on language testing (Fox and Cheng 2008).

Literature review

Messick (1989) pushed the field of language assessment to consider the importance of social context when evaluating test validity, not only by challenging the notion of a value-free validation and testing process, but also of the need to take the social consequences of the test into consideration during the design. McNamara (2001:336) has built on Messick's work, stating that this consideration 'requires us to engage explicitly with the fundamentally social character of assessment at every point'. This work has led the field of assessment and test validation to consider both the social context of testing, as well as the viewpoints of multiple stakeholders (Moss 1996).

The area of study that investigates test taker and stakeholder viewpoints is CIV. Haladyna and Downing (2004) presented four types of systematic errors associated with CIV. The four types are 1) uniformity and types of test preparation; 2) test development, administration and scoring (such as, item quality, conditions, and rating practices); 3) students (individual characteristics, such as verbal ability and anxiety); and 4) cheating. They argue for an increase in research that systematically investigates sources of construct irrelevant variation.

In the field of language assessment, research on test taker perceptions has focused primarily on testing experiences, which include the behavior and reactions of test takers during the process of test taking (Elder, Iwashita and McNamara 2002, Lewkowicz 2000) and the strategies they use to answer test items or complete tasks (Cohen and Upton 2006, Rupp, Ferne and Choi 2006). In relation to the types of CIV-related errors, studies have examined uniformity in test preparation (Sasaki 2000), as well as affective factors and students' perceptions of themselves as language learners (Huhta, Kalaja, Pitkänen-Huhta 2006, Shohamy, Donitsa-Schmidt and Ferman 1996, Xiao and Carless 2013). However, the use of test taker perceptions to validate language tests has been less common in language assessment research (Cheng and DeLuca 2011, Hamp-Lyons 2000).

Bachman (2005) discussed the lack of connection between validity and test use, and Cheng and DeLuca (2011) investigated potential links between test taker perceptions of test validity and use and test validation itself. Their analysis of 59 test takers on English writing assessments uncovered eight themes related to their perceptions of test use and validity. These themes included factors related to the experience of testing, how the test items were scored and used, the purpose of the test, and psychological factors. They argue that test takers' perceptions contribute to validation arguments, especially when these perceptions are conceived of multi-dimensionally, and that test designers and classroom teachers can use these perceptions to ensure the accuracy of how test scores are used and interpreted. However, their work on high-stakes English testing in China may not be generalizable to US community college settings that serve large numbers of immigrant-origin students.

Test taker perceptions

One way to investigate the perceptions of test takers is to use the concept of causal thoughts or the underlying structures or ways that students conceptualize the reasons for outcomes. The question of how students interpret their success or failure on exams (i.e. their causal thoughts) underlies a psychologically based line of research first outlined by Heider (1958). He identified four prominent causes – ability, effort, luck, and the difficulty of the task. Other research has found that causal interpretation may be motivated by the need to sustain self-image (Anderson 1991, Bempechat and Mirny 2005), which can be related to conceptions of identity. This research has focused primarily on grades in courses or reactions to specific content exams or tasks rather than on standardized exams (Forsythe, Story, Kelley and McMillan 2009). In addition, most studies occurred in the context of 4-year universities without any mention of cultural or linguistic diversity.

Research on community college students' perceptions and experiences of testing has revealed other issues with writing tests. Salas's (2008) ethnographic research of Latino students in community colleges revealed that problems with standardized assessment, course placement and complicated and intricate institutional requirements seemed to conspire to frustrate students, making students particularly vulnerable to test use error, which poses a threat to consequential validity (i.e. that the social consequences of using a particular test would be negative) and results in student disengagement. Other researchers have documented further connections between assessments and disengagement, such as Venezia, Reeves Bracco and Nodine (2010), who found that students did not feel prepared or well informed about assessment and placement, which they perceived to be a 'one shot deal' instead of an extended process. Mott-Smith (2009) found that students became frustrated when test results did not align with their performance in courses, and this frustration led some of the students in her study to drop out of their courses.

The intersection of perceptions and beliefs about language identity and proficiency is especially salient for multilingual and immigrant-origin students who are in the process of complex identity formation (Leung, Harris and Rampton 1997, Rampton 1990, Suárez-Orozco 2004). Identity, especially as it relates to how students position themselves in relation to the languages they speak, plays an important role in the perceptions of community college students and how they react to placement into English as a Second Language (ESL) or remedial English (Bunch and Panayotova 2008, Marshall 2009). Fox and Cheng (2008) found that first language users and second language learners perceive tests differently, suggesting that testing can be an important space for identity work.

Language testing in community colleges

Assessment and placement practices at community colleges play key roles in shaping the academic pathways of students (Hughes and Scott-Clayton 2011). However, studies have found a variety of threats to validity when tests are used with diverse populations, such as in community college contexts (Bunch and Panayotova 2008, Curry 2004, García and Menken 2006, Hodara 2013, Llosa and Bunch 2011). For example, the tests may not have been designed for the purpose of dividing students into the categories of ESL and native English or proficient bilingual speakers, which is one of their primary functions in community college settings (Bunch and Panayotova 2008, di Gennaro 2008). Additionally, many students acquiring English have previous experiences with English as a second or foreign language instruction that often focuses on discrete skills and grammar instruction that may not prepare students for the types of academic writing and reading tasks found on high-stakes writing tests (Curry 2004). Lastly, there are threats to scoring validity, as immigrant-origin students may have acquired non-standard English forms from their peers or other community members, and untrained raters who are not familiar with these forms may judge them as non-native (Valdés 1992).

Research has also found that the assessments used by community colleges suffer from particularly low predictive validity (i.e. the extent to which a score on a test predicts a score or performance on some other criterion measured). For example, Scott-Clayton (2012) found that the correlation of English placement exam scores with a grade B or higher in their college English class was only 0.147 in comparison to math scores, which had a correlation of 0.3. Both Scott-Clayton (2012) and Belfield and Crosta (2012) found that using English test scores was likely to result in a severe error rate (i.e. a situation in which students may be placed in classes that are either over or under their predicted abilities). In fact, there was no overall reduction in severe error rate when Scott-Clayton compared using English test scores with placing all students in college-level classes, due to the high percentage of under-placements generated by using exam scores. These researchers recommended using a combination of high school grades and test results to counteract the lack of predictive validity of the English exams and reduce severe error rate without compromising college success rates. Tests are only as valid as the chain of decisions that inform their use (Llosa and Bunch 2011).

Overall, this body of work underscores the importance of researching student perspectives of their abilities and identities. Incorporating how the acquisition of academic language and experiences of testing intersect with institutional practices will lead to more nuanced understandings of the immigrant community college population, and can inform test development, policy and practices to improve educational outcomes.

Research questions

Using a critical language-testing framework (McNanamara 2006, Shohamy 1998), I examine community college students' causal thoughts (i.e. perceptions) and experiences of community college English tests. Students' perceptions of their linguistic proficiency, the development of their identity, the institutional tests, and their subsequent experiences should be looked at collectively as part of an overall dynamic process. To this end, the current study addressed the following research questions:

1. How do community college immigrant-origin test takers perceive the validity of English tests?
2. How do community college immigrant origin test takers experience testing?
3. What are the causal thoughts (i.e. the perceptions) that they use to explain their experiences?
4. What are possible implications of their causal thoughts for test validity?
5. What are the possible psychological or social impacts of causal thoughts for their academic trajectories?

This chapter examines the experiences of linguistically and culturally diverse test takers as a social process. This approach incorporates qualitative inquiry into the process to provide a perspective that is different from the generalizable, quantitative studies that currently dominate the field of language assessment (Cumming 2004:10). It is, therefore, of particular importance to see the processes of community college testing and placement as complex and dynamic and to examine if and how immigrant-origin student populations experience these processes.

Data collection procedures

The data reported on in this chapter come from a larger umbrella study, Research on Immigrants in Community Colleges (RICC), which was a multiphase, mixed methods study in which data were collected from three community college settings. Over the course of two years, a team of researchers collected survey data, classroom observations, artifacts, and interviews of students and staff. This study draws on data from the qualitative interviews of community college students, as well as analyses of community college artifacts related to testing.

Research context

Data were collected in three different community colleges with varying densities of immigrant-origin students who were from diverse countries, had different educational emphases, and studied in distinctive physical settings all referred to with pseudonyms – two urban sites, Taino and Domino, and one suburban site Oakmont. All three sites served a mix of native-born bilingual students, foreign-born immigrant students, and foreign-born student visa-holders (the latter group of students were excluded from the study). All three campuses had both ESL and remedial programs. Taino had the largest number of recently arrived immigrant students, and Domino served the largest number of native-born bilingual students. Oakmont, though predominantly white and native born, had most recently experienced increasing numbers of culturally and linguistically diverse students.

Writing tests in community college settings

The three campuses included in the study differed in their testing and placement practices in English writing. These different practices provided useful points of comparison among the institutions. At the two urban campuses, Taino and Domino, students could be exempted from taking placement exams by surpassing cut-off scores on standardized college readiness exams. All other students were required to take the Computer-Adaptive Placement, Assessment and Support System (COMPASS) reading exam, which was designed to assess native speakers, and a locally developed writing test. Cut-off scores and other testing procedures were set centrally. At these campuses, writing exams were scored off campus by trained raters who used a rubric developed by American College Testing (ACT). The earlier exam used a 6-point, holistically scored rubric; the latter exam assigned scores of 1–6 along various dimensions, such as development or control of language, which were then combined for a total score. If readers detected that the writers exhibited 'ESL features' in their writing, the readers were instructed to assign an 'E' to the essay, which marked the test taker as a potential ESL student.

Oakmont, the suburban campus, had a very different procedure. This campus employed practices more similar to those of the State of California (Bunch et al 2011), as students largely self-selected into either a non-credit-bearing ESL program or the regular college matriculation process. This college used a computer-adaptive reading exam called Accuplacer, which is a suite of tests that can be used to assess reading, writing, math, and computer skills, and locally developed writing exams, which differ based on whether students were being placed into non-credit ESL courses, credit-bearing ESL, or remedial English. The rubrics for the latter exams weighed local and global grammatical and lexical errors more explicitly in the rubric, and maintained a practice of referring students with such errors to ESL.

Participants

Ten percent (N = 60) of the students surveyed were interviewed. Names were drawn randomly from a list of willing participants representing each gender and ethnic group. Table 9.1 summarizes the demographics of participants by campus, gender, race or ethnicity, and immigrant generation. Also listed are the primary home languages spoken by bilingual participants. The home languages are Spanish, Albanian, Arabic, Bangla, Cambodian, Chinese, French, Ga, Greek, Haitian Creole, Hebrew, Japanese, Portuguese, Russian, Tagalog and Twi.

Table 9.1 Qualitative interview demographics

	N	Male	Female	Asian	Black	Latino	White	First gen.	Second gen.
Taino	21	11	10	2	6	12	1	12	9
Domino	19	12	7	5	4	10	0	9	10
Oakmont	20	8	12	1	6	10	3	12	8
Total	60	31	29	8	16	32	4	33	27

Student interviews and survey data also provided information about the perceived bilingual abilities of the participants. In this sample, 15 students identified themselves as monolingual English speakers. Of the remaining 45 bilingual students, 23 identified themselves as being in the process of learning English or indicated they had second language difficulties at the time of their entrance into community college.

Semi-structured student interviews

I selected a specific subset of questions from the comprehensive RICC protocol. The protocol was semi-structured and covered a wide range of topics about student experiences in community colleges. The subset of questions analyzed for this paper included the following: testing, placement, misplacement, experiences in developmental and/or ESL courses, and challenges faced, and support received in academic courses (see the Appendix).

Data analysis

The qualitative research design of the study was informed by grounded theory (Charmaz 2003, 2006), which allows concepts to emerge from the participants. As such, these concepts might be new to the field and would not reify or reproduce existing frameworks that may not apply to the experiences under study. Codes were developed using an open-coding process, which used ‘active codes’ (gerund phrases pulled from the data), ‘in vivo’ codes (respondents’ exact words), and codes derived from the sensitizing concepts related to test types. Once all interviews were coded using an open-coding process, axial coding (i.e. comparing codes, concepts, and categories to see the data in new ways, refining categories and relationships, and looking for negative or falsifying cases) was employed. The last stage of coding was selective coding, which involved selecting and collapsing axial codes to identify themes that cut across the data. All coding was aided by the use of the qualitative software package MAXQDA.

In qualitative data analyses, Lincoln and Guba (1985) recommend attending to issues of *credibility* (i.e. the results are believable), *dependability* (i.e. the findings are consistent and could be repeated), and *transferability* (i.e. the results could transfer to other contexts). I attended to credibility by triangulating data, using difference data sources, including quantitative student data, school administrative data and archives, observation field notes, and interviews with faculty and administrators from the campuses under study. I attended to dependability by employing two other coders to review codes and coding at various stages of the analysis to guard against researcher bias. Lastly, I attended to transferability by providing a chain of evidence that others can follow and replicate in future studies (Merriam 2009).

Results

Student perceptions of community college assessments can best be understood within the context of the issues with the tests that are used to construct the proficiencies and readiness of the test takers; conversely, the validity of these tests is further informed by the experiences of test takers whose outcomes are determined by these tests. Analyses revealed distinctive patterns related to perceived language proficiency and these patterns will be discussed in the following section.

Student perceptions of test validity

When asked if initial placement exams were a good measure of what they actually knew, students were in agreement that the tests were fair and a good indicator of their abilities. It is interesting to note that students did not question the validity of the tests they were given or the value of the knowledge that was being tested. Oscar, a Taino student who arrived at age 17 from the Dominican Republic said:

> I had a professor, an American government professor, he used to argue that the SATs and the ACTs were designed to keep black students from going to Ivy League schools, and he would say that they made it difficult so people from low budget school districts could not go to good schools. But I think this test was not like that, because I think what they did is that they measure what you know and then they try to teach you exactly what you don't know.

The perceptions of test fairness were equally distributed across all campuses, which means that the type of test did not really affect students' perceptions of the test because the tests were different on each campus. In addition, there were only two instances of students complaining about the test itself in the 60 interviews analyzed.

Another theme that emerged from these data was what I call 'testing tautology', namely that the test is a good measure because the score reflected students' abilities, and students know what their abilities are because the test showed them. Most students did not use other sources of information (such as grades or teacher comments) when they evaluated their own abilities but often relied heavily on test results as the way of knowing what their abilities were. Evelyn, a first generation student from Ecuador, asserted that the test was fair because it showed her what her abilities were. When asked why she believed the test was a good measure, she responded: 'After I took the test, it told me how good I did and how bad I did and in both of the tests; it said I did really good except the English one'. Sophal, who immigrated from Cambodia at age 17, echoed this by stating: 'It meets the level in order for you to pass, like they make it not . . . I am not going to say easy because I didn't find it easy, but if you pass the test, that proves that you really are good enough to go to the upper levels, so yeah, it is fair'. It is important to note that causal thoughts that shift the locus of causality away from one's individual responsibility to an external cause did not occur or occurred very rarely in these data. For the majority of students in the sample, the causal thoughts were directed squarely at themselves and their abilities. There was no mention of external factors, such as on the quality and appropriateness of previous instruction or the fairness of the test.

The responses revealed that these students accepted the validity of the test and the underlying construct of valuable knowledge that the test represented. Student perceptions of test fairness should be contextualized in the findings. In other words, because these exams have low predictive validity, use of exam scores alone can result in high rates of over- and under-placement in community college settings. The unquestioning acceptance of the test has important consequences, as will be discussed in more detail.

Experiences of testing

The implicit trust in test results meant that students looked to other reasons besides test validity or fairness when asked to explain the reasons for their success or failure. The following sections will examine how immigrant-origin students in the qualitative sample perceived the reasons for the results of the assessments, or why they received passing or failing scores, especially in relation to possible sources of construct irrelevant variance (CIV). From a psychological standpoint, an analysis of causal thoughts or students' perceptions helps researchers understand student-level variables, such as motivation and preparation, which relate to CIV.

Causal thoughts for success were not commonly found in the narratives developed by this interview protocol. Most of the students who passed the exam on the first try did not find this experience to be very salient, and did not say much about it, though some noted the easiness of the exam or the fact that they felt prepared from high school. Those that passed after remediation often attributed passing to personal effort or improvement from remedial classwork and instructor help. Of the 60 students of immigrant origin in the qualitative sample, almost half (29) reported passing the writing exam upon entry, and for these students testing and placement was not a salient experience.

For students who did not pass, testing was a salient experience, one that prompted more discussion about the reasons for their failure. An analysis of these causal thoughts revealed different patterns for students who self-identified as being in the process of learning English during community college (hereafter referred to as 'English learners'), and those who identified as either native speakers and/or proficient bilinguals.

The theme of 'I forgot' was very prevalent in the narratives of English proficient students. These students ascribed their failure to taking the test after a break from high school, which meant that the information was 'not fresh' in their minds. For these students, failure was due not to a lack of abilities or competence, but to a delay period that caused them to forget what they had learned. It is interesting to note that the delay period could mean anything from a few weeks to years. 'But probably because I took the tests in June, I was already done with my core classes so I didn't really have that knowledge fresh in my mind,' said Marisol, a second-generation student from Oakmont who took her test before graduating high school. For these students, the way to pass the test was to 'relearn' the material and master the skills they had forgotten, which were primarily expressed as skills-based rules of punctuation and grammar.

English proficient students also reported confusion as to the purpose and consequences of the writing test, although this was not common.. Two students in the sample ascribed their performance to not being told that the test would be used to place them into remedial, and so they did not take it seriously. Luis from Taino said:

> Oh I really didn't did good because it was my first day of college so I thought it was just an exam, whatever, to see whatever. But I didn't really pay mind to it and then when they told me I had to take remedial class because I didn't have good grades on the tests. Then I felt, you know, like, oh really?

This confusion about the high-stakes nature of this test affected test motivation, a threat to the validity of their scores.

Students across sub-groups also alluded to psychological or mood factors that could affect their performance. Some attributed their failure to lack of focus, which could be caused by personal problems at the time of the exam, distracting test conditions, or physical problems. However, approximately half of the students who were in the process of learning English more often reported high test anxiety due to the high-stakes consequences of the test and an over-awareness of the consequences of failure. Nelli, who immigrated to the USA from Peru at the age of 14, stated the following:

> Then when I took it, the first exam, I said, if they would give me, they gave me to choose, then, what do I do now? Then I was nervous and I didn't write much, and because of that I did badly on that.

In this sample there was a marked difference in the causal thoughts of English learners, who stated explicitly that they had difficulties with the test due to speaking English as a second language. Eleni, a Taino student, said: 'Oh, I didn't pass it. As I told you, I just came from Cyprus and even though in the high school I was taking English classes, I don't know why, I had that hard time to learn English.' Another causal thought that English learners identified was due to the format of the test. For example, they attributed failure to not following the form they believed was necessary to pass the writing test, reflecting a belief that they should pass it once they master an essay form. Heydi explained her failure during a retake of the test as follows: 'When you first start this [the essay], you have to start with some steps, and I didn't follow them. So that's why I think I failed them. I did it my way'. English learners also experienced more problems with the time limit of the writing test than other test takers.

In addition, some test takers reported experiencing utter confusion when confronted by the writing test. They were completely unfamiliar with this type of test and could not respond to the prompts. As Felix, a recently arrived Dominican student from Domino stated:

> Maybe I didn't even know what an essay was at the time that I took the test. When I read the instructions, "Write an essay", I was, 'What do they want me to do? What is an essay?' So I feel I had failed that one, [laughs] because I didn't know what I had to do.

Because test takers who were English learners were unfamiliar with the test format and the expectations for writing in the USA college system, they often saw their ESL classes as providing the test preparation that they needed. In fact, when they did pass the test, they often attributed this to the amount of test preparation they were given by their ESL instructors. These students also overwhelmingly perceived these tests to be difficult to pass, in contrast with other immigrant-origin students in the sample.

Due to the perceived difficulty and importance of the test, as well as the fact that multiple retakes were often required, the test became a central feature that structured their initial entrance into college and dominated the experiences of English learners. It is worth noting that none of the students in the sample who identified as English learners passed both the reading and writing tests at their initial college entry. These tests were designed, through the use of previously unpracticed writing prompts, to assess a student's ability to do college-level work, especially their ability to write well enough to pass college-level English courses. However, students often perceived the ability to pass these tests to be the result of targeted preparation, and ascribed their failure to not studying the pre-determined format enough rather than to the time it takes to sufficiently develop English written and academic skills. Noemi, who had failed the writing test multiple times over the course of three years at Taino, stated: 'So when I failed the test the last semester, I think I said "Oh my God" I have to keep write [sic], practice a lot, a lot, a lot of times and then I can pass'.

These findings show that immigrant-origin students may be constructing causal thoughts in different ways due to their positionality and experiences. Reactions to test scores and causal thoughts are the first step in the assessment and placement process, and will inform the students' trajectory after this initial entry into the system.

Discussion

The analysis of students' perceptions of assessment and placement practices revealed a number of important findings. First, students perceived assessments to be valid and did not question the results, which meant that students did not often articulate concerns about the exam or about misplacements in their English courses. Overall, students placed the blame for poor performance on themselves rather than on the college's assessment and placement practices. Only two students in the sample reported actively advocating for another placement. Although student perceptions seemed to support the validity of the test, there is the question of testing tautology, which has implications for policy and practice. Because students expressed high levels of trust in the test results, community colleges should not rely solely on the students' perceptions to initiate challenges to placements in English courses. Colleges and programs also need to be proactive in identifying misplaced students and establish clear procedures for allowing students to make changes in their initial English course placements. A careful balance needs to be achieved. Even in cases when there are problems with the test itself – the processes, how the constructs are being measured, or the constructs themselves – students may not question placements.

An analysis of test taker responses also revealed a number of possible threats related to construct-irrelevant variance. Students had various experiences during the testing process, including issues with distraction, text anxiety, and lack of understanding of test formats or the purposes of the test. These issues have been documented in other studies; however, in contrast to research conducted in Chinese contexts (Cheng and DeLuca 2011, Xiao and Carless 2013) this study found little reference to explicit test preparation as a cause for performance. 'Test prep' explanations for test takers in this study were only mentioned in reference to the need for remedial coursework.

The findings show that English proficient immigrant-origin students experienced testing differently from English learners. In particular, English learners in this study experienced high levels of test anxiety, coupled with other factors that made passing the test of high importance, such as losing financial aid or asserting their identity and belonging. English learners also reported more confusion about the test format. In addition, they had less access to test preparation and experienced more difficulty with timed writing tasks. Findings suggest that students learning English may be experiencing more issues related to CIV.

This study adds to the literature on the testing of bilingual populations (Valdés and Figueroa 1994) by showing that non-linguistic factors also contribute to test validation and should be investigated systematically (Fox and Cheng 2008). Testing and placement practices intersect in complex ways with the perceptions and identities of multilingual students. Because it is difficult to tease apart whether the lower score of a student learning English is due to their lower proficiency level or CIV, future studies should combine qualitative and quantitative data to better understand how CIV can affect test validity with this population of learners. It is important to understand how the exam scores result in such high severe error rates, and tease apart issues of construct validity, construct-irrelevant variance, and other threats to validity. Such studies could investigate whether CIV has a disproportional impact on sub-populations such as students learning English, which provides evidence for the consequential validity of these tests and testing procedures. The results of this study documented how students experienced testing, repeated test failure and suggested a misalignment of policies to the realities of second language development. Assessments and placement practices are agents in forming student perceptions, feelings of belonging, and, ultimately, identity. The experience of attending a community college should be about opening gates and supporting aspirations, and tests should not be a source of adverse social consequences. Research that examines how students experience testing policies and practices can be used to create open, equitable, and valid testing practices.

Acknowledgements

The data for this research came from the Research on Immigrants in Community Colleges project, which was funded by the W T Grant Foundation and the Ford Foundation. The principal investigators for this project were Carola Súarez-Orozco, Marcelo Súarez-Orozco, Robert Teranishi and Margery Martin. This chapter, which is based on my dissertation, was also made possible by a grant from The International Research Foundation for English Language Education (TIRF).

References

Anderson, C A (1991) How people think about causes: Examination of the typical phenomenal organization of attributions for success and failure, *Social Cognition* 9 (4), 295–329.

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

Belfield, C and Crosta, P M (2012) *Predicting Success in College: The Importance of Placement Tests and High School Transcripts*, CCRC Working Paper number 42, Assessment of Evidence Series, New York: Teachers College, Columbia University, Community College Research Center.

Bempechat, J and Mirny, A (2005) Contemporary theories of achievement motivation, in Farenga, S J and Ness, D (Eds) *Encyclopedia of Education and Human Development*, Armonk: ME Sharpe, 433–443.

Broadfoot, P and Black, P (2004) Redefining assessment? The first ten years of Assessment in Education, *Assessment in Education* 11, 7–27.

Bunch, G C and Panayotova, D (2008) Latinos, language minority students, and the construction of ESL: Language testing and placement from high school to community college, *Journal of Hispanic Higher Education* 7 (1), 6–30.

Bunch, G C, Endris, A, Panayotova, D, Romero, M and Llosa, L (2011) *Mapping the terrain: Language testing and placement for US-educated language minority students in California's community colleges*, available online: www.escholarship.org /uc/item/31m3q6tb

Charmaz, K (2003) Grounded theory: Objectivist and constructivist methods, in Denzin, N K and Lincoln, Y S (Eds) *Strategies of Qualitative Inquiry*, Thousand Oaks: Sage Publications, 219–291.

Charmaz, K (2006) *Constructing Grounded Theory*, Thousand Oaks: Sage Publications.

Cheng, L and DeLuca, C (2011) Voices from test-takers: Further evidence for language assessment validation and use, *Educational Assessment* 16 (2), 104–122.

Cohen, A D and Upton, T A (2006) *Strategies in Responding to New TOEFL Reading Tasks*, Princeton: Educational Testing Service.

Cumming, A (2004) Broadening, deepening, and consolidating, *Language Assessment Quarterly* 1 (1), 5–18.

Curry, M J (2004) UCLA Community college review: Academic literacy for English language learners, *Community College Review* 32 (2), 51–68.

Deil-Amen, R and Tevis, T (2010) Circumscribed agency: The relevance of standardized college entrance exams for low SES high school students, *Review of Higher Education* 33 (2), 141–175.

di Gennaro, K (2008) Assessment of Generation 1.5 learners for placement into college writing courses, *Journal of Basic Writing* 27, 61–79.

Elder, C, Iwashita, N and McNamara, T (2002) Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing* 19, 347–368.

Erisman, W and Looney, S (2007) *Opening the Door to the American Dream: Increasing Higher Education Access and Success for Immigrants*, Washington, DC: Institute for Higher Education Policy.

Ferrier, D E, Lovett, B J and Jordan, A H (2011) Construct-irrelevant variance in achievement test scores: A social cognitive perspective, in Madson, L E (Ed) *Achievement Tests: Types, Interpretations, and Uses*, New York: Nova, 89–108.

Forsythe, D, Story, P, Kelley, K and McMillan, J (2009) What causes failure and success? Students' perceptions of their academic outcomes, *Social Psychology and Education* 12 (2), 157–174.

Fox, J and Cheng, L (2008) Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test takers, *Assessment in Education: Principles, Policy and Practice* 14 (1), 9–26.

García, O and Menken, K (2006) The English of Latinos from a plurilingual, transcultural angle: Implications for assessment and schools, in Nero, S (Ed) *Dialects, Englishes, Creole and Education*, Mahwah: Lawrence Erlbaum, 167–183.

Haladyna, T M and Downing, S M (2004) Construct-irrelevant variance in high-stakes testing, *Educational Measurement: Issues and Practice* 23 (1), 17–27.

Hamp-Lyons, L (2000) Social, professional, and individual responsibility in language testing, *System* 28, 578–591.

Heider, F (1958) *The Psychology of Interpersonal Relations*, New York: Wiley.

Hodara, M (2013) *Heterogeneous effects of English as a Second Language compared to developmental English coursework*, paper presented at AERA 2013, San Francisco.

Hughes, K L and Scott-Clayton, J (2011) *Assessing Developmental Education Assessment in Community Colleges*, CCRC Working Paper number 19, Assessment of Evidence Series, New York: Teachers College, Columbia University, Community College Research Center.

Huhta, A, Kalaja, P and Pitkänen-Huhta, A (2006) Discursive construction of a high-stakes test: The many faces of a test-taker, *Language Testing* 23 (3), 326–350.

Leung, C, Harris, R and Rampton, B (1997) The idealized native speaker, reified ethnicities, and classroom realities, *TESOL Quarterly* 31 (3), 543–560.

Lewkowicz, D J (2000) The development of intersensory temporal perception: An epigenetic systems/limitations view, *Psychological Bulletin* 126, 281–308.

Lincoln, Y and Guba, E (1985) *Naturalistic Inquiry*, Thousand Oaks: Sage.

Llosa, L and Bunch, G (2011) *What's in a Test? ESL and English Placement Tests in California's Community Colleges and Implications for US-educated Language Minority Students*, available online: knowledgecenter.completionbydesign.org/sites/default/files/95%20Llosa%202011.pdf

Marshall, S (2009) Re-becoming ESL: Multilingual university students and a deficit identity, *Language and Education* 24 (1), 41–56.

McInerney, D M, Brown, G T and Liem, A D (2009) *Student Perspectives on Assessment: What Students Can Tell Us About Assessment for Learning*, Charlotte: Information Age Publishing.

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

McNamara, T (2006) *Language Testing: The Social Dimension*, Malden: Blackwell Publishing.

Merriam, S B (2009) *Qualitative Research: A Guide to Design and Implementation*, San Francisco: Jossey-Bass.

Messick, S (1989) Meaning and values in test validation: The science and ethics of assessment, *Educational Researcher* 18 (2), 5–11.

Moss, P A (1996) Enlarging the dialogue in educational measurement: Voices from interpretive research traditions, *Educational Researcher* 25 (1), 20–28.

Moss, P A, Girard, B J and Haniford, L C (2006) Validity in educational measurement, *Review of Research in Education* 30, 109–162.

Mott-Smith, J A (2009) Responding to high-stakes writing assessment: A case study of five generation 1.5 learners, in Roberge, M, Siegal, M and Harklau, L (Eds) *Generation 1.5 in College Composition*, New York: Routledge, 120–134.

Perry, M, Bahr, P R, Rosin, M and Woodward, K M (2010) *Course-taking patterns, policies, and practices in developmental education in the California community colleges*, available online: edsource.org/wp-content/publications/FULL-CC-DevelopmentalCoursetaking.pdf

Rampton, B (1990) Displacing the 'native speaker': Expertise, affiliation and inheritance, *ELT Journal* 44 (2), 97–101.

Rupp, A A, Ferne, T and Choi, H (2006) How assessing reading comprehension shapes the construct: A cognitive processing perspective, *Language Testing* 23, 441–474.

Salas, S (2008) Roberta; or, the ambiguities: Tough love and high stakes assessment in a two-year college in North Georgia, *Journal of Basic Writing* 27 (2), 5–28.

Sasaki, M (2000) Effects of cultural schemata on students' test-taking for cloze tests: A multiple data source approach, *Language Testing* 17, 8–114.

Scott-Clayton, J (2012) *Do High-stakes Placement Exams Predict College Success?* CCRC Working Paper number 41, Assessment of Evidence Series, New York: Teachers College, Columbia University, Community College Research Center.

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

Shohamy, E (2001) *The Power of Tests: A Critical Perspective on the Uses of Language Tests*, London: Pearson.

Shohamy, E (2007) The power of language, the power of English language, and the role of ELT, in Cummins, J and Davison, C (Eds) *International Handbook of English Language Teaching Volume 15*, New York: Springer, 521–531.

Shohamy, E, Donitsa-Schmidt, S and Ferman, I (1996) Test impact revisited: Washback over time, *Language Testing* 12 (3), 298–317.

Suárez-Orozco, C (2004) Formulating identity in a globalized world, in Suárez-Orozco, M and Qin-Hilliard, D B (Eds) *Globalization: Culture and Education in the New Millennium*, Los Angeles: University of California Press, 173–202.

Szelenyi, K and Chang, J C (2002) Educating immigrants: The community college role, *Community College Review* 30 (2), 55–73.

Teranishi, R, Suárez-Orozco, C and Suárez-Orozco, M (2011) Immigrants in community colleges, *Future of Children* 21 (1), 153–169.

Valdés, G (1992) Bilingual minorities and language issues in writing: Toward profession-wide responses to a new challenge, *Written Communication* 9 (1), 85–136.

Valdés, G and Figueroa, R (1994) *Bilingualism and Testing: A Special Case of Bias*, Westport: Ablex Publishing.

Venezia, A, Reeves Bracco, K and Nodine, T (2010) *One-shot deal? Students' perceptions of assessment and course placement in California community colleges*, available online: www.wested.org/online_pubs/oneshotdeal.pdf

Xiao, Y and Carless, D R (2013) Illustrating students' perceptions of English language assessment: Voices from China, *RELC Journal: A Journal of Language Teaching and Research* 44 (3), 319–340.

Appendix

Interview protocol

SECTION D: ASSESSMENT EXPERIENCES and EXPERIENCES (10 minutes) IN DEVELOPMENTAL AND ESL COURSES
Intent: To understand experiences with assessment and placement in developmental and ESL classes (if any).

Think about the tests you took when you entered college, such as COMPASS or other writing assessment tests or other placement exams like Accuplacer. These tests are designed to assess your abilities to do college level work.

D1. How did you do on the exam?
[PROBE for: How difficult did you find the test to be? Did you feel the test measured what you actually knew? Why or why not?]
D2. How did you respond once you learned about the results of the testing?
[IF they say they did not pass a test, ASK]
a. In what class were you placed after you took the test?
[For English/Math, GO TO D3 Developmental questions and for ESL, GO TO D4 ESL questions].

Developmental questions

D3. Tell me about the class you went to after you took the test.
a. How would you describe your overall experience in this class?
[PROBE for: Liked, learned a lot, bored, struggled, misplacement …]
b. Can you give me an example of this experience?

ESL questions

D4. Tell me about the first ESL class that you took after the test.
a. How would you describe your overall experience in this class? [PROBE for: liked, learned a lot, bored, struggled …]
b. Can you give me an example of this experience?

10 Washback and the reformed CET-4: Insights from students

Zhiling Wu

Indiana University of Pennsylvania, US

Motivation for the research

At the beginning of the 21st century, English has been widely recognized as the language of commerce, politics, academia, pop culture, and tourism, and more and more people worldwide are learning English. In China, more than 400 million people, about one third of the total population, are English learners and about 27 million of them are university students (Cheng and Curtis 2009). The National College English Test Band 4 (CET-4) is the only national test for non-English majors at the college level in China. Since its inception in 1987, it has drawn the largest number of test takers in the world (Jin and Yang 2006). In 2005, as part of the national Higher Education Undergraduate Level Teaching Quality and Teaching Reform Project proposed by the Chinese Ministry of Education (MoE) and Ministry of Finance (MoF), a reform of college English teaching and assessment began. Even though the CET-4 has been modified multiple times in the past years, the MoE launched a major reform for the CET-4 as part of the Higher Education Undergraduate Level Teaching Quality and Teaching Reform Project. The current study examines washback in three Chinese universities of different national rankings as a result of the reformed CET-4.

Review of the literature

Washback is defined simply as 'the effect of testing on teaching and learning' (Bailey 1996, Pearson 1988). Messick (1996:241) depicted washback as 'the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning'. The concept is also known as *backwash* in the literature (Alderson and Wall 1993). However, washback is the term that has currently gained prominence in language testing and applied linguistics circles, and it is now generally agreed that washback and backwash are synonymous with each other (Cheng 2005, Davies 1990, Green 2007).

As long as there is testing, there will be washback, and it can be either positive or negative. Reforming a test to maximize positive washback and minimize negative washback is important for all stakeholders and is an ongoing process as long as a test exists. Alderson and Wall (1993) hypothesize about the effects of testing on content, methodology, rate, degree, and attitude in language teaching and learning. Hughes (1993) then suggested that washback involves three core components: participants, processes, and products.

Hamp-Lyons (1997) has stressed that tests bring different meanings to different stakeholders. She made an appeal for more research on washback, especially research on students' views, to enhance test writers' professional responsibility in language testing. Green (2007:314) has concluded that even now students' perspectives remain 'under-investigated in the literature'. Green (2007) also emphasized that the meaning of washback for teachers was not equal to the meaning of washback for learners.

Of the limited number of empirical studies that have appeared on washback so far, the majority focus primarily on teachers' perspectives and on how their English teaching has been influenced by high-stakes tests, both in terms of what they teach and how they teach (e.g. see Manjarrés 2005, Wall and Alderson 1993, Watanabe 1996). In contrast, the perspectives of the test takers, who are the most immediate stakeholders, have been largely ignored. Shohamy (2001:97) has pointed out that test writers are not particularly interested in the test takers' voices, for 'in the testing literature test takers are often kept silent; their personal experiences are not heard or shared'. Fullan (2007:15) also lamented that in educational change, the students were rarely considered 'participants' but were rather considered 'potential beneficiaries of change'.

Even if students are assumed to be empty receptacles waiting for teachers to fill them with knowledge, the knowledge still needs to be digested and internalized by the students themselves. The goal for some tests is to promote Communicative Language Teaching (CLT) and to help learners develop communicative language competence, for which the expectation is that students be treated as active participants in the process. To provide a correct or complete picture of washback for any given test, it is important to balance the role of test takers, teachers, and other variables.

Most of the published studies on washback focus primarily on tests administered at the middle school or high school level (e.g. Andrews, Fullilove and Wong 2002, Cheng 2005, Luxia 2005, Shohamy 2001, Wall and Alderson 1993); consequently, washback on high-stakes, national English tests at the college level is underresearched (Yang and Weir 1998). One of the few studies that was done at college level was by Shih (2007). He examined the washback on the General English Proficiency Test (GEPT) in two private institutions in Taiwan. He found that the intensity of washback was stronger in the institution that requested its day division students to pass either the first stage of the GEPT's intermediate level or the school administered make-up exam to fulfill the degree requirement. The other institution did not have a GEPT degree requirement. The relationship between washback and school rankings has never been addressed. By setting the current study in three universities of different rankings, this research becomes one of the first to explore whether school rankings play a part in washback.

Theoretically, evidence of washback is related to the consequential aspect of construct validation enquiry, according to the unified validity argument made by Messick (1996). In this sense, the findings in the current study will contribute to the ongoing validation process of the reformed CET-4. Ross (2008) expresses concern about negative washback, particularly as it relates to the assumption that a foreign language test can predict aptitude or achievement. This chapter could also be of particular interest to large-scale norm-referenced language test developers, policy makers, researchers, and English teachers and learners at tertiary level not only in China, but also many other relevant EFL contexts.

The research context

The three universities involved in this research were all located in Shanghai, and they were selected based on a convenience sampling. They shall be referred to throughout the chapter as University A, University B and University C. As the biggest and most populous city in China, Shanghai attracts hundreds and thousands of students nationwide to its 30 universities and colleges. All the three selected universities in this study are science- and engineering-oriented universities. Their undergraduate population ranges from 10,000 to 20,000. According to www.cuaa.net, a website that is famous for its university rankings in China, in 2010, University A was ranked among the top 5, and University B and C were in the top 170 and 480 respectively among Chinese universities. Rankings were made on the basis of the number of prestigious scholars, publications, student quality, teacher quality, and availability of tangible resources, such as research funding, library resources, and university size. In other words, the three universities in the study in some way represent three levels of prestige in the Chinese hierarchy of higher education – highly reputed (University A), intermediate (University B), and ordinary (University C).

As in other Chinese universities, College English does not stand as a separate department but forms a part of the language department, which offers fundamental English courses for all freshmen and sophomores who are not English majors. All three universities applied for and participated in a 1-year pilot study of the College English Teaching and Learning Reform from February 2004 to February 2005 by the MoE. They all successfully passed the evaluation at the end of the study. In response to the reform, the three universities have different requirements as to when their students are eligible to take the CET-4. At University B and University C, all freshmen are divided into fast classes and ordinary classes based on an English placement test result. Students in the fast class are allowed to take the CET-4 in June of their first year, whereas students in the ordinary class must wait until their third semester (i.e. December of their second year). At University A, however, all students, without exception, must wait until the third semester to take the CET-4 for the first time.

Research questions

The research questions that framed this study were the following:

1. From students' perspectives, will washback from the CET-4 be the same or different in the three universities of different rankings?
2. If it is different, what is the difference and why?

Data collection

Based on the different administrative practices in the three universities and the fact that CET-4 was given only twice a year in June and December, a 4-phase study was conducted. In Phase 1, a researcher-constructed survey was given to classes that had agreed to participate in the study. The survey was designed to collect information about the effects of the CET-4. In Phase 2, students who had indicated their interest in participating further were selected for follow-up interviews. The selection was based upon variables such as gender, academic major, English classes registered, and their matriculation English grades. Because the test administration schedule at University A was different from Universities B and C, the study was expanded to include Phases 3 and 4 of data collection. In Phase 3, two more sophomore classes at University A were surveyed and four students from each class were selected to participate in the interviews and self-recordings in Phase 4. In addition to these students, six participants who would take the test in December from Phase 2 (two from each university) had a further follow-up interview and four of them also agreed to do self-recordings. Portable digital recorders were used. In total, 414 participants were surveyed, 34 were interviewed, and seven submitted their self-recordings. Please see the Appendix for the overall timeline of the study.

Results

Participants

A total of 414 students participated in the survey: 173 from University A, 125 from University B and 116 from University C. Ranging from the ages of 17 to

22, 68.12% of them were male (282), and 31.88% were female (132), which is a true reflection of the student population of science- and engineering-oriented universities in China.

Geographically, all of the eight official administrative districts in China are represented. Nevertheless, because all three universities are located in Shanghai, which is in Eastern China, 61.4% of the participants were from Eastern China. There were 12.6% from nearby Central China, and the percentages from other districts were all under 10%. Ninety-seven majors were involved in the study. The top three majors were telecommunications (11.4%), environmental studies (9.2%), and software engineering (6%).

Coming from different administrative districts, the participants had spent six to 20 years studying English, 23.4% of whom had studied English for seven years. Another 21.5% had studied English for 10 years. When participants' matriculation English exam grades are converted to a number out of a total of 100 points, the mean scores of the three universities are 85.70, 78.63 and 76.17 for Universities A, B, and C respectively, which matches the corresponding rankings of the three universities. At University A the lowest English score was 67, which is higher than the lowest score at University B (58) and University C (60).

The participants expressed different purposes for learning English at the college level. These purposes or goals were coded as follows: (a) to get high scores in the CET-4; (b) to pass required English courses and get high scores; (c) to learn further about English cultures to satisfy their own learning interests; (d) to improve comprehensive communicative ability in English; (e) to lay a good foundation in English to study abroad in the future; and (f) other. Descriptive data suggest that the majority of the students at University A were learning English at the college level because they were interested in improving their comprehensive communicative ability in English (d), whereas the majority of students at University B and University C were studying English to improve their scores on the CET-4 (a).

Washback in three universities

Students' perceptions of the effects of the CET-4 at the three universities are presented in the following areas: 1) overall CET-4 effect on the teaching of College English; 2) the frequency with which CET-4 was mentioned in classes; 3) the frequency of taking mock CET-4 or practice tests in the classroom and/or independently outside of class; 4) fast reading as a classroom practice; 5) intensive reading practice; and 6) knowledge of the CET-Spoken English Test (SET).

Overall CET-4 effect on the teaching of College English

In the survey, participants were asked to respond to a general question about the amount of influence they believed the CET-4 had on teaching English at the college level, and 51 out of 414 participants (12.3%) claimed that the CET-4 did not have any influence on how English was taught in their classrooms. From University A 16.8%, 10.4% from University B, and 5.2% from University C said that the CET-4 had no influence on their English learning, and 17.3%, 12.8% and 4.3% from Universities A, B, and C respectively reported that English instruction in their college classrooms had not been affected by the CET-4. In other words, the perceived impact of the reformed CET-4 on English teaching and learning from students' perspectives appeared strongest at University C and weakest at University A.

Frequency of the CET-4 mentioned in class

Participants were also asked to comment on the frequency with which CET-4 was mentioned in their classes. University C was the university that most frequently mentioned the CET-4 in its classes with 11%, 20.8%, and 39.7% of the participants responding that it was *often* mentioned at Universities A, B and C respectively, while 0%, 4.8% and 6.0% indicated it was *always* mentioned in classes.

ANOVA was run using participant data from the June CET-4 test administration to determine if there were significant differences among students' responses related to the frequency of teachers' mentioning the CET-4 in class at the three universities. Post hoc comparisons indicated that the mean score for University A (Mean (M) = 2.71, Standard deviation (SD) = 0.725) was significantly different from University C (M = 3.58, SD = 0.811), $p < 0.001$. The mean score for University C was significantly different from University B (M = 2.70, SD = 0.749), $p < 0.001$. There was no significant difference between participants' responses for Universities A and B. Therefore, among students who registered to take the CET-4 in June at University C, teachers mentioned the CET-4 significantly more often in class than their counterparts at University A and University B. This finding indicates that the reformed CET-4 seems to exert much stronger washback on English classes at University C than Universities A and B, the highly reputed and intermediate universities. The probable reason for this observation will be explored in the 'Discussion' section.

Mock CET-4 and practice tests in class

The survey also queried participants at the three universities about the frequency of doing mock CET-4 tests or taking past CET-4 tests for practice in classes. For students who registered to take the CET-4 in June, ANOVA indicated that there was a statistically significant difference among participants at the three universities. Post hoc comparisons indicated that the mean score for University A (M = 2.10, SD = 0.725) was significantly different from University C (M = 3.23, SD = 0.820), $p < 0.001$. The mean score for University C was significantly different from University B (M = 1.85, SD = 0.744), $p < 0.001$. There was no significant difference between Universities A and B. Similar results concerning mock CET-4 tests practice were found for participants who registered to take the CET-4 in December. Therefore, students who registered for the reformed CET-4 in both June and December at University C did significantly more mock CET-4 tests or past CET-4 tests for practice in class than their counterparts at University A and University B.

This finding further suggests that the reformed CET-4 had stronger washback on students and English teachers in classes at University C. English teachers at University C not only mentioned the CET-4 more often, but also set aside more class time for students to practice mock CET-4 tests.

Mock CET-4 and practice tests out of class

Data were also collected about the frequency of taking CET-4 mock tests or past CET-4 tests out of class. The survey data suggest that 41.4% of University C, 24.8% of University B and 24.3% of University A participants *sometimes* take CET-4 mock tests or past CET-4 tests by themselves. 17.2% from University C, 9.6% from University B, and 4% at University A indicated they *often* took the tests; 2.6% at University C, 0.8% at University B and 0.6% at University A indicated they *always* took them.

ANOVA indicated that there was a statistically significant difference in the three universities in terms of the frequency of doing the CET-4 mock tests or taking past CET-4 tests out of class ($p < 0.001$). Post hoc comparisons showed that the mean score for University A (M = 2.06, SD = 0.857) was significantly different from University C (M = 2.71, SD = 0.987), $p < 0.001$. The mean score for University C was significantly different from University B (M = 2.13, SD = 1.016), $p < 0.001$, but there was no significant difference between University A and University B.

Therefore, at University C it is not just the teachers who pushed the students to become familiarized with the reformed CET-4; the students also chose to spend a considerable amount of time preparing for the CET-4 outside of class. Once again, the data suggest that the reformed CET-4 exerts much less washback effect on English classes in highly reputed or intermediate universities, particularly the former, than universities with ordinary ranking.

Fast reading practice

Fast reading is one of the newly added items in the reformed CET-4. Fast reading is a skill that the MoE expects test takers to practice and master through the reform of the CET-4. In class, the survey data showed that 4.6% of students at University A indicated they *often* had fast reading practice in class, as compared to 22.4% at University B and 28.4% at University C, which demonstrates that there was less washback from the CET-4 change to include the fast reading practice at University A.

An ANOVA test showed that there was a statistically significant difference when it came to the participants' perceptions about the frequency of doing fast reading in class between the three universities (p < 0.001). Post hoc comparisons indicated that the mean score for University A (M = 2.32, SD = 0.789) was significantly different from University B (M = 2.69, SD = 1.139), p = 0.003. The mean score for University A was also significantly different from University C (M = 2.92, SD = 0.988), p < 0.001. There was no significant difference between University B and University C.

Because students had registered for the CET-4 either in June or December, the data were also compared again by registration month. For students who registered for the CET-4 in December, students at University C still did significantly more fast reading than their counterparts at University A and University B (p < 0.001). Likewise, among students who registered for the CET-4 in June, students at University A did significantly less fast reading in class than their counterparts at University B and University C. There was no significant difference between University B and University C (p < 0.001).

Students from Universities B and C in the follow-up interviews explained that fast reading is often served as a warm-up exercise at the beginning of a class. It did not take too much class time if teachers had many other things to cover in one class. Given the nature of its questions, teachers did not need to analyze why a certain answer was correct other than providing the students with the answers. The interviewees said this was a question type that both teachers and students found rewarding in the course of the CET-4 preparation, especially those who wanted to boost their score in a short time. Next, the washback effect of the practice of the other type of reading – intensive reading – is discussed.

Intensive reading practice

In the reformed CET-4, the weight of intensive reading has been decreased and makes up 20% of the total score in the reformed CET-4. Survey data suggested that 6.9% of students at University A, but 16.8% at University B, and 27.6% at University C indicated they *often* did intensive reading in class. 1.7% at University A, 2.4% at University B, and 1.7% at University C indicated they *always* did intensive reading in class.

An ANOVA test indicated that there was a statistically significant difference in participants' responses to the frequency of doing intensive reading in class among the three universities (p < 0.001). Post hoc comparisons indicated that the mean score for University A (M = 2.36, SD = 0.883) was significantly different from University B (M = 2.62, SD = 0.989), p = 0.039. The mean score for University A was significantly different from University C (M = 3.01, SD = 0.839), p < 0.001. The mean score for University B was also significantly different from University C, *p* = 0.003. Therefore, students at University A indicated that they did significantly less intensive reading in class than their counterparts at University B and University C. Students at University C, however, did significantly more intensive reading than students at Universities A and B. These differences suggest that a slight decrease in the weight of the intensive reading comprehension in the reformed CET-4 did not result in changing the emphasis in the English classes to speaking skills in universities of lower ranking. According to the interviewees, the reading skill was easier to practice compared with writing and speaking skills, since the latter needed much composing time and one-on-one attention.

Knowledge of the CET-SET test

Even though the MoE wanted to improve students' level of spoken English, few English teachers seemed to provide students with necessary information about the test or encourage them to take the CET-SET test, a separate component of the CET which measures oral English proficiency. About 35.3% of the interviewees knew that a certain score on the CET-4 was required, yet none could give the exact required score during the interview. Many interviewees at University C openly claimed that they did not have confidence in their skills to pass the CET-SET. For example, C Gao said that her oral English would be too poor to pass. C Liu said that he did not want to 'embarrass himself by going through it' since he had little practice in oral English. C R Zhao, however, confessed that she had never heard of the CET-SET. Apparently, without including the CET-SET as an essential part of the reformed CET-4 and making it accessible to all college students, it would be hard to motivate teachers and students to spend more time on spoken English. The MoE expected positive washback from the reformed CET-4 in this aspect, but it has not yet materialized.

Discussion

Data analyses clearly showed that the reformed CET-4 created much less washback at University A in the specific areas investigated than those at Universities B and C. This finding could be interpreted that the reformed CET-4 exerted much less washback in highly reputed universities. The question that researchers and test developers should tackle then is why the reformed CET-4 had rather different effects on students at University A from the other two universities. The following are possible reasons for this outcome based on the surveys and follow-up interviews.

First of all, the long-term goals for the participants are quite different for University A and University C. At University C, 63.8% indicated they would go to work directly after getting their bachelor's degree. Upon graduation at University A, about one third of the graduates go directly to work, while two thirds choose to continue their studies, with about one third going on to graduate school in China and one third going to graduate school abroad. At University A, 35.8% of the students and at University B 39.2% of students planned to go to graduate school in China. When students plan to go to a graduate school in China, the CET-6 is more important than the CET-4 because the CET-6 is a test required for graduate level studies.

The percentage of students who believed that the CET-4 was more important than the CET-6 is considerably smaller in Universities A and B than in University C, with 13.3% students from University A, 22.4% from University B and 40.5% from University C responding to the importance of the CET-4. Students who plan to go to a graduate school abroad have to score well in the Test of English as a Foreign Language (TOEFL$_{®}$), or *International English Language Testing System (IELTS)*, and/or the Graduate Record Examinations (GRE). These are the main English language tests that are required for foreign students to study and/or get assistantships or scholarships in English-speaking countries; consequently, students planning to study abroad would likely focus on those exams, not the CET-4.

Second, the teaching and learning resources at University A are much richer than Universities B and C. All student dorms at University A had access to the high-speed internet, as well as to English TV channels, such as China Central Television Channel-9 (CCTV-9) and the Discovery Channel (programming in English). The freshmen at University C, however, did not have internet access in student dorms. In addition, almost all classrooms at University A were equipped with a computer connected to the internet and a projector for the teacher's use. At Universities B and C, in contrast, not every classroom had a computer. Even if the classroom had a computer, it might not be connected to the internet. Without easy access to online technologies, students at University C would have fewer opportunities in class or out of class to take advantage of the unlimited resources online to experience using English for communicative purposes, or to learn English beyond passing tests and cultivate more genuine interest in English learning.

Third, even though most of the students are required to study English for the first two years in the three universities, sophomores at University A are free to choose from a range of more specific English courses, such as English Literature, Public Speaking, English Writing, English Translation, and Audio and Visual English. Compared to the one-size-fits-all general College English course offered at the other universities, these tailor-made English courses have the potential to better maintain students' interest in English learning while preparing for the reformed CET-4.

Fourth, University A is the only university among the three that offered courses taught in English partially or completely in the students' major fields of study. Nine participants from University A had actually either taken semi-bilingual courses in their majors, or courses that were fully bilingual, such as

Fundamentals of Digital Electronics, Physics, Introduction to Engineering, Microeconomics, and Signal and System, etc. The teachers of these courses are native Chinese speakers who studied and/or graduated from universities overseas and have a good command of English. The required textbooks and resources of these courses were all in English. Teachers' lectures were also delivered through PowerPoint in English, and students were often encouraged to complete their assignments and/or major tests in English. Even though some students complained about their teachers' accent, the university has sent students a message that learning English is not just about passing tests. Rather, it is a means to open more windows for thinking, absorbing new knowledge, becoming a member of a larger discourse community, and eventually creating an opportunity for making one's life more meaningful and interesting. When students at University A were exposed to these opportunities, it is possible that they became less concerned about passing English tests and instead became more interested in using English to increase their knowledge in their major fields of study and begin to relate to the international academic community.

In contrast, except the general College English courses that are required by the national curriculum, no other interest-based English courses or bilingual courses are provided at Universities B and C. Rather, a CET-4 Test Preparation course for credit is offered at University B, and an after-class private intensive CET-4 preparation class is available at University C. Even though the MoE reiterated the national policy – that the passing of the CET-4 and the college diploma should never be tied together – in a news conference in February 2005, in the interviews, many students, particularly those from Universities B and C, stated that they knew nothing about the national policy. This perception persists, although the three universities in this study have also publicly announced that they would no longer require their graduates to pass the CET-4 to qualify for their diplomas/bachelor's degrees.

In fact, as many as 10 interviewees from Universities B and C still believed that there was a connection between the CET-4 score and the awarding of the college diploma/degree. They claimed that officials at the two universities provided no official clarification for them. Instead, they had confirmed with other sources whom they considered reliable: B J Li from University B said that his first year College English teacher had told him that every college student had to pass the CET-4 to be granted a diploma. C Gao then recalled that on various occasions at University C, both the university president and teachers had stressed that they could not graduate without passing the CET-4.

Additional information regarding graduation requirements has also been distorted. For instance, B Song from University B maintained that his college friends told him that at University B, passing both Higher Mathematics and the CET-4 were required to get the diploma. Likewise, C Feng said his

English teacher had told the students that they would be 'safe' once they passed the CET-4 and that the CET-4 was still an indispensable prerequisite to getting a diploma at college. He even argued that University C had lower requirements, because according to what he had heard, in other universities, the upper-level CET-6 was the required test to pass for graduation.

It is not clear exactly why some administrators or teachers spread outdated and incorrect information on the relationship of CET-4 scores and graduation. As might be expected, the students tended to believe the authorities when faced with conflicting statements. Because students felt that the reformed CET-4 was still a high-stakes test, as high as 38.4% surveyed at University B and 40.5% at University C set getting high scores in the CET-4 as their primary goal of English learning at college. Consequently, there was more washback from the reformed CET-4 on students at less prestigious universities, particularly ordinary-ranking universities. Correct information regarding the CET-4, nevertheless, should be provided to all students.

Last, but certainly not least, the relaxed CET-4 preparation environment at University A could also have played a part in washback as there seemed less pressure exerted on students by the CET-4. By admitting top students from each province, University A was likely to have selected students with a good foundation in English. Six students from University A in Phase 1 and five students in Phase 4 stated that they were certain they would have no difficulty attaining the minimum required CET-4 score. Among them, five students said that either English teachers or senior students had told them that the reformed CET-4 was an easy test. Some even claimed that the reformed CET-4 was no more difficult than the English matriculation exam, and that they could have passed it in senior high school. A Liang from University A said that around him, he only saw students preparing for the TOEFL$_{®}$ test, not the reformed CET-4. On the contrary, at Universities B and C, only two interviewees in these two universities expressed absolute confidence in getting a satisfactory score. On the other end of the spectrum, one student from University C even mentioned that he knew three seniors who had failed the CET-4 twice.

It is quite possible that the overall campus environment might have had an influence on the amount of time students spent on preparation for the CET-4. In his washback model, Green (2007) predicted that if participants do not perceive the importance of a test and consider the test not particularly difficult (as is the case for the reformed CET-4 for many interviewees at University A), it will not exert the expected influence. Negative stories about difficulty in passing the CET-4, and the belief that failing to score well on the CET-4 could result in graduation without a diploma, made the CET-4 preparation atmosphere at University C more stressful than that at University A or B.

In order to see more positive washback at universities of higher ranking, especially those highly reputed universities, a computerized adaptive CET-4 test could be created, given that the computer-based CET-4 has already been pilot tested and predicted to take the place of the paper-based version in the future. At the same time, a speaking test should be included as an essential component in the CET-4 to substantially change the English classes that are largely reading/vocabulary based, as claimed by most participants. Many of them admit that they are afraid of speaking English in front of other people; yet speaking is the skill they deem as the most important. Of course, more training should be provided to teachers to help them make a smooth transition to a curriculum more centered on the development of speaking skills.

In addition, some students from Shanghai even reported that they had been given the CET-4 reading, listening, and writing exercises at their senior high school when they were preparing for the entrance exam to college. One student from Shanghai mentioned that many of his high school friends had registered for summer preparation classes at the New Oriental English Training School after graduating from high school. These classes included English interpretation, oral English development for intermediate or advanced level students, and the TOEFL® or *IELTS* preparation classes. Some took one or more English tests while they were just attending senior high school. At the time of the interview, one participant who is not from Shanghai at University A was actually taking an intermediate level English interpretation class on weekends. She said she was surprised to find out that most of her classmates there were high school students in Shanghai. On the other hand, participants from Northwestern China said that they had little access to resources described by Shanghai students and that their high school English classes did not prepare them adequately. One interviewee from Gansu in Northwestern China at University A said that he did not understand a single sentence spoken in his first English class at college, and two other students from the same area from Universities B and C admitted that they were at a loss in their College English classes at the beginning because their high school English classes had always been taught in Chinese. While listening is not a component on the entrance exam to the colleges in Northwest China, it is one of the two emphasized skills according the reformed CET-4; consequently, there is a lot of pressure on students with lower English proficiency levels from the less developed Northwest China to do well on the CET-4, regardless of which university he/she is attending.

At present it is quite hard for students at lower-ranked universities to compete with students at higher-ranked universities. The differences in availability of and access to educational resources between the inner cities and coastal cities, Southern and Northern China, and urban and rural areas, have created big gaps in students' English foundations and are indirectly linked to some of the negative washback of the CET-4. Uneven resources provided by universities of different rankings further widen the gap between advanced and less advanced students. It is generally true that students recruited by universities of the first tier in the nation have a better foundation in English and, consequently, achieve higher scores on exams. Some majors require higher levels of English proficiency, so students with higher proficiency levels in English have access to the most sought-after majors and more opportunities for upward mobility in Chinese society.

Trying to improve English language education mainly through positive washback from a test such as the reformed CET-4 is a difficult task and a long-term endeavor. Changes in the overall English language ability profile in China cannot be executed successfully without the consideration of factors other than positive washback from the changes implemented in the reformed CET-4. However, because there is no easy way to resolve the issues of resource availability and access prior to entrance into universities, taking a closer look at the practices regarding English education at the university level is still a logical place to begin.

Tests, such as the CET-4, can be tools to stimulate learning and create positive washback. However, at no time should passing a test be the only reason for students to learn. University students' interest in learning English for communication purposes should not be sacrificed to a 'passing' score in one test. In this study, University A offers an array of courses in English for students to learn about new topics and content. Some major classes are also taught in English and may be effective in improving English skills. At Universities B and C, however, the focus is on College English courses and preparation for the CET-4. In order to compete with higher-ranked universities, lower-ranked ones might consider an expansion of the English curriculum to include English courses offering specific content. Changing the English curricula for universities is a manageable task and one that may have the potential to stimulate students and capitalize on intrinsic motivational factors that occur naturally when students are participating in activities they have chosen and in which they have an interest. In the long term, these steps could help create more desired positive washback as a result of the reformed CET-4.

Acknowledgements

Special thanks go to The International Research Foundation for English Language Education (TIRF) for its financial support, to Dr Michael M Williamson, my dissertation advisor, for his intellectual guidance and humanity, and to all the people who helped and believed in me in this writing process.

References

Alderson, J C and Wall, D (1993) Does washback exist? *Applied Linguistics* 14, 115–129.

Andrews, S, Fullilove, J and Wong, Y (2002) Targeting washback: A case-study, *System* 30, 207–223.

Bailey, K M (1996) Working for washback: A review of the washback concept, *Language Testing* 13, 257–279.

Cheng, L (2005) *Changing Language Teaching Through Language Testing: A Washback Study*, Studies in Language Testing volume 21, Cambridge: UCLES/Cambridge University Press.

Cheng, L and Curtis, A (2009) The realities of English language assessment and the Chinese learner in China and beyond, in Cheng, L and Curtis, A (Eds) *English Language Assessment and the Chinese Learner*, New York: Routledge, 3–12.

Davies, A (1990) *Principles of Language Testing*, Oxford: Blackwell.

Fullan, M (2007) *The New Meaning of Educational Change* (4th edition), New York: Teachers College Press.

Green, A (2007) *IELTS Washback in Context: Preparation for Academic Writing in Higher Education*, Studies in Language Testing volume 25, Cambridge: UCLES/Cambridge University Press.

Hamp-Lyons, L (1997) Washback, impact and validity: Ethical concerns, *Language Testing* 14, 295–303.

Hughes, A (1993) *Backwash and TOEFL 2000*, unpublished manuscript, University of Reading.

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

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

Manjarrés, N B (2005) Washback of the foreign language test of the state examinations in Colombia: A case study, *Arizona Working Papers in SLAT* 12, 1–19.

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

Pearson, I (1988) Tests as levers for change, in Chamberlain D, and Baumgardner, R J (Eds) *ESP in the Classroom: Practice and Evaluation*, London: Modern English, 98–107.

Ross, J S (2008) Language testing in Asia: Evolution, innovation, and policy challenges, *Language Testing* 25, 5–13.

Shih, C (2007) A new washback model of students' learning, *The Canadian Modern Language Review* 64, 135–161.

Shohamy, E (2001) *The Power of Tests: A Critical Perspective of the Uses of Language Tests*, London: Longman.

Wall, D and Alderson, J C (1993) Examining washback: The Sri Lankan impact study, *Language Testing* 10, 41–69.

Watanabe, Y (1996) Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research, *Language Testing* 13, 318–333.

Yang, H and Weir, C J (1998) *Empirical Bases for Construct Validation: The College English Test – A Case Study*, Shanghai: Shanghai Foreign Language Education Press.

Appendix

An overall timeline of the study

Phase 1 and Phase 2

University	Class type	Phase 1 – Survey 5/8/2010 – 5/18/2010	Phase 2 – Interview 5/19/2010 – 6/30/2010
		Number of participants	
University A	CET-4 in June	29	10
	CET-4 in December Teacher 1	38	1
	CET-4 in December Teacher 2	35	1
	Total	102	12
University B	CET-4 ability training Practice class	55	2
	Fast class	31	4
	Ordinary class	39	2
	Total	125	8
University C	Fast class Teacher 1	41	4
	Fast class Teacher 2	39	3
	Ordinary class	36	2
	Total	116	9
Phase 1 and 2	Grand total	343	29

Phase 3 and 4 in University A

Phase	Phase 3 Survey	Phase 4 Interview	Phase 4 Self-recording
Time	11/1/2010 – 11/18/2010	11/19/2010 – 12/20/2010	
Class	Number of participants		
Teacher 1	45	3	2
Teacher 2	26	2	1
Total	71	5	3

Phase 3 and 4: A further follow-up of participants in Phase 2

Phase	Phase 3 Survey returned	Phase 4 Interviews	Phase 4 Self-recordings
Time	**11/1/2010 – 11/18/2010**	**11/19/2010 – 12/20/2010**	
University	**Number of participants**		
University A	2	2	2
University B	2	2	1
University C	1	2	1
Total	5	6	4

11 The impact of strategic processing on lexico-grammar test performance

Nick Zhiwei Bi

University of Shanghai for Science and Technology, China

Motivation for the research

Communicative language ability

In the past 35 years views on how to interpret language ability have shifted from a unitary/global level to a multi-dimensional level (Bachman 2014). The most influential multi-dimensional model has been the *communicative language ability* (CLA) model originally proposed by Bachman (1990) and later refined and modified by Bachman and Palmer (1996). The CLA model specifies both the linguistic and non-linguistic components underlying language use (Purpura 2008). Within the multi-dimensional CLA framework and focusing specifically on assessment, Purpura (2004, 2013) proposes a theoretical framework that accounts for the relationships among grammatical knowledge, ability, and performance. Purpura's model was developed on the basis of Bachman and Palmer's CLA model (1996) but focused on the grammatical performance of second language (L2) learners. Purpura defines grammatical ability as composed of *grammatical knowledge* (i.e. a range of linguistic forms including the -s affix, word order and the semantic meanings associated with these forms, such as the plurality in nouns), and *strategic competence* (i.e. a set of cognitive and metacognitive strategies associated with the use or activation of grammatical knowledge in context). The operation of strategic competence with grammatical knowledge is one of the most important factors affecting test takers' grammar test performance.

This framework highlights the importance of strategic competence as an integral part of students' grammatical ability. Unlike grammatical knowledge, strategic competence is not directly scored, but it is assumed to influence a grammatical score. We, therefore, need to know the extent to which test takers use cognitive and metacognitive strategic processes in ways that might account for their performance on a grammar test. In language assessment, it is important to gather the so-called cognitive validity of a language test. *Cognitive validity* is the extent to which a language test involves the cognitive processes or skills associated with the test constructs and context of language use (O'Sullivan and Weir 2011, Weir 2005). Cohen (2011, 2014) describes test takers' strategy use during test tasks and stresses the importance of relating cognitive processing to identify test constructs as a method of validation.

To date, strategic competence has remained absent from the scoring rubrics for the majority of language tests, though language testing researchers have recognized that test takers' strategic processing can provide insights regarding test validity. Empirical research has so far been lacking in relation to the precise nature of strategic competence as applied in language testing contexts (Swain, Huang, Barkaoui, Brooks and Lapkin 2009). With regard to grammar assessment, though Purpura (2004, 2013) has presented a comprehensive model of grammatical ability in communicative language use and assessment methods, he has not provided an approach to assessing and inferring strategic processes in grammar assessment. Similarly, according to the literature, few studies have used empirical data to validate Purpura's (2004) model. Hence, there is a need for researchers to explore and explain this issue further through empirical investigations.

The insightful ideas from the language testing literature mentioned previously provided the initial motive for the current study. The study presented in this chapter goes on to examine one important strand of language testing research, as suggested by Bachman (2000), by looking at the factors affecting language testing performance. In particular, further empirical studies were conducted as part of the research to validate Bachman's (1990) theoretical model of CLA, because CLA is considered the key contributor to language performance. Furthermore, this study aims to complement Purpura's grammatical assessment framework by proposing a *human information processing* approach for assessing the lexico-grammatical strategies use of L2 test takers. This approach is built upon the current theory of language learner strategy use (e.g. Cohen 2011, Macaro 2006, Oxford 2011) and a model of strategic processing (e.g. Bachman and Palmer 2010, Phakiti 2007).

Strategic processing in test taking

Strategic processing or strategy use research has been studied for more than 30 years; however, there seem to be a number of unresolved issues and questions regarding strategy use research in general language acquisition and in test-taking contexts. In relation to L2 learning strategy research, Macaro (2006) has suggested that a lack of consensus with regard to the constructs of strategy use has contributed significantly to undermining the development of L2 learner strategy use research. By the same token, research into strategy use in test-taking situations also faces this dilemma because it seems that researchers have not yet reached an agreement on what kinds of strategic processes are related to students' test performances. In particular, as language testing is a special case of L2 use based on certain tasks, strategy use during test completion should reflect L2 test takers' knowledge of strategy use in their long-term and working memories. That is to say, research should include the constructs of strategy use in both general and specific language use activities, such as language testing. Although the ways in which strategic processing is related to L2 test performances have been presented and tested in several empirical studies, research in this area is still limited. As strategic processing is highly complex and skill specific, a lack of knowledge in the area of how strategy use affects lexico-grammatical performance has shown the need to conduct further empirical investigation in this area.

Cohen (1987) points out that L2 learners' reported strategic behaviors might be quite different from their actual learning and use activities. In other words, L2 learners might inaccurately report their strategic behaviors in such situations. To make more accurate inferences about L2 learners' strategic processing, there is a need to make a distinction between different types of strategies in general L2 language learning and specific language use activities (Cohen 2011, Oxford 2011). Such a distinction would also be helpful for the understanding of strategic processing in test-taking situations. Generally, three types of strategies may be considered to have influence on a language test.

The first type of strategy use is referred to as a general learning strategy use. It was proposed by Cohen (2011) and refers to test takers' typical or general use of L2 language strategies. This type of strategy use includes one's perceived knowledge of general language strategies or strategy use that is free of context or not related to a specific context. Empirical evidence still suggests that strategy use may have both a direct and an indirect effect on test performance (e.g. Phakiti 2003, Purpura 1999, Song 2005). The second type of strategy use includes test takers' perceived knowledge of the strategies they would employ when facing a specific task, such as a test. Phakiti (2008a) used the term *trait strategies* to define L2 test takers' perceived strategy use in skill-specific tasks in non-specific contexts. The third type of strategy use is defined as the perceived knowledge of actual strategy use in a specific context. This type of strategy use is captured in Phakiti's definition of them as *state strategies* (Phakiti 2007).

The relationships among each of the above three strategies and L2 test performance have been documented in a number of empirical investigations, and there have been two ways of looking at strategy use in language assessment research. Some researchers (e.g. Purpura 1999, Song 2005) have adopted their conceptualization of strategy use from general perceived long-term language strategy use and related these strategies to test performance, while other researchers have sought to look at strategic processing solely under test-taking conditions (e.g. Phakiti 2003, 2007).

So far there has been no research that brings the three types of strategy use together as a cluster of strategies and examines the complex relationships among them. The purpose of the current study is to use empirical data to investigate the nature of these relationships. The current study proposes a structural model to establish the nature of the strategic processes as measured by different strategy use variables. Initially, the theoretical model proposed by Bachman and Palmer (1996) suggested that strategic competence should include a set of higher-order metacognitive strategy uses. However, empirical evidence presented by Purpura (1999) and Phakiti (2007, 2008a, 2008b) has shown that the concept of strategic competence in Bachman and Palmer's (1996) CLA model needs to be expanded and that the study of strategic competence needs to be researched with reference to metacognitive theory (McNamara 1996). In addition, Phakiti (2007, 2008a, 2008b) and Purpura (1999) state that the construct of strategic competence should include both metacognitive and cognitive factors. Phakiti (2008a) also asserts that employing both trait and state types of strategy use to explain strategic competence marks a step forward in this area of research. However, previous studies still have not taken into consideration all possible strategy use variables. The complex nature of strategic processing suggests that a comprehensive account of strategy use variables in long-term memory may be explained under the broad umbrella of strategic competence. Lastly, as Bachman and Palmer's CLA model is theoretically based and lacks empirical support (Phakiti 2007, Purpura 1999), there is a great need to use actual empirical data to investigate the relationship between strategic processing and L2 test performance.

Research questions

Based on the existing research outlined in the discussion above, the current study offers several improvements over previous research designs to help us better understand the notions of strategic competence and strategy use in language test performance. The following three issues, relatively unexplored in the existing literature, are of particular interest in this study:

- the relation between general learning strategy use strategies and test-taking strategies
- the nature of strategic processing (i.e. a consideration of general learner use, trait, and state variables), and
- the relationship of strategic processing to lexico-grammar test performance.

In order to investigate these issues, the current study set out to conduct a preliminary investigation to answer the following two research questions:

1. What are the relationships between L2 test takers' lexico-grammar use, trait and state strategic processing?
2. To what extent are different kinds of strategic processing related to test takers' lexico-grammar test performance?

Research design

Setting and participants

The data collection was conducted at a Chinese university in Guangzhou, mainland China, where English is considered a foreign language and is compulsory at various levels of education. This study focused on intermediate level EFL students at a Chinese university who were considered suitable for this study for two reasons. First, they had studied English grammar by the time they had completed their high school certificate (Ministry of Education 2004). Second, they had just taken the National Matriculation English Test (NMET) at the end of their high school and were about to take the College English Tests (CET) 4 and 6, which emphasize (30% of the tests) students' lexico-grammatical knowledge (National College English Testing Committee 2006). The lexico-grammar test in the present study was considered suitable to measure their English grammatical ability that would then be linked to the questionnaire data.

There were more than 460 participants who were recruited to voluntarily participate in the research. The number of students was reduced to 416 after the preliminary analyses and Rasch analysis were done for misfitting and outlier test takers. These students ranged in age from 19 to 22 (mean = 20.50, standard deviation (SD) = 0.67). They were from different departments in one faculty and were majoring in accounting, business enterprise management, international trade and business, e-commerce, business English, and marketing.

Research instruments

The current study used a questionnaire as the main method to assess test takers' lexico-grammatical strategy use. It is clear that we can examine L2 lexico-grammatical strategies through various testing techniques (think-aloud, interviews); however, no matter what methods we draw on to examine perceived strategy use in either general L2 learning or specific use conditions, we can never fully understand the extent to which L2 learners employ certain strategies (Phakiti 2007). Therefore, the best scenario seems to be gathering data that documents possible strategic processes that may assist test takers in completing their test tasks. The aim in the current study is to have a more comprehensive understanding of test takers' strategic processes. To this end, three types of questionnaires were used: 1) Learner Strategy Use Questionnaires, 2) Trait Strategy Use Questionnaires, and 3) State Strategy Use Questionnaires.

The Learner Strategy Use Questionnaire

The first type of questionnaire, the Learner Strategy Use Questionnaire, measures test takers' knowledge about their strategy use when applying lexico-grammar knowledge. This questionnaire was designed for use before test takers completed the lexico-grammar test. The questionnaire allows for inferences to be made about test takers' use of general strategic processes that are used in general contexts. For example, questions or items are written in the simple present tense to determine what L2 learners normally think and do in language use situations. A 38-item questionnaire was designed that targeted five categories of cognitive strategies (i.e. comprehending, retrieval, memory, grammar, and vocabulary) and three categories of metacognitive strategies (i.e. planning, monitoring, and evaluating) (see Table 11.1 for taxonomies and examples from the Learner Strategy Use Questionnaire).

Table 11.1 Taxonomies and examples of cognitive and metacognitive strategy use in general language use (38 items)

Strategy type	Sub-class	Examples
Cognitive strategies	Comprehending	I use correct grammar rules or vocabulary by looking at examples from various sources of how to use grammar or a word or expression.
	Retrieval	I use my knowledge of how structures change their forms (e.g. from a noun to an adjective, from an adjective to an adverb) when producing correct words.
	Memory	I relate knowledge I have previously learned in order to improve my grammar and vocabulary use.
	Grammar	I make sure the most appropriate and meaningful grammatical structure is used.
	Vocabulary	I choose the most appropriate word to fit in a sentence.
Metacognitive strategies	Planning	I fully plan how to use appropriate grammar rules and vocabulary in making new sentences.
	Monitoring	I know when I make mistakes in grammar or vocabulary.
	Evaluating	I check over in my mind how I have performed in grammar and vocabulary use.

The Trait Strategy Use Questionnaire

The second questionnaire, the Trait Strategy Use Questionnaire, measures test takers' knowledge about their strategic processing strategies in test-taking situations. This questionnaire is to be used before test takers complete a lexico-grammar test and allows us to infer test takers' general strategic processes outside the context of a given lexico-grammar test. A 43-item questionnaire was produced targeting five categories of cognitive strategies (i.e. comprehending, retrieval, memory, grammar, and vocabulary) and four categories of metacognitive strategies (i.e. test taking, planning, monitoring, and evaluating) (see Table 11.2 for taxonomies and examples from the Trait Strategy Use Questionnaire).

Table 11.2 Taxonomies of trait cognitive and metacognitive strategies (43 items)

Strategy type	Sub-class	Examples
Cognitive strategies	Comprehending	I paraphrase rules or words given in test tasks, because I can understand better in my own words.
	Retrieval	I use my memory of English grammar rules or words and apply them in test tasks.
	Memory	I underline important information in test tasks.
	Grammar	I identify relationships within and between sentences in test tasks.
	Vocabulary	I search for words that make my answer meaningful and appropriate to test tasks.
Metacognitive strategies	Test taking	I look through the whole test and try to get a general idea of the topics or the weighting of each task.
	Planning	I make sure I understand what has to be done and how to do it.
	Monitoring	During the whole test, I concentrate on what I am doing.
	Evaluating	I double-check all the answers before submitting the test paper.

The State Strategy Use Questionnaire

The third questionnaire, the State Strategy Use Questionnaire, measures the degree to which test takers perceive themselves to use cognitive and metacognitive strategies during a lexico-grammar test. It was administered after test takers had completed the lexico-grammar test, and hence was a retrospective questionnaire. The questionnaire measures test takers' strategic regulation so questions or items are written in the past simple tense. There were 50 items in the state questionnaire, and the underlying categories of state cognitive and metacognitive strategies were the same as in the Trait Strategy Use Questionnaire (see Table 11.3 for taxonomies and examples from the State Strategy Use Questionnaire).

Table 11.3 Taxonomies of state cognitive and metacognitive strategies (50 items)

Strategy type	Sub-class	Examples
Cognitive strategies	Comprehending	I tried to understand the context in which the text occurs in test tasks.
	Retrieval	I used the information I knew about the test tasks to help me to find suitable words or sentence structures.
	Memory	I memorized sentence structures that are often repeated in the test tasks in order to help me better complete the test tasks.
	Grammar	I made sure that I applied suitable grammatical knowledge to deal with the different test tasks.
	Vocabulary	I used knowledge of word stems (e.g. friendship) and prefixes (e.g. unhappy) or suffixes (e.g. childhood) to complete the test tasks.
Metacognitive strategies	Test taking	I looked up the test examples to familiarize myself with what I had to do.
	Planning	I knew what to do if my plans did not work efficiently.
	Monitoring	I was aware of my previous mistakes in order to avoid repeated mistakes.
	Evaluating	I checked my answers and saw whether they were accurate or reasonable.

Data collection

Students were asked to complete the preliminary Learner Strategy Use and Trait Strategy Use Questionnaires one week before taking the lexico-grammar test (average completion time: 40 minutes). A week later, students took the grammar test, which lasted 45 minutes, followed by the State Strategy Use Questionnaire (average completion time: 30 minutes). The lexico-grammar test will be described below. Table 11.4 summarizes the planned procedure for achieving the research goal of the current study.

Table 11.4 Data collection procedures

Strategic processing and lexico-grammar test performance
Stage 1: Test takers answer general Learner Strategy Use Trait Strategy Use Questionnaires one week before completing a lexico-grammar test.
After 1 week
Stage 2: Test takers complete a lexico-grammar test and answer a State Strategy Use Questionnaire immediately after the lexico-grammar test is completed.

The lexico-grammatical test

In order to capture the strategic processes related to a lexico-grammatical test, the study used a retired *First Certificate in English* (*FCE*; now known as *Cambridge English: First*) examination section developed by Cambridge English – Use of English – for intermediate-level students. The Use of English section, which assesses lexico-grammatical knowledge, consists of four parts (a total of 42 questions). Test takers were given 45 minutes to complete this test. The details of the test are as follows:

- Section 1: Cloze items (to measure knowledge of lexical meaning)
- Section 2: Gap-filling items (to measure knowledge of morphosyntactic forms and meanings and lexical meanings)
- Section 3: Word formation items (to measure knowledge of lexical forms)
- Section 4: Sentence transformation items (to measure knowledge of morphosyntactic forms and meanings and lexical forms).

There are several reasons why sections of the retired *FCE* were chosen for this study. First, the formats of the selected response and limited production tasks extend beyond a typical multiple-choice format and, hence, provide wide ranges of sampling of grammatical ability and enhance the generalizability of its score-based interpretation (see Purpura 2013). Second, after considering other existing grammar tests, the researcher was convinced that this section of *FCE* was one of the best tests to assess students' grammatical knowledge in both forms and meaning, thereby following Purpura's framework. Purpura (2004) also specifically points out that the tasks in the Use of English section in the exam measure test takers' knowledge with regard to lexical, morphosyntactic and cohesive form and meaning. Furthermore, the tasks in the Use of English section extend to a discourse level, which requires students to use their grammatical knowledge meaningfully and appropriately to produce the correct answers.

Results

Analytical procedures

The IBM SPSS Program Version 21 was used to check for missing data, to compute descriptive statistics, and to conduct reliability analyses by estimating Cronbach's Alpha (a) coefficients (for the questionnaires). The test and questionnaire data sets were inputted to the EQS 6.2 program (Bentler 1985–2015) for confirmatory factor analyses (CFAs) and structural equation modeling (SEM) analyses. The test and questionnaire data were matched for each test taker and were checked for missing and incomplete data prior to the descriptive analysis.

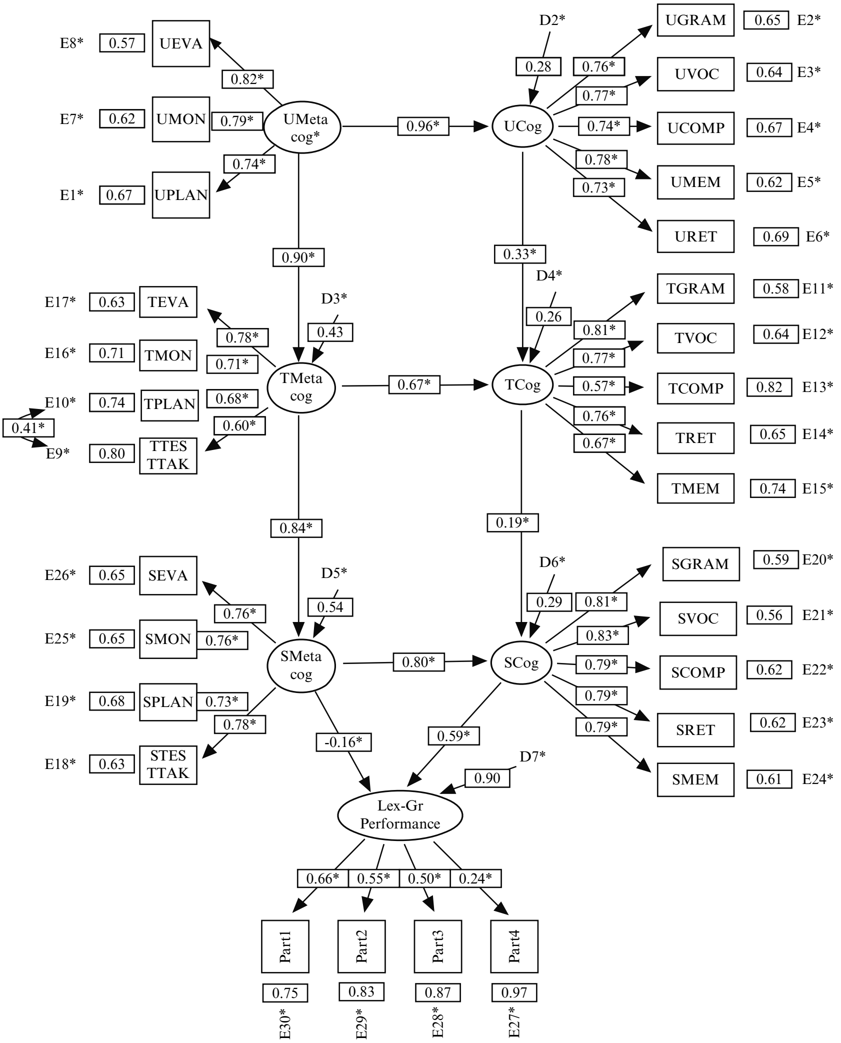
With the data distribution, normality, and internal consistency estimates of the lexico-grammar test and questionnaire data sets having been checked, a structural model to address the research questions was established (see Figure 11.1). After examining the relationship between observed and latent variables in its measurement models, which identify how well strategic processing variables can be measured by each reported strategy use variable, the study went on to explore the relationships between underlying constructs by means of latent structure analyses. These analyses were done by testing a number of hypothesized models to check their statistical and substantive plausibility. In this chapter, I present one structural model, discuss the findings, and answer the relevant research questions based on the statistical results. The model (see Figure 11.1) provides an adequate fit to the data (Chi-square (c^2 (388)) = 1169.99, $p = 0.00$, Comparative fit index (CFI) = 0.97, Root Mean Square Error of Approximation (RMSEA) = 0.06 (90% of the Confidence Interval (CI) = 0.06, 0.07), supporting the relationship in the model between strategic processing variables and lexico-grammar test performance variables. All parameter estimates were significant at the 0.05 level ($p < 0.05$). It should be noted that in principle, the chi-square statistic should be non-significant ($p > 0.05$) to indicate good model fit. However, it is well documented that this statistic is sensitive to sample size, and CFI and RMSEA have been developed to address this limitation.

Research Question 1

Relationships among metacognitive strategic processing strategies

The components of the model in Figure 11.1 consist of measures of L2 test takers' perceptions of their general cognitive and metacognitive strategy use, as well as their perceived trait and state cognitive and metacognitive strategy use in test taking. The full latent model indicates that general metacognitive language use strategic processing has a direct positive impact on trait metacognitive strategic processing and an indirect effect on state metacognitive strategic processing in test taking. From the model, the regression coefficient for the path from general learner use metacognitive strategic processing (UMetacog) to trait metacognitive strategic processing (TMetacog) was 0.90 ($R^2 = 0.85$; large effect size [ES]), and the path from TMetacog to state metacognitive strategic processing (SMetacog) was 0.84 ($R^2 = 0.77$; large ES). The results suggest that UMetacog in language use has an executive function over other strategic processes and has a positive indirect impact on SMetacog (0.90 x 0.84; $R^2 = 0.58$; large ES). The regression coefficients show that relationships among the three metacognitive strategies were all positive, direct and/or indirect. The results indicated that individuals who perceived they extensively employed

Figure 11.1 The structural relationship between strategic processing and lexico-grammar test performance (N = 416)



Chi-square (χ^2 (388)) = 1169.99, P = 0.00, CFI = 0.97, RMSEA = 0.06 (90% CI = 0.06, 0.07)

Note: U = Learner use; T = Trait; S = State; Meta = Metacognitive; Cog = Cognitive; GRAM = Grammar strategy; VOC = Vocabulary strategy; COMP = Comprehending strategy; RET = Retrieval strategy; MEM = Memory strategy; EVA = Evaluating strategy; PLAN = Planning strategy; MON = Monitoring strategy; TESTTAK = Test-taking strategy; Lex-Gr Performance = Lexico-grammatical performance metacognitive strategy use in general grammar and vocabulary situations also reported a high use of metacognitive strategy use in an actual lexico-grammar test situation. Although large portions of relevant variance among the three inter-correlated facets of strategic processing are shared across contexts, the limited extent to which each of the three types of metacognitive strategic processing strategies – general use, trait, and state metacognitive – can account for the other two factors indicates that they appear to be different in nature.

Relationships among cognitive strategic processing strategies

The model proposed in Figure 11.1 also related cognitive strategic processing of general use, trait, and state together and found that the regression coefficient from general use cognitive strategic processing (UCog) to trait cognitive strategic processing (TCog) was 0.33 ($R^2 = 0.11$; small ES), and that from TCog to state cognitive strategic processing (SCog) was 0.19 ($R^2 = 0.03$; small ES). This indicated that general cognitive language use and trait cognitive strategic processes do not have executive functions over state cognitive strategic processing. This result was consistent with Phakiti's (2008a) study, which sees cognitive strategies as context specific. Therefore, based on this study, it can be inferred that language learners' general cognitive strategy use does not have much to do with their cognitive strategy use in specific language use. For instance, learners' perceptions of cognitive strategy use in the lexico-grammar test seems not to be greatly influenced by general cognitive strategy use in general grammar-learning contexts. Therefore, it would be methodologically flawed to assume that from what test takers think they are doing one could infer what they actually do.

The relationship between metacognitive and cognitive strategic processing

The model in Figure 11.1 suggests that the regression coefficient for the pathway from UMetacog to UCog was 0.96 ($R^2 = 0.92$; large effect size [ES]), that from TMetacog to TCog was 0.67 ($R^2 = 0.45$; large ES) and that from SMetacog to SCOG was 0.80 ($R^2 = 0.64$; large ES). The model indicated that UMetacog explained 92% of learner use of cognitive strategic processing. TMetacog explained 45% of TCog variance, and SMetacog explained 64% of SCog variance. This finding suggested that metacognitive strategic processing might regulate and control cognitive strategic processing in both general and specific language use situations.

On the basis of the model presented in Figure 11.1, the findings were consistent with previous studies, which regard metacognitive strategies as having a higher order executive function over other human information processing. Because there have so far been no studies looking at the relationships between language learners' perceptions of strategy use and the actual strategies employed by L2 learners in test-taking situations, this study, therefore, makes an important contribution. Based on the model it seems that we are able to distinguish language learner general strategy use from specific strategy use in test-taking situations. In addition, it can also be noted that L2 learners' general learner strategy use seems to have a strong influence on L2 learners' strategy use in specific situations, as for example in testing contexts. The model also explains that general learner metacognitive strategy use is a strong indicator of specific strategy use.

Research Question 2

The model in Figure 11.1 also helped us understand how online strategic processing (i.e. processes that are employed during the actual process of taking a test) are related to lexico-grammar test performance. State cognitive and metacognitive processing strategies do not seem to play the same role in test performance. This idea is supported by the results of SEM, which showed that state cognitive processing was directly related to lexico-grammar test performance. Overall, the direct and indirect influences of strategic processing were found to explain test takers' test performance.

Direct influences

Based on the model presented in Figure 11.1, it was found that state cognitive processing has a direct effect on lexico-grammar test performance ($a = 0.59$; $R^2 = 0.35$; medium ES). In other words, state cognitive processing explained 35% of lexico-grammatical performance variance in this model, suggesting that strategic processing is moderately related to lexico-grammar test performance. It is not surprising to see only a moderate influence of strategic processing on lexico-grammatical test performance because language test performance can be influenced by many other factors (Bachman and Palmer 2010). It seems that the performance of test takers on the lexico-grammar test largely relies on their lexico-grammar knowledge.

Indirect influences

In the model you see that a negative relationship ($b = -0.16$) was found between test takers' metacognitive processing and their lexico-grammar test performance. This result suggests that state metacognitive processing does not effectively help test takers' performance, but it is possible that they may influence test performance indirectly. This finding seems to provide evidence against the commonly held belief that the more metacognitive strategies test takers can employ the better their test performance will be. Although previous research does not provide us with clear empirical evidence suggesting the relationship between metacognitive strategic processes and test performance, in line with Purpura (1999) and Phakiti (2008a, 2008b), in this study an indirect relationship (b = 0.80 × 0.59) was found between state metacognitive strategy and test performance, indicating that metacognitive strategies alone do not influence test performance; rather the way this influence was accomplished was through metacognitive processing exerting an effect over cognitive processing, with both kinds of processing working together closely to impact on L2 test performance.

Discussion

Overall, the current study has provided some new insights into our understanding of strategic processing research in language testing. For instance, although the positive effect of online strategic processing on test performance has been documented in a few studies (e.g. Phakiti 2007, 2008a), the function of offline strategic processing on test performance has not been well addressed. In particular, there have been few studies that have attempted to look at the broad picture of how both offline (e.g. pre-task planning) and online strategic processes might influence test performance. In other words, the majority of previous investigations simply looked at one aspect of the influence, while the real nature of strategic competence has not been comprehensively explained. Phakiti's innovative approach to bringing trait and state conception together represents a way forward in strategic processing research. However, the findings of the present investigation suggest that apart from the contribution of trait and state strategic processes to our understanding of strategic competence, it seems that the general learner use strategic processing definitely needs to be distinguished from the other two processes (i.e. trait and state) and included in the construct of strategic competence.

Though Purpura (1999) realized that it is impossible to isolate strategic competence in testing from strategic competence in second language learning and use, his study only revealed part of the story (i.e. only learner strategies) of the influence of strategic competence on test performance. The current study helps us understand the precise nature of Purpura's suggestion that strategic competence in testing, learning and use work together to directly and indirectly affect L2 test performance.

Types of strategic processing in lexico-grammar tests

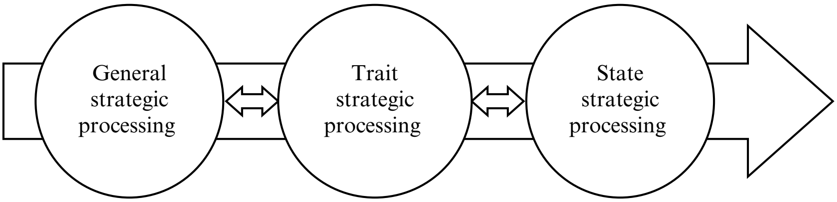
As language testing is a special case of L2 use based on certain given tasks, strategic processing during completion of a test should reflect L2 test takers' knowledge of strategy use in their long-term and working memory (Phakiti 2007). In other words, strategic processing research in language assessment should include the constructs of strategy use variables for both general and specific language use activities. For instance, findings in this study suggest that strategic processing in test taking may involve 1) test takers' general knowledge or awareness of strategy use in language use activities; 2) test takers' perceived knowledge of strategy use when taking a given test; and 3) actual strategy use in test taking.

Because it appears that the majority of previous studies (e.g. Purpura 1999, Song 2005, Song and Cheng 2006) did not differentiate between strategic processing in test taking and strategic processing in language use and acquisition, the current study highlighted the importance of identifying the differences between them. For example, findings from this study suggest that the three types of strategic processing identified from SEM (see Figure 11.1) are actually different in nature. Accordingly, to make more accurate inferences about L2 test takers' strategic processing, there is a need to make a distinction between different types of strategic processing that learners can employ in general L2 language learning and specific test-taking contexts.

Another question in strategic behavior research is how different kinds of strategic processing in different contexts are related to one another. Purpura (1999) suggested that strategic competence in test taking is related to strategic competence in second language use and acquisition. Some researchers (Bachman, Davison, Ryan and Choi 1995, Purpura 1999) argue that the effect of strategic processing on test performance is direct in certain circumstances, but may be indirect in other circumstances. It may be impossible to separate strategic competence in test taking from strategic competence in second language use and acquisition in general (Purpura 1999). It seems reasonable to posit that strategic processes used in general contexts and in specific contexts are closely related because the current study found empirical evidence that metacognitive strategic processes (i.e. general use, trait, and state) are closely related to one another. Hence, the empirical evidence in the study suggested that it is necessary to recognize a close relationship between the different kinds of strategic processing, as well as to make clear distinctions between strategic behaviors in these three contexts, in order to make more precise inferences about their influences on test performance. For instance, cognitive processing in this study was found to be contextual and related to specific language use activities.

Although Phakiti (2007) successfully brings the concepts of state and trait strategic processing into test taking in his pioneering attempts to understand strategic competence in test taking, his studies may have not documented the full array of strategies that affect test takers' test performance. For example, Phakiti's studies did not consider test takers' strategic processing in general language use situations as an individual factor that has influence on test performance. Hence, studies like the current one looking at strategy use in language testing at a macro level (i.e. in both general language use and specific testing contexts) are needed to provide a more comprehensive picture of the strategic processing construct variables. The results from the current study revealed that the degree of conscious employment of the three kinds of strategic processes in test taking can be considered as residing on a continuum (see Figure 11.2). Under test conditions, test takers' conscious selection of strategic behaviors might be in a sequential and interactive relationship (Cohen 2011). In other words, test takers usually begin their strategic thinking from general awareness in their long-term memory and then call on highly focused strategic processes to deal with specific test tasks. In Figure 11.2, the double arrows between each variable denote that these three types of strategic processing are closely related. However, there is also a sequential direction in test takers' strategic behaviors. That is, a test taker's general awareness of what they normally do in general language use activities informs both what they think they do and what they actually do, in specific situations.

Figure 11.2 Sequential and interactive relationships among strategic processes in language testing


General strategic processing
Trait strategic processing
State strategic processing

Because the current study has shown that the three types of macro-level strategic processing do exist in test takers' strategic activities during test taking, the next question that needs to be asked is how they influence test performance. Nevertheless, little empirical evidence has so far been documented to reveal the nature of strategic competence or to help model the relationships between strategic competence and language test performance. In other words, the understanding of strategic competence and its effect on language test performance is still very much based on theoretical hypothesizing. In particular, research has so far been unable to properly explain: 1) the different facets of strategic competence; 2) the nature of strategic competence based on the three different types of observable strategic processes; and 3) possible shared variance between strategic processing and test performance. The current study has shed further light on these questions.

Relationships between strategic processing and lexico-grammar test performance

The model employed in the current study (see Figure 11.1) is consistent with the theoretical model proposed by Bachman and Palmer (2010), which postulates that metacognitive strategies directly regulate cognitive strategies, and cognitive strategies are directly related to test takers' language test performance. In metacognition theory, strategic awareness is similar to metacognitive awareness. In fact, Schraw and Moshman (1995) posit the following: 1) knowledge of cognition is considered static judgment because the process requires individuals to assess their knowledge or ability in hypothetical situations; 2) regulation of cognition is metacognition in action as it enables an individual to orchestrate different mental processes during problem solving; and 3) general awareness or knowledge regulates the processes for specific contexts.

The model that I propose in this study also considered test takers' general strategic processing, and recognized the role that general offline strategic thinking could play in online situational performance. By looking at learners' perceptions of strategic processing in both general and specific contexts, its true nature might be better explained. Therefore, it is possible that the findings from this study may indicate that general strategic knowledge is directly related to online metacognitive processes, but only indirectly related to strategic cognitive processing via online metacognitive processes. The results show that general learner use processing and trait are strong indicators of test takers' other strategic processes.

In the current study, a close relationship was found between metacognitive processing and cognitive processing, with the regression coefficient ranging from 0.67 to 0.96. The high regression coefficients in these studies between metacognitive strategies and cognitive strategies can be interpreted as they have been in many previous studies (e.g. O'Malley and Chamot 1990, Phakiti 2003, 2008b, Purpura 1999, Zhang and Zhang 2013). That is to say, the current study provided further empirical evidence that test takers' metacognitive strategic processing performs an executive function over their cognitive strategic processing. Zhang, Goh and Kunnan (2014) point out that in the research on strategic competence in language testing, researchers have focused on investigating metacognitive strategies, as they are regarded as the focal attributes of strategic competence. The component of cognitive strategies newly included in the theoretical model by Bachman and Palmer (2010) has not been well researched. Zhang et al (2014) particularly suggest that further empirical research of how metacognitive and cognitive strategies are related is needed in validating strategic competence theory. The current study has provided further empirical evidence to address the above research issues.

Conclusions and implications for further research

This study followed the literature in suggesting that there has been a lack of empirical research aiming to validate Bachman and Palmer's (2010) strategic competence theory, and that there is a need to understand L2 learners' less known strategic processing in language abilities, such as lexico-grammar. This study also aimed to investigate the complex relations between different strategy use constructs in general language use contexts and language use in specific situations, such as test taking. In line with Phakiti (2007, 2008a, 2008b), this study classified strategy variables into trait and state types. The study also added general learner use strategy to the constructs of possible strategy use affecting test performance and found that general learner strategy use has a direct and positive influence on trait strategy use in test taking, and an indirect and positive effect on strategy use in the context of a specific language use activity. Unlike previous investigations, the present study first distinguished and modeled learner strategy use outside of a specific context and then modeled strategy use in test taking. The results found that there is a strong relationship between the variables of general learner use strategy and testing strategy use. Generally, metacognitive strategy use was shown to act in an executive role to control and regulate cognitive strategy use, which once again offered empirical evidence to support the theory that metacognitive strategy use can be considered to consist of higher order strategies that provided a control or management function in language use (Bachman and Palmer 2010).

Given that the main purpose of strategy research is to look at how the use of strategies can facilitate language learning and use, the relationships between strategic competence, strategy use, and language test performance were the ultimate concern of this study. The present study suggests that general strategy use and state strategy use are positively, directly and indirectly related to language test performance. The findings suggested that there is only a moderate relationship between the two variables and this finding is still consistent with the theory of Bachman and Palmer's general model of communicative language ability, and more specifically with the framework of factors affecting grammar test performance proposed by Purpura (2004).

Finally, in terms of the limitations of this study, it is important to note that a one-time correlational research study can only provide a snapshot into the complexities of strategy use in L2 language use and assessment. The effect of the questionnaire questions and the validity of the responses from the test takers are a significant concern for the present study. Apart from not being able to capture the full range of test takers' strategic behaviors, due to some overlap in the taxonomies of strategy use items, students might tend to overgeneralize their answers in general and in test-taking situations, thus questioning both the reliability and validity of their answers. The study results would be stronger if other research techniques were to be used to understand strategy use in both general language use and under assessment conditions. Therefore, follow-up qualitative or longitudinal quantitative investigations are needed to provide more insightful additions to our knowledge. Also, as discussed earlier, both test takers' language test performance and strategic abilities are highly complex, multi-dimensional, and variable across contexts and specific situations. Therefore, a longitudinal investigation, which looks at the stability of strategy use, could provide us with more information about the complexity of L2 learners' mental processing.

References

Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bachman, L F (2000) Modern language testing at the turn of the century: Assuring that what we count counts, *Language Testing* 17 (1), 1–42.

Bachman, L F (2014) Ongoing challenges in language assessment, in Kunnan, A J (Ed) *The Companion to Language Assessment III*, Oxford: John Wiley & Sons Inc, 1,586–1,604.

Bachman, L F and Palmer, A S (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford: Oxford University Press.

Bachman, L F and Palmer, A S (2010) *Language Testing in Practice*, Oxford: Oxford University Press.

Bachman, L F, Davison, F, Ryan, K and Choi, I-C (1995) *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge–TOEFL Comparability Study*, Studies in Language Testing volume 1, Cambridge: UCLES/Cambridge University Press.

Bentler, P M (1985–2015) *EQS Version 6.2 for Windows*, Encino: Multivariate Software.

Cohen, A D (1987) The use of verbal and imagery mnemonics in second language learning, *Studies in Second Language Acquisition* 9 (1), 43–61.

Cohen, A D (2011) *Strategies in Learning and Using a Second Language* (Second edition), Harlow: Pearson Education.

Cohen, A D (2014) Using test-wiseness strategy research in task development, in Kunnan, A J (Ed) *The Companion to Language Assessment II*, Boston: Wiley Blackwell, 893–905.

Macaro, E (2006) Strategies for language learning and for language use: Revising the theoretical framework, *The Modern Language Journal* 90 (3), 320–337.

McNamara, T F (1996) *Measuring Second Language Performance*, London: Longman.

Ministry of Education (2004) *Full-time High School English Syllabus* (Trial revised edition), Beijing: People's Education Press.

National College English Testing Committee (2006) *CET-4 Test Syllabus and Sample Papers*, Shanghai: Shanghai Foreign Language Education Press.

O'Malley, J M and Chamot, A U (1990) *Learning Strategies in Second Language Acquisition*, Cambridge: Cambridge University Press.

O'Sullivan, B and Weir, C J (2011) Test development and validation, in O'Sullivan, B (Ed) *Language Testing: Theories and Practices*, Basingstoke: Palgrave Macmillan, 13–32.

Oxford, R L (2011) *Teaching and Researching Language Learning Strategies*, Harlow: Pearson Education.

Phakiti, A (2003) A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance, *Language Testing* 20 (1), 26–56.

Phakiti, A (2007) *Strategic Competence and EFL Reading Test Performance: A Structural Equation Modelling Approach*, Frankfurt: Peter Lang.

Phakiti, A (2008a) Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests, *Language Testing* 25 (2), 237–272.

Phakiti, A (2008b) Strategic competence as a four-order factor model: A structural equation modeling approach, *Language Assessment Quarterly* 5 (1), 20–42.

Purpura, J E (1999) *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modelling Approach*, Cambridge: Cambridge University Press.

Purpura, J E (2004) *Assessing Grammar*, Cambridge: Cambridge University Press.

Purpura, J E (2008) Assessing communicative language ability: Models and their components, in Hornberger, N H (Ed) *Encyclopedia of Language Education*, New York: Springer, 2,198–2,213.

Purpura, J E (2013) Assessment of grammar, in Chapelle, C A (Ed) *The Encyclopedia of Applied Linguistics*, Malden: Blackwell Publishing, 195–204.

Schraw, G and Moshman, D (1995) Metacognitive theories, *Educational Psychology Review* 7 (4), 351–371.

Song, X (2005) Language learner strategy use and English proficiency on the Michigan English Language Assessment Battery, *Spaan Fellow Working Papers in Second or Foreign Language Assessment* 3, 1–26.

Song, X and Cheng, L (2006) Language learner strategy use and test performance of Chinese learners of English. *Language Assessment Quarterly* 3 (3), 243–246.

Swain, M, Huang, L-S, Barkaoui, K, Brooks, L and Lapkin, S (2009) The speaking section of the TOEFL iBT™ (SSTiBT): Test-takers' reported strategic behaviors, *TOEFL iBT ™ Research Report*, available online: onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2009.tb02187.x/pdf

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.

Zhang, L and Zhang, L J (2013) Relationship between Chinese college test takers' strategy use and EFL reading test performance: A structural equation modeling approach, *RELC Journal* 44 (1), 35–37.

Zhang, L, Goh, C C M and Kunnan, A J (2014) Analysis of test takers' metacognitive and cognitive strategy use and EFL reading test performance: A multiple-sample SEM approach, *Language Assessment Quarterly* 11 (1), 76–102.

Section 4
Epilogue

12 Epilogue: A teacher educator's personal and professional perspective

Kathleen M Bailey

Middlebury Institute of International Studies at Monterey (MIIS), US

As a language teacher, test developer, and teacher educator, I am fascinated by the chapters in this volume because they raise a number of interesting issues for me. I am particularly happy to see these studies in print, because for several years I have been offering a seminar on language assessment for MA candidates in language teaching. In this epilogue, I will comment on each chapter in terms of what it offers me as a teacher educator, and also put the various topics discussed here into their historical contexts.

In the mid-1990s, J D Brown and I conducted a survey of language testing specialists in order to 'investigate the existing teacher training courses in language testing so that we [could] eventually understand how the ideologies, value systems, and normative practices of language testing . . . are informing language teaching and vice versa' (Bailey and Brown 1996:238). The questionnaires elicited information on areas covered in teacher training courses on language assessment, including course structure, general topics covered, item writing, item analysis, test consistency, and test validity.

In 2007, that study was replicated (Brown and Bailey 2008) with 20 new Likert-scale items added to the questionnaire. The survey included Likert scale and open-ended response items, and there were 76 Likert scale items that appeared on both the 1995 version and the 2007 version of the questionnaire. In the second study, Brown and Bailey reported that:

> . . . the Pearson product-moment correlation coefficient between the means on the 76 common Likert-scale items in 1996 and the [2007] study was .961 and r^2 was .924. Thus 92.4% of the variation in means on the 1996 questionnaire was shared by the variation in means on those same items in the [later] study, indicating, at least for the Likert scale items, remarkable stability in relative ratings assigned by teachers of language testing over more than a decade (Brown and Bailey 2008:365).

Furthermore, 'the similarities in responses to items included in both the old and new version of the questionnaire indicate the presence of a stable knowledge base that is evolving and expanding, rather than shifting radically' (Brown and Bailey 2008:371). In spite of the apparent similarities in the two data sets, however, we should note that 20 new Likert scale items were added to the survey in 2007 because of developments in the field. Those new items addressed the following topics:

> . . . test analysis, washback, test bias, testing in relationship to curriculum, standards (cut-point) setting, critical approaches to language testing, language program evaluation, classroom testing practices, Rasch analysis, computer-based TOEFL$_{®}$ (CBT) scores, internet-based TOEFL$_{®}$ (IBT) scores, generalizability theory, consequential validity, values implications in validity, multiple regression, structural equation modeling, analysis of variance (ANOVA), many-faceted Rasch (FACETS) analysis, and validity as a unitary concept (Brown and Bailey 2008:370).

I share these survey results with you here to provide the context for my remarks about the research reported in this volume, which covers many of these same topics – washback, classroom testing practices, internet-based scoring practices, structural equation modeling, consequential validity, critical approaches to language testing, and so forth. My comments represent the perspective of someone who trains teachers about language assessment rather than someone who works primarily as a test developer or an assessment researcher.

Section 1: Assessing the productive skills in post-secondary contexts

The unifying theme of the first section of this volume has long been a primary focus for language testers. Because many of my graduates go on to teach languages and/or direct programs in post-secondary contexts, the chapters in this section were of great interest to me. The chapters in this section by di Gennaro, Zhao, and Dikli all address issues of written discourse in academic contexts, while those by Cheng and Hsieh focus on spoken language.

In Chapter 1, di Gennaro reports on her investigation of the performance of international students writing in English (their second language, L2) with that of L2 writers who had attended at least secondary school in the USA. The analysis examined the writers' control over five different components of writing ability: grammatical, cohesive, rhetorical, sociopragmatic, and content control. The results showed that the international L2 writers outscored the resident L2 writers on all five of the components listed above. For teachers of writing courses in which both sorts of students enroll, the question arises as to how best to help various types of L2 learners improve their academic writing. The author concludes:

> Unless studies can show that resident L2 learners' writing improves more when they are placed into composition courses for native English L1 writers than when they are placed into composition courses for L2 students, diminishing resident learners' L2 status seems ingenuous at best and irresponsible at worst, especially in cases where composition courses have been designed with L2 learners' strengths and weaknesses in mind (see section on 'Implications and discussion', di Gennaro, this volume).

In Chapter 2, Zhao addresses writing in post-secondary contexts, but with a focus on voice in argumentative writing by L2 writers. She investigates a scoring rubric based on Hyland's (2008) model of authorial voice in academic writing, in order to determine the rubric's reliability and validity. Zhao uses a definition of *voice* in writing as 'the expression of the essential individuality of a particular writer' (Stewart 1992:283).

Hyland's (2008) model of voice consists of two main categories: stance and engagement. *Stance* is further subcategorized as consisting of hedges, boosters, attitude markers, and self-mention. *Engagement* consists of reader mention, personal asides, knowledge reference, directives, and questions. The rubric investigated in this study consisted of the 10 components of Hyland's (2008) model listed above, with the addition of a category for central point articulation and another for rating the overall voice. Zhao's findings include her quantitative results (based on confirmatory factor analysis and regression analysis) and interesting quotes from the raters (derived from think-aloud protocols and interviews about what influenced their ratings). The author notes that 'a rater's perception or feeling of overall voice strength in a writing sample is less related to how many times individual voice elements are used but more related to how they are used' (see section on 'Phase 2: Validation of the revised voice rubric', Zhao, this volume).

This chapter provides much food for thought for my teacher trainees and for me as an instructor of a language assessment course. Both the construction and the validation processes described here provide examples of careful rubric development. This investigation of voice intrigues me because it is such an elusive construct. Many years ago, J D Brown and I investigated the use of an analytic scoring system for assessing compositions written by upper intermediate university English as a Second Language (ESL) students (Brown and Bailey 1984). The five categories in that rating instrument were 1) organization (the introduction, body, and conclusion), 2) the logical development of ideas (a focus on content), 3) grammar, 4) punctuation, mechanics, and spelling, and 5) style and quality of expression.

This last category was the least well defined of all the categories on the analytic scoring rubric. In fact, the explanation of the style component was only one-third to one-half the length of the other categories. And yet – ironically – when we correlated the scores on the various categories with the total points awarded, we found that the criterion of style and quality of expression correlated most highly with total scores. The Pearson's correlation coefficient was 0.91, which means the r^2 value was 0.8281. In other words, the ratings on this least well defined criterion contributed about 83% of the variance in the total scores. For this reason, I am pleased to read about Zhao's research on voice because it offers my teacher trainees and me specific practical ways of discussing this important component of style with language learners.

In Chapter 3, Dikli takes a case study approach to investigating how two ESL writers incorporated feedback from an automated essay scoring (AES) system into their academic writing. Two other ESL writers received teacher feedback on their drafts. Dikli examined how the two types of input were used as these learners produced multiple drafts of their work. Dikli also looked at whether there were differences between the two pairs of students (those who got teacher feedback and those who received AES input) in terms of how they utilized their respective types of feedback in their revisions. Finally, Dikli elicited the students' perceptions of the two types of feedback.

Of course, anyone who has ever taught a composition course, or even a content-based course or a four-skills course with a substantial writing component, knows the heavy workload associated with reading and commenting on students' written work, so it has long been hoped that the use of technology might spare teachers some of the work involved in giving feedback while providing improved consistency due to the absence of human rater biases. In addition, using technological tools to do the rating in large-scale testing contexts was thought to be a way to increase the practicality of scoring direct tests of writing. However, not everyone is sanguine about the use of machine scoring. In 2004, the Conference on College Composition and Communication (CCCC, or 'The Four Cs') issued the following statement about machine scoring:

> Writing-to-a-machine violates the essentially social nature of writing: We write to others for social purposes. If a student's first writing experience at an institution is writing to a machine, for instance, this sends a message: Writing at this institution is not valued as human communication – and this in turn reduces the validity of the assessment (Weigle 2013:87).

As Dikli notes, in spite of such concerns, AES is widely used in testing organizations, higher education, and school systems for large-scale high-stakes assessments, not only because machine scoring is practical but also because the scores typically correlate well with those assigned by human raters.

In Dikli's study, essays written by the four learners were evaluated using five categories of an analytic scoring system: focus and meaning, content and development, organization, language use and style, and mechanics and conventions. The L2 writers who received AES input got more feedback about form and meaning than did the L2 writers who received teacher feedback. There were also substantial differences in terms of the teacher feedback and the AES feedback on the category of language use and style, with the AES input being much greater than the teacher feedback on this category. The AES feedback on mechanics and style was also greater than the teacher feedback. However, even though the AES gave more feedback than the teacher did, the author concluded that only the teacher 'provided feedback based on the individual needs of each student for the particular draft on which they were working' (see the section 'Discussion', Dikli, this volume). Thus, although it is important for my graduate students to learn about machine scoring and how it may influence their own students' test preparation, it is also important for teacher trainees to be able to give appropriate tailored individual feedback on learners' work at any given point in the writing process.

Chapter 4 reports on Cheng's investigation of requests spoken by L2 learners of English. She used the concepts of power, distance, and rank of imposition difference between interlocutors, as well as the learners' proficiency (high or low) and context (ESL versus English as a Foreign Language (EFL)) as variables in her research. Four discourse completion tasks were computer delivered as a semi-direct test: 1) borrowing a pen from a friend, 2) requesting a job interview at what might be an inconvenient time for the potential employer, 3) asking one's brother for the TV remote control, and 4) asking a professor if one could take an exam a day later than when it is scheduled. One dependent variable was *response latency*, which was defined as 'the lapse in time between stimulus and response, which was normally the gap between when the audio prompt ended . . . and when the participant pushed the "record" button' (see the section 'Data analyses', Cheng, this volume). Cheng found that regardless of the speakers' context or proficiency level, the response latency for the tasks regarding rescheduling the job interview and postponing the exam were consistently longer than for the tasks about the pen and the remote control. Furthermore, the speech rate was slower for the interview and exam tasks, regardless of the EFL/ESL context or the learners' proficiency level. Cheng's analysis of variance results showed that there was a large, statistically significant, main effect for both speech rate and response latency in terms of the discourse features of the tasks. There were also statistically significant differences in response latency and speech rate between the high and low proficiency learners. However, there were no significant differences between learners in the EFL and ESL situations – a surprising result, because it is often thought that pragmatic competence can develop more readily in a second language (compared to a foreign language) context.

I would definitely want my teacher trainees to read Cheng's study because pragmatic competence is an important part of communicating appropriately in a second language. Making an inappropriate request is a speech act that can easily cause offense, so language teachers must be skilled in ways to help learners develop their awareness and abilities in this area. Many years ago, the assessment of request-making in an L2 was investigated by Farhady (1980) using an indirect multiple-choice format. Cheng's use of discourse completion tasks and computer recordings of learners' utterances brings this important focus into the 21st century.

In Chapter 5, Hsieh discusses an assessment issue that has been a matter of concern since the early 1980s: the oral proficiency of international teaching assistants (ITAs) (see Bailey 1983, 1985, Bailey, Pialorsi and Zukowski-Faust (Eds) 1984). Teaching assistants in US universities are typically graduate students who either supervise laboratory sessions (e.g. in physics, chemistry, biology, etc.) or lead discussion sessions for students following professors' lectures. However, the term *teaching assistant* can be rather misleading because in some contexts graduate students function as the primary instructor, in courses such as basic mathematics, foreign languages, composition, and ESL.

ITAs are graduate students whose native language is not the language of instruction – typically English in the available research reports. Hsieh's research specifically investigates undergraduates' reactions to the ITAs – a topic which has received attention for at least three decades. (For related studies, please see Briggs and Hofer 1991, Hinofotis and Bailey 1981, Plakans 1997.) Hsieh's study focuses specifically on undergraduate students' perceptions of ITAs' accentedness, comprehensibility and oral proficiency.

I find this chapter valuable, not only because it addresses one of my personal research interests, but also because some of my graduates have gone on to do ITA training and assessment. In addition, this issue continues to resurface because ITAs move on as they complete their graduate studies and new ITAs take their places. Turnover is thus inherent in recruiting and training ITAs, and as a result, there is an ongoing need for appropriate assessment procedures.

In comparing the ratings given by undergraduates and ESL teachers to ITAs' speech, the author found that the undergraduates were 'significantly more severe in their ratings of accentedness and comprehensibility than the ESL teachers' (see the section 'FACETS analyses', Hsieh, this volume), but that the ITAs' phonology was the issue most frequently commented on by both groups of raters. For ESL teachers who work with ITAs, it is important to understand what the undergraduate students think about their non-native speaking instructors and to be able to provide appropriate language improvement and support strategies as needed.

Section 2: Assessing young learners in school contexts

The three chapters in this section report on research in a burgeoning area of language testing – that of assessing young language learners. Recent publications attest to the importance of this topic (see, e.g. Coombe and Davidson 2012, Hasselgreen 2005, Johnstone 2000, McKay 2005, 2006, Rea-Dickins 2000a, 2000b). This trend may be a result of various countries lowering the age at which foreign languages are taught in schools.

In Chapter 6, Clark-Gareca explores young second language test takers' views about the accommodations ESL children experienced when taking content area tests. Using work by Acosta, Rivera and Shafer-Willner (2008:vii), Clark-Gareca defines *accommodations* for English language learners as 'changes to testing procedures, testing materials, or the testing situation in order to allow students' meaningful participation in the assessment' (see the section 'Literature review', Clark-Gareca, this volume). The information about students' views of such accommodations reminded me of dynamic assessment – an approach which 'challenges conventional views on teaching and assessment by arguing that these should not be seen as separate activities but should instead be fully integrated' (Poehner 2010:5). In dynamic assessment, teachers and students interact, and teachers intentionally provide the test takers with the scaffolding they need in order to accomplish the test tasks. In traditional assessment procedures, this support would be seen as inappropriate because

> . . . interacting with students during a test, providing feedback on performance before test-takers have finished, and modifying the test administration procedure for individual learners are usually considered unfair because the resulting score no longer represents a learner's solo performance (Poehner 2010:12).

However, Poehner notes that 'the provision of such assistance simultaneously aids development, and so assessment itself becomes an instructional intervention' (2010:5).

In dynamic assessment, testing procedures are used to help students learn, rather than solely to measure how much they have learned up to the moment of testing. In dynamic assessment, 'children are given hints or training to enable them to show individual differences in progress made during the process of solving a variety of cognitive tasks' (Elliot, Grigorenko and Resing 2010:220). My question is whether the kinds of accommodations Clark-Gareca has described in this chapter would fall under the umbrella of dynamic assessment.

In Chapter 7, Smith reports on her investigations of the knowledge of multi-word phrases (MWPs) among 7- to 10-year-old school children who are L2 speakers of English in the UK. She used a multi-word task that included transparent, semi-transparent, and non-transparent phrases (e.g. 'break a bone', 'break the silence', and 'break the ice', respectively). The children completed sentence prompts in a game-like context. The results showed that the children responded more accurately and more frequently to the transparent items than to the semi-transparent items, as well as to the semi-transparent items compared to the non-transparent items. Both of these comparisons revealed statistically significant differences.

Smith's results are important for me, because some of my teacher trainees do intend to work with young language learners, so Smith's findings about transparency and MWPs are likely to be relevant to them as future primary school teachers. In terms of classroom assessment, I will encourage my graduate students when they are assessing vocabulary knowledge to examine MWPs as well as individual lexical items. In addition, I hope they will develop creative ways to help children acquire semi-transparent and non-transparent MWPs to support the learners in the development of reading (and perhaps listening) comprehension skills. As Smith notes: 'It is likely that there is a reciprocal relationship between reading and MWP knowledge; learners who read more might encounter greater numbers of MWPs' (see the section 'Implications and further research', Smith, this volume).

In Chapter 8, Kimberly K Woo reports on the results of her survey research about how teachers of young EFL learners view social language and its assessment, in contrast with the assessment of academic language. Her main concern was to determine how kindergarten and first-grade teachers in New York's Chinatown define and assess the construct of social language. Four of the 30 survey respondents were also interviewed. The respondents used the descriptors of *conversational* and *interactive* most often to describe social language. The teachers who were interviewed noted that appropriacy was a key feature of social language and that younger learners had more difficulty with this issue than older learners. The survey items and interview questions asked teachers how they assessed social language and how often such assessments were used. The teachers reported using (in descending order of frequency) in-lesson observations, individual conferences, students' classroom presentations (including sharing), and classroom tests as their four most frequently employed assessment procedures. In general, the assessment of social language was reported to be quite limited, since teachers viewed it as happening largely during play, lunch, snack time, and recess, when they themselves often were not present. Having insufficient time to observe and assess students' social language emerged as a prevalent theme in the teachers' comments.

Woo's study was motivated in part by social language tasks on high-stakes English tests. She concludes her chapter by saying that 'to help bridge the disconnection, there is a need for both tests and teachers to acknowledge and address areas of overlap' between social and academic language (see the section 'Discussion', Woo, this volume). To me this point seems like very good advice for elementary school teachers-in-training. Further research is needed on this topic – including classroom observation research – to help teachers develop effective and efficient means of assessing both children's academic and social language skills.

Section 3: Language assessment concerns in local contexts

The third section of this volume addresses language assessment concerns in local contexts, an issue that was recently added to The International Research Foundation for English Language Education (TIRF) research priorities on language assessment. The three chapters by Darbes, Wu, and Bi represent different contexts but all contribute to our understanding of assessment issues in important ways.

The study by Darbes (Chapter 9) investigates test takers' perceptions of test validity in a community college environment and discusses the issue of construct irrelevant variation. Relatively little research has been conducted in community college contexts, compared to the plethora of studies conducted in 4-year colleges, universities, and elementary school contexts. (For exceptions see Schuemann 2009, and Bailey and Santos (Eds) 2009.) Darbes takes the notion of causal thought as an investigatory concept. That is, she has tried to understand the linkages between student-level variables, such as motivation and self-regulation, and test-based academic outcomes. She relates the concept of causal thoughts to the broader issue of test taker perceptions, which include their test response strategies and how they react when taking tests. They documented the causal thoughts students used in discussing their exam results, and considered possible social and psychological outcomes for these learners in pursuit of their academic goals.

Darbes found that the immigrant students didn't question the validity of the community college tests. In their interview data, only two out of the 42 students complained about the test. The vast majority of the students' causal thoughts were about the students themselves. The author notes that the students exhibited 'implicit trust in test results' (see the section 'Experiences of testing', Darbes, this volume). For those students who did not pass the exam, the idea that the knowledge being tested was not fresh was a common explanation. Some students reported that they had not taken the test very seriously. Others referred to personal factors, such as anxiety or lack of focus. Almost half of the students reported experiencing anxiety due to the high-stakes nature of the test(s) they were facing. Some reported being unclear about the genre they were to produce. For instance, when faced with instructions to write an essay, one student said he wasn't even sure what an essay was.

This chapter evoked a strong reaction from me, because many of my graduates go on to work in community colleges in the USA and Canada. As a result, it is vital that they understand language assessment issues in these contexts. It is quite likely that my former trainees will be responsible for preparing their ESL students to take these sorts of tests, and they may be responsible for administering (or even developing) such tests and interpreting the scores.

The Chapters by Wu and Bi both report on research conducted with university EFL students in China. This context is important internationally due to the huge number of language learners involved and the recent developments in the assessment of such learners (see Cheng and Curtis 2010), but also because so many of my students wind up teaching English in China.

Washback is the main topic of Wu's report in Chapter 10. She examines the apparent effects of the National College English Test Band 4 (CET-4) in three different science- and engineering-oriented universities of different rankings in Shanghai. University A was the highest ranked, University B the second ranked, and University C the lowest ranked in this study. Wu wanted to know if the washback from the CET-4 would differ across these three university contexts, and if so, why. Data were collected via a survey (N = 414) and follow-up interviews with 34 students. Wu notes that at University A the main reason students gave for learning English was that they wanted to increase their ability to communicate in English. In contrast, the need to improve their CET-4 scores was the main reason given by students at University B and University C.

Wu found that the perceptions of washback were strongest among students at University C and weakest at University A. Wu discusses students' perceptions of various test preparation and learning activities, including fast reading, intensive reading, and taking mock examinations for practice. Overall she found that the CET-4 exerted greater washback on the lower ranked universities than at the higher ranked university in her study. She related the findings to the learners' goals (e.g. students at University A were more likely to want to travel in English-speaking countries than were those at University B or University C). She points out that students recruited by these various universities come into higher education contexts with differing levels of English ability to start with. In addition, in the higher-ranked university in this study there were more English facilities and resources available to learners than at the other two universities.

I would want my teacher trainees to read Wu's chapter because many of them go to China to teach English, including those who teach university English courses through the US Peace Corps. The importance of language assessment there, and the use of the CET-4 in particular, cannot be underestimated. Wu's chapter should be helpful to any English teacher who plans to work in higher education in China.

In Chapter 11, Nick Zhiwei Bi reports on his research on strategic processing in test-taking contexts. Drawing on previous research, he investigated learners' self-reports of their perceptions regarding 1) their general learning strategies, 2) strategies they thought they used in carrying out specific tasks (called *trait strategies*), and 3) their actual use of such strategies immediately after completing a task (known as *state strategies*).

Bi's data collection procedures involved having students complete two questionnaires a week before taking a test of lexico-grammatical knowledge (a retired version of *First Certificate in English* (*FCE*; now known as *Cambridge English: First*)). They then completed a third questionnaire immediately after taking the test, which had four types of items: cloze, gap-filling, word formation, and sentence transformation. The questionnaires asked the learners about their cognitive and metacognitive strategy use.

Among other findings, the results of Bi's research suggest that learners' reported use of general strategies strongly influences their reported use of strategies in specific contexts, such as language testing. In other words, 'state cognitive processing has a direct effect on lexico-grammar test performance' (see the section 'Direct influences', Bi, this volume). However, he cautions readers that what language learners say they do when taking tests may or may not be what they actually do.

As a teacher educator, what I gather from Bi's chapter that will be helpful for my teacher trainees is the importance of helping language learners develop and deploy their strategic competence, in both test and non-test situations. In addition, it is not uncommon for graduates to be assigned to teach test preparation courses in language programs. Helping learners develop their test-taking strategies can sometimes seem just as important – particularly in high-stakes assessment contexts – as helping them improve their target language skills. Therefore, I would want my graduates to be knowledgeable and confident in this regard.

Concluding comments

The importance of research on language assessment cannot be ignored by any organization concerned with language teaching and learning, and/or the preparation of language teachers. In various contexts around the world, high-stakes language tests influence decisions that affect students' primary and secondary opportunities for further education. In many countries, language test results are used in corporate contexts to determine which employees might be promoted, provided with further training opportunities, and/or posted overseas. In addition, test results are used to judge teacher effectiveness and to evaluate programs. Finally, washback – generally speaking, the effects of a test on teaching and/or learning – can influence lesson planning and execution, curriculum design, and materials development.

The chapters in this volume have addressed many important issues in language assessment, and have raised important questions about several others (Duff and Bailey 2001). If you would like to access reference lists about such issues, please visit www.tirfonline.org/resources/references. There you will find free downloadable Word documents with citations on the following topics, many of which have been addressed in these chapters: anxiety, automated essay scoring, authentic assessment, community colleges and ESL, dynamic assessment, immigrant issues, oral proficiency interviews, raters and rating scales, reading assessment, self-assessment, speaking assessment, technology in language assessment, validity and validation, vocabulary, voice in writing, washback, writing assessment, and assessing young learners. I hope these resources will be helpful to you and to your students if you are teaching courses on language assessment.

References

Acosta, B, Rivera, C and Shafer-Willner, L (2008) *Best Practices in State Assessment Policies for Accommodating English Language Learners: A Delphi Study*, available online: files.eric.ed.gov/fulltext/ED539759.pdf

Bailey, K M (1983) Foreign teaching assistants at US universities: Problems in interaction and communication, *TESOL Quarterly* 17 (2), 308–310.

Bailey, K M (1985) If I had known then what I know now – performance testing of foreign teaching assistants, in Hauptman, P C, LeBlanc, R and Wesche, M B (Eds) *Second Language Performance Testing: Le Testing de Performance en Language Seconde*, Ottawa: University of Ottawa, 153–180.

Bailey, K M and Brown, J D (1996) Language testing courses: What are they? in Cummings, A and Berwick, R (Eds) *Validation in Language Testing*, Clevedon: Multilingual Matters, 236–256.

Bailey, K M and Santos, M G (Eds) (2009) *Research on English as a Second Language in US Community Colleges: People, Programs and Potential*, Ann Arbor: University of Michigan Press.

Bailey, K M, Pialorsi, F and Zukowski-Faust, J (Eds) (1984) *Foreign Teaching Assistants in US Universities*, Washington, DC: National Association for Foreign Student Affairs.

Briggs, S and Hofer, B (1991) Undergraduate perceptions of ITA effectiveness, in Nyquist, J D, Abbott, R D, Wulff, D H and Sprague, J (Eds) *Preparing the Professionals of Tomorrow to Teach*, Dubuque: Kendall-Hunt, 435–445.

Brown, J D and Bailey, K M (1984) A categorical instrument for scoring second language writing skills, *Language Learning* 34 (4), 21–42.

Brown, J D and Bailey, K M (2008) Language testing courses: What are they in 2007? *Language Testing* 25 (3), 348–383.

Cheng, L and Curtis, A (2010) *English Language Assessment and the Chinese Learner*, New York: Routledge.

Coombe, C and Davidson, P (2012) Assessing young language learners: Issues, principles and practices, in Emery, H and Gardiner-Hyland, F (Eds) *Contextualizing EFL for Young Learners: International Perspectives on Policy, Practice and Procedure*, Dubai: TESOL Arabia, 283–296.

Duff, P A and Bailey, K M (2001) Identifying research priorities: Themes and directions for the TESOL International Research Foundation (guest editor and contributor), *TESOL Quarterly* 35 (4), 595–616.

Elliot, J G, Grigorenko, E L and Resing, W C M (2010) Dynamic assessment, in Peterson, P, Baker, E and McGaw, B (Eds) *International Encyclopedia of Education*, Oxford: Elsevier, 220–225.

Farhady, H (1980) *Justification, development, and validation of functional language testing*, unpublished doctoral dissertation in applied linguistics, University of California, Los Angeles.

Hasselgreen, A (2005) Assessing the language of young learners, *Language Testing* 22 (3), 337–354.

Hinofotis, F B and Bailey, K M (1981) American undergraduates' reactions to the communication skills of foreign teaching assistants, in Fisher, J C, Clark, M A and Schachter, J (Eds) *On TESOL '80, Building Bridges: Research and Practice in Teaching English as a Second Language*, Washington, DC: TESOL, 120–136.

Hyland, K (2008) Disciplinary voices: Interactions in research writing, *English Text Construction* 1 (1), 5–22.

Johnstone, R (2000) Context-sensitive assessment of modern languages in primary (elementary) and secondary education: Scotland and the European experience, *Language Testing* 17 (2), 123–143.

McKay, P (2005) Research into the assessment of school-age language learners, *Annual Review of Applied Linguistics* 25, 243–263.

McKay, P (2006) *Assessing Young Language Learners*, New York: Cambridge University Press.

Plakans, B S (1997) Undergraduates' experiences with and attitudes toward international teaching assistants, *TESOL Quarterly* 31 (1), 95–118.

Poehner, M (2010) *Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting Second Language Development*, Berlin: Springer.

Rea-Dickins, P (2000a) Assessment in early years language learning contexts, *Language Testing* 17 (2), 115–122.

Rea-Dickins, P (2000b) Current research and professional practice: Reports of work in progress in the assessment of young language learners, *Language Testing* 17 (2), 245–249.

Schuemann, C (2009) Access to freshman composition at stake: Comparing student performance on two measures of writing, in Bailey, K M and Santos, M G (Eds) *Research on English as a Second Language in US Community Colleges: People, Programs and Potential*, Ann Arbor: University of Michigan Press, 170–185.

Stewart, D C (1992) Cognitive psychologists, social constructionists, and three nineteenth-century advocates of authentic voice, *Journal of Advanced Composition* 12 (2), 279–290.

Weigle, S C (2013) English language learners and automated scoring of essays: Critical considerations, *Assessing Writing* 18 (1), 85–99.