Read the latest journal articles and reports on the TOEFL® test.


# TABLE OF CONTENTS

Welcome Message From The ILTA President 01  
Samuel J. Messick Memorial Lecture Award 03  
Cambridge/ILTA Lifetime Achievement Award 05  
Sponsors 07  
Exhibitors 08  
Conference Support 09  
ILTA Board and Officers 2012 10  
Selection Committees 2012 10  
Award Winners 2012 11  
Conference Overview 13  
Conference Schedule 14  
Workshops 26  
Paper Abstracts 29  
Symposium Abstracts 48  
Poster Abstracts 60  
Works-In-Progress Abstracts 73  
Index 85

LTRC 2012

ASSESSMENT AND LEARNING: BRIDGING DIVERSE DISCIPLINES AND DOMAINS
Studies in Language Testing

The essential reference for professionals and researchers in language testing and assessment

*Studies in Language Testing* (SiLT) is a series of academic volumes focusing on important issues and new developments in language testing.

They are published collaboratively by two departments of the University of Cambridge: Cambridge ESOL and Cambridge University Press.

SiLT profiles work on topics such as:

- the assessment of writing, reading, speaking and listening
- language testing for educational and professional purposes
- traditional and computer-based testing
- test impact.

The volumes form a comprehensive resource for anyone with a professional or research interest in language learning, teaching or assessment.

For more information on the Studies in Language Testing series visit: [www.CambridgeESOL.org/silt](http://www.CambridgeESOL.org/silt)
To order copies from Cambridge University Press visit: [www.Cambridge.org](http://www.Cambridge.org)
Welcome Message from the ILTA President

In the second year of my term as President of the International Language Testing Association, I am very pleased again to welcome you to our 34th annual Language Testing Research Colloquium in Princeton, New Jersey.

The Colloquium was last held in Princeton in 1991, which seems a long time ago to those of us who attended that meeting. At the time the conference could comfortably fit into the Chauncey Conference Center on the campus of the Educational Testing Service (ETS). It is a measure of how LTRC has grown that this time around we are at a larger venue on the other side of town, the Hyatt Regency Princeton.

With growth there often come growing pains. In the case of LTRC, one kind of growing pain has expressed itself in an ongoing debate about how to structure the conference program. The relatively small group of language testers who founded LTRC and sustained it through the early years in the 1980s highly valued the intimacy of the meetings, in which all the papers were presented to the whole group and were then subject to vigorous debate on the merits of the work. For graduate students whose papers were accepted, it could be both a harrowing and a highly gratifying experience to have their work come under the critical scrutiny of a number of the top scholars in the field.

As the number of participants has increased over the years, it has become more difficult to maintain the same kind of intimacy. Many regular LTRC goers still have a strong preference for plenary paper sessions, but that comes at the cost of restricting the number of papers that can be accepted within a three-day program. Thus, some recent LTRC organizing committees have introduced parallel paper sessions to varying degrees, although a survey of colloquium participants in 2009 showed a strong division of opinion on the desirability of this development. For this year’s colloquium, the conference program chairs have sought – in consultation with the ILTA officers – to achieve a good balance in the overall program structure, with a judicious number of parallel sessions as part of the mix. We welcome your feedback on how well it works this time.

Partly as a trade-off for the limited number of paper presentations, the LTRC program has long given prominence to the poster and work-in-progress sessions. I often go to conferences where posters are assigned to an out-of-the-way location in the lunch hour where they are easy to forget about, but the poster sessions at LTRC have always been highly visible and built into the main conference schedule. Similarly, the work-in-progress sessions have frequently been a highlight for me, both as a presenter and a participant, by providing focused discussion of research issues in small groups and valuable feedback on studies that have yet to be completed.

Another symptom of growth has been the introduction in recent years of a Newcomers’ Session before the main colloquium program begins. We are particularly pleased to greet those who are joining us for the first time and want you to make the most of the experience. Graduate students are in a real sense our future, and we have a special welcome for you. The ILTA Executive Board and the conference chairs look forward to meeting you all at the session, which is scheduled on Monday at 5.10 pm, just before the Opening Reception.
The last LTRC in Princeton in 1991 was significant as the venue for an organizational meeting which led to the formal establishment of the International Language Testing Association (ILTA) the following year. Our Association has also grown since then and achieved a milestone in recording a core membership of 300 in January, covering the regular, student, institutional and lifetime categories. These numbers are augmented by those who belong to the 12 regional affiliates around the world. In the last year we have been happy to accept two new affiliates, the Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) and The Canadian Association of Language Assessment /L'association canadienne pour l'évaluation des langues (CALA/ACEL). They both represent models of good organization and commitment to the promotion of high standards in testing and assessment. More generally, there is scope for more regular communication between ILTA and the affiliates, as well as the fostering of links between the affiliates themselves.

In ILTA our activities have both an inward and an outward focus. Within the field we attach great importance to the recognition of excellence in research and professional practice through the range of awards that we give (with Cambridge ESOL, Educational Testing Service, and Sage Publications as valued sponsors) for outstanding publications, presentations and overall achievement. At the Gala Banquet on Thursday evening, we will be honoring the winners of these awards, as well as the recipients of thesis and dissertation awards from Educational Testing Service, Cambridge ESOL, and The International Research Foundation for English Language Education.

We also seek an outward focus, to encourage better understanding of best practice in language testing and assessment beyond our own community. We have resumed our joint colloquia at the annual conference of the American Association for Applied Linguistics, and we contribute an invited lecture to the program of the TESOL International convention. We provide grants each year for workshops and meetings in parts of the world where knowledge and expertise in testing and assessment are lacking. And, within our limited resources, we are exploring other innovative ways of promoting the concept of assessment literacy among language teachers, educational administrators, policymakers and other test users. We will be reporting on our work at the Annual Business Meeting of ILTA, to be held on Thursday at 12.15 pm. I urge all ILTA members to attend, and I encourage others to take up ILTA membership as a way to lend support to our broader objectives.

Finally, I want to express on behalf of all of us our great appreciation to the conference chair Xiaoming Xi, and the program co-chairs Rob Schoonen, Steve Ross, and Yasuyo Sawaki, along with all the other people with roles in the conference organization, for the endless hours of effort they have put into making this conference a success. Only those who have been at the center of the planning for a meeting like this know how demanding a task it is and we can have nothing but admiration for their dedication and their sheer hard work.

John Read
President, International Language Testing Association, 2011-12
Values, computer technology and validation in language assessment

Recent approaches to validation build upon Messick’s seminal work to provide praxis-oriented concepts and tools for planning, conducting, and interpreting validation research (e.g., Kane, 2006). Language testers (e.g., McNamara & Roever, 2006; Bachman & Palmer, 2010) embrace such practice-oriented approaches, but point out the need to retain Messick’s (1989) message that validation should include an analysis of the value implications of assessment practice: “Given the pervasiveness and the subtlety of the impact of values and ideologies on test interpretation, we need to explore some ways and means of uncovering tacit value premises and of coping with their consequences for test validation” (p. 62). McNamara (2006) saw Messick’s assertion that “all test constructs embody values” as “challenging in the most creative sense” (p. 46).

This lecture takes up the challenge of uncovering tacit value premises intertwined within validation practices for language assessments that are delivered and scored using computer technology. I begin with a brief review of Messick’s stance on values in validity, which ties values to construct definition. I argue that because definition of constructs in language assessment often entails consideration of contexts of language use beyond the test setting, analysis of values also entails such contexts. Focusing on one aspect of language use contexts—computer technologies for information and communication—I explore how validation is affected by language testers’ value premises about the role of technology in contexts of language use and in language assessment tasks. I draw upon reported validation projects and my own experience to examine the effects on validation of value-laden assumptions regarding normal communication practices, relevant qualities of test-takers’ language performance, and benefits of technology for language test design and delivery. With the aim of incorporating such an analysis into argument-based validation practices, I suggest how discovery of value-laden assumptions about technology might inform development of validity arguments.

Biography

Fifty years as Japan’s leading English-language assessment provider.

www.eiken.or.jp
On behalf of the committee for the 2012 Cambridge/ILTA Lifetime Achievement Award (consisting of John Read, Jo Lewkowicz, Hanan Khalifa, and myself), it is a great pleasure to announce that, after carefully considering the several highly meritorious nominations received from the field, the committee has selected Professor Carol Chapelle to receive the Lifetime Achievement Award to be presented at the 34th Language Testing Research Colloquium, which will be held in Princeton, USA in April 2012.

Professor Carol Chapelle is a Distinguished Professor in TESL/Applied Linguistics in the Department of English of Iowa State University. Throughout her career she has undertaken a program of research and publications in language testing which has been deeply embedded in a wider range of interests in applied linguistics and TESOL. She has made notable contributions in two major areas of the field. The first is in the use of computer technology for language testing, growing originally out of her experience with computer-assisted language learning (CALL) as an ESL teacher. Since then, she has been a leader in investigating the potential of the technology to enhance language assessment, while at the same time maintaining a critical perspective by acknowledging problem areas and challenges. She is the author (with Dan Douglas) of Assessing Language Through Computer Technology (Cambridge, 2006), a comprehensive survey of the area. More broadly, Carol’s work on computer-based assessment should be viewed as a key component of her primary interest in issues at the intersection of computer technology and applied linguistics, as reflected in her books Computer Applications in Second Language Acquisition (Cambridge, 2001) and English Language Learning and Technology (Benjamins, 2003).

The second area in which Carol has made outstanding contributions is the construct validation of language tests. Through a series of very influential papers in the late 1990s and early 2000s, she explored how modern validity theory could be applied in the analysis and development of language tests, particularly but not exclusively those designed to assess vocabulary knowledge and ability. At the same time she was deeply involved in building the conceptual framework for what has become the internet-based Test of English as a Foreign Language (iBT). This was the basis for a sophisticated validity argument, as presented in the volume for which she was first co-editor and a prominent author, Building a Validity Argument for the Test of English as a Foreign Language (Routledge, 2008) – a book described by Alister Cumming in a Language Testing review as a “monumental achievement”.

Apart from her theoretical contributions, Carol was co-director with Joan Jamieson of the project that led to the Longman English Assessment and has developed innovative language tests for her own institution. She was also co-author of ESOL Tests and Testing: A Resource for Teachers and Administrators (TESOL, 2005), a noteworthy initiative to promote assessment literacy among the target readership.

Carol has served on the Executive Board of ILTA and been a frequent presenter at LTRC, as well as an active member of the Midwest affiliate MwALT. She won the ILTA Best Article Award in 1998 for her book chapter “Construct definition and validity inquiry in SLA research”. She has also served the wider field as Editor of TESOL Quarterly and President of AAAL – roles which have brought her to international prominence and provided opportunities to communicate her work on language testing and related areas to a much broader audience. Her current major project is the multi-volume Wiley-Blackwell Encyclopedia of Applied Linguistics, for which she is not only General Editor but also editor of the Assessment volume.

Because of her many significant contributions over the years, the committee is delighted to select Professor Carol Chapelle to receive the 2012 Cambridge/ILTA Lifetime Achievement Award.

Elana Shohamy
Chair, 2012 Cambridge/ILTA Lifetime Achievement Award Committee
LRN is recognised as an awarding body by Ofqual, the qualification regulator in England.

Learning Resource Network is a flexible, secure and reliable awarding body with approved test centres around the world that offers a diverse range of examinations for speakers of English worldwide.

Currently on offer:
Skills for Life – speaking & Listening modules
Entry 1 – Entry 3.

Currently under development:
Financial English, Business English, ESOL International and awards to cater for the workplace, international immigration entrants, university entrants, exams for schools.

Become a registered test centre
before May 2012 FREE
6 working day turnaround on certificates

contact us to find out how to become approved centre

0870 6258 408 enquiries@lrnglobal.org
CONFERENCE SPONSORS

ETS

The creator of the TOEFL® and TOEIC® Tests

Opening Reception
Pre-Conference Workshop 3
ETS TOEFL and TOEIC
www.ets.org

UNIVERSITY of CAMBRIDGE
ESOL Examinations

Gala Banquet
University of Cambridge
ESOL Examinations
www.cambridgeesol.org

PEARSON

Refreshment Breaks
(April 3 & April 5)
Pearson Tests of English
www.pearsonpte.com

ETC TOEIC

Conference Bag
ETS TOEIC
www.ets.org/toeic

EIKEN

Student Registration Fee Subsidies
The Society for Testing English Proficiency, Inc.
www.eiken.or.jp

WiDA

Refreshment Breaks
(April 4)
WIDA
www.wida.us

Learning Resource Network

Pre-Conference Workshop Refreshments
Learning Resource Network
www.lrnglobal.org

2LT SECOND LANGUAGE TESTING INC.

A Berlitz Company

Summer Internship Session
Second Language Testing, Inc.
www.2LTI.com
CONFERENCE EXHIBITORS

ETS

The creator of the TOEFL® and TOEIC® Tests
ETS TOEFL and TOEIC
www.ets.org

UNIVERSITY of CAMBRIDGE
ESOL Examinations
University of Cambridge ESOL Examinations
www.cambridgeesol.org

Routledge

Routledge
Taylor & Francis Group
www.routledge.com

Cambridge University Press
Cambridge University Press
http://www.cambridge.org/us/esl/

Cambridge Michigan Language Assessments
www.cambridgemichigan.org
Organizing Committee

Chairs
Steve Ross, University of Maryland
Rob Schoonen, University of Amsterdam
Xiaoming Xi, Educational Testing Service

Conference Program
Steve Ross, University of Maryland
Yasuyo Sawaki, Waseda University
Rob Schoonen, University of Amsterdam
Xiaoming Xi, Educational Testing Service

Conference Website
Yeonsuk Cho, Educational Testing Service
Ellie Patounas, Educational Testing Service

Logistics
Marcel Ionescu, Educational Testing Service
Alexis Lopez, Educational Testing Service
Spiros Papageorgiou, Educational Testing Service
April Shaw, Educational Testing Service

Pre-Conference Workshops
Marcel Ionescu, Educational Testing Service
Rob Schoonen, University of Amsterdam
April Shaw, Educational Testing Service
Youngsoo So, Educational Testing Service
Mikyung Kim Wolf, Educational Testing Service

Professional and Social Events
Yeonsuk Cho, Educational Testing Service
Marcel Ionescu, Educational Testing Service
Steve Ross, University of Maryland
April Shaw, Educational Testing Service
Cynthia Tucci, Temple University

Program Book Production
Ian Blood, Educational Testing Service
Lin Gu, Educational Testing Service
Sara Kirsich, Educational Testing Service
Spiros Papageorgiou, Educational Testing Service
Kurt Schmenger, Educational Testing Service
April Shaw, Educational Testing Service
Xiaomei Song (student volunteer), Queen’s University
Sultan Turkan, Educational Testing Service
Erik Voss (student volunteer), Iowa State University

Sponsorship
Yeonsuk Cho, Educational Testing Service
Yasuyo Sawaki, Waseda University

Student Events
Meg Malone, Center for Applied Linguistics
Rebecca Present-Thomas (student volunteer), University of Amsterdam
Jonathan Schmidgall, University of California at Los Angeles
Sultan Turkan, Educational Testing Service

Abstract Reviewers
(in alphabetic order)
Lyle Bachman
Alison Bailey
Beverly Baker
Jayanti Banerjee
David Beglar
Vivien Berry
Annie Brown
Yuko Butler
Micheline Chalhoub-Deville
Carol Chapelle
Yeonsuk Cho
Martyn Clark
Alister Cumming
Fred Davidson
John de Jong
Barbara Dobson
Dan Douglas
Glenn Fulcher
Ardeshir Geranpayeh
April Ginther
Liz Hamp-Lyons
Ari Huhta
Talia Isaacs
Eunice Jang
Dorry Kenyon
Ute Knoch
Antony Kunnan
Constant Leung
Lorena Llosa
Sari Luoma
Meg Malone
Tim McNamara
Fumiyo Nakatsuhara
John Norris
Gary Ockey
Spiros Papageorgiou
Lia Plakans
David Qian
John Read
Dan Reed
Nick Saville
Elana Shohamy
Toshihiko Shoitsu
Charles Stansfield
Lynda Taylor
Randy Thrasher
Carolyn Turner
Elvis Wagner
Yoshinori Watanabe
Sara Weigle
Gillian Wigglesworth
Mikyung Kim Wolf
Guoxing Yu
ILTA Board and Officers

ILTA Executive Board 2012

President
John Read
University of Auckland

Vice-President
Gillian Wigglesworth
The University of Melbourne

Secretary
Margaret (Meg) Malone
Center for Applied Linguistics

Treasurer
Alistair van Moere
Pearson

Immediate Past President
Carolyn Turner
McGill University

ILTA Members at Large 2012

Ute Knoch
The University of Melbourne

Lorena Llosa
New York University

Spiros Papageorgiou
Educational Testing Service

Yasuyo Sawaki
Waseda University

ILTA Nominating Committee 2012

Chair
India Plough
Michigan State University

Kirby Grabowski
Teachers College, Columbia University

Rachel Brooks
Federal Bureau of Investigation

Tineke Brunfaut
Lancaster University

2012 Selection Committees

Cambridge/ILTA Lifetime Achievement Award

Chair
Elana Shohamy
Tel Aviv University

Hanan Khalifa
University of Cambridge ESOL Examinations

Jo Lewkowicz
University of Warsaw

John Read
University of Auckland

Committees cont’d.

Sage/ILTA Book Award

Chair
Annie Brown
Australian Council for Educational Research

Alistair Cumming
University of Toronto

Craig Deville
Measurement Incorporated

Meg Malone
Center for Applied Linguistics

Lia Plakans
University of Iowa

Best Article Award

Chair
Ari Huhta
University of Jyväskylä

Nathan Carr
California State University, Fullerton

Fred Davidson
University of Illinois at Urbana-Champaign

Barbara Dobson
Cambridge Michigan Language Assessments

Talia Isaacs
University of Bristol

Robert Lado Memorial Award for Best Graduate Student Paper

Chair
Yan Jin
Shanghai Jiao Tong University

Jessica Wu
Language Testing and Training Centre, Taipei

Ron Martinez
University of Nottingham

Gad Lim
University of Cambridge ESOL Examinations

ILTA Student Travel Awards

Chair
Carolyn Turner
McGill University

Lorena Llosa
New York University

Spiros Papageorgiou
Educational Testing Service

ILTA Workshops and Meetings

Chair
John Read
University of Auckland

Ute Knoch
The University of Melbourne

Lorena Llosa
New York University
AWARD WINNERS 2012

Sage/ILTA Book Award

ILTA Best Article Award

ILTA Student Travel Awards
Winner
John Pill, The University of Melbourne
Using health professionals’ feedback commentary to inform the validation of an ESP rating scale

Runners-up
(alphabetical order, receiving funding portions)
Ikkyu Choi, University of California, Los Angeles
Modeling the structure of passage-based tests: An application of a two-tier full information item factor analysis

Huei-Lien Hsu, University of Illinois at Urbana-Champaign
The impact of World Englishes on oral proficiency assessment: rater attitude, rating tendency and challenges

Jing Xu, Iowa State University
Collocations in learner speech: Accuracy, complexity and context

2012 Jacqueline Ross TOEFL Dissertation Award
Cecilia Guanfang Zhao, New York University
Dissertation Title: The Role of Voice in High-stakes Second Language Writing Assessment
Supervisor: Dr. Lorena Llosa

Caroline Clapham IELTS Masters Award
Kellie Frost, The University of Melbourne
Title: Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers’ oral performances
Supervisor: Dr. Cathie Elder

TIRF Doctoral Dissertation Awards
Larry Davis, University of Hawai‘i
Dissertation Title: Rater Cognition and Behavior in Speaking Assessment
Award Name: TIRF Doctoral Dissertation Grant
Supervisor: Dr. John Norris

Liang Guo, Georgia State University
Dissertation Title: Product and Process in TOEFL iBT Independent and Integrated Writing Tasks: A Validation Study
Award Name: TIRF Doctoral Dissertation Grant
Supervisor: Dr. Sara Weigle

Sara Ashley Smith, Oxford University
Dissertation Title: Exploring Differential Collocational Knowledge among Native and Nonnative English Speaking Children
Award Name: TIRF Doctoral Dissertation Grant
Supervisor: Dr. Victoria Murphy

Beth Clark Gareca, New York University
Dissertation Title: Assessing ELL Content Knowledge in the Mainstream Classroom
Award Name: TIRF Russell N. Campbell Doctoral Dissertation Award
Supervisor: Dr. Lorena Llosa
The Amsterdam Center for Language and Communication (ACLC) University of Amsterdam

is pleased to announce that it will host EUROSLA 23, the 23rd conference of the European Second Language Association from 28 to 31 August, 2013.

In due time, all relevant information will be posted at http://www.hum.uva.nl/EUROSLA23
CONFERENCE OVERVIEW

SUNDAY, APRIL 1
8:15–5:00  Conference Registration
9:00–5:00  Workshops

MONDAY, APRIL 2
8:15–6:30  Conference Registration
9:00–5:00  Workshops
12:00–5:00 ILTA Executive Board Session
1:00–5:00  ETS Campus and Operations Tour (Princeton and Ewing)
5:10–6:10 LTRC Newcomers Session
6:30–8:00  Opening Reception

TUESDAY, APRIL 3
8:15–6:00  Conference Registration
8:30–5:00  Exhibition
8:45–9:00  Opening Welcome
9:00–10:00 Samuel J. Messick Memorial Lecture Award
10:00–10:20 Break
10:20–12:00 Paper Session 1 (2 parallel)
12:00–1:15 Lunch Break
12:00–1:00 Professional Event 1: Summer Internship Opportunities
12:00–1:15 Language Assessment Quarterly Editorial Board Meeting
1:15–2:25 Poster Session
2:30–4:10 Symposia (2 Parallel)
4:10–4:30 Break
4:30–6:00 Paper Session 2 (Plenary)
6:00–7:00 Professional Event 2: Researchers’ Experiences in Academia and Government, Nonprofit, for-Profit Organizations

WEDNESDAY, APRIL 4
8:15–6:00  Conference Registration
8:30–5:00  Exhibition
8:45–10:15 Paper Session 3 (Plenary)
10:15–10:35 Break
10:35–12:15 Paper Session 4 (3 Parallel)
12:15–2:15 Lunch Break
12:15–2:15 ILTA Business Meeting
2:15–3:30 Works–In–Progress Session
3:30–5:10 Symposium (Plenary)
5:10–5:30 Break
5:30–6:30 Paper Session 5 (Plenary)
6:30–7:15 Professional Event 3: Research Grants from Different Organizations

THURSDAY, APRIL 5
8:15–6:00  Conference Registration
9:00–5:00  Exhibition
8:45–10:15 Paper Session 6 (Plenary)
10:15–10:35 Break
10:35–12:15 Paper Session 7 (3 Parallel)
12:15–1:30 Lunch Break
12:15–1:30 Language Testing Editorial Board Meeting
1:30–3:00 Paper Session 8 (Plenary)
3:00–3:30 Break/Group Photo
3:30–5:15 Symposium – Cambridge/ILTA Lifetime Achievement Award Event
5:15–5:30 Closing Comments
6:30–10:00 Banquet and Award Presentation at Mercer Oaks
**CONFERENCE SCHEDULE**

<table>
<thead>
<tr>
<th>Time</th>
<th>SUNDAY, APRIL 1</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:15–5:00</td>
<td>Conference Registration</td>
<td>Regency Foyer</td>
</tr>
<tr>
<td>9:00–5:00</td>
<td><strong>Assessing the Ability to Convey Semantic and Pragmatic Meanings in Language Assessments (Part 1)</strong>&lt;br&gt;James E. Purpura, Columbia University&lt;br&gt;Kirby C. Grabowski, Columbia University</td>
<td>Lexington</td>
</tr>
<tr>
<td>9:00–5:00</td>
<td><strong>Latent Growth Modeling for Language Testing Research</strong>&lt;br&gt;Gregory R. Hancock, University of Maryland</td>
<td>Baltimore</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>MONDAY, APRIL 2</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:15–6:30</td>
<td>Conference Registration</td>
<td>Regency Foyer</td>
</tr>
<tr>
<td>9:00–12:15</td>
<td><strong>Assessing the Ability to Convey Semantic and Pragmatic Meanings in Language Assessments (Part 2)</strong>&lt;br&gt;James E. Purpura, Columbia University&lt;br&gt;Kirby C. Grabowski, Columbia University</td>
<td>Lexington</td>
</tr>
<tr>
<td>9:00–5:00</td>
<td><strong>Scaling and Equating Test Scores</strong>&lt;br&gt;Samuel A. (&quot;Skip&quot;) Livingston, Educational Testing Service&lt;br&gt;Shuhong Li, Educational Testing Service</td>
<td>Baltimore</td>
</tr>
<tr>
<td>12:00–1:00</td>
<td><strong>LUNCH BREAK</strong></td>
<td></td>
</tr>
<tr>
<td>12:00–5:00</td>
<td>ILTA Executive Board Session</td>
<td>Cleveland</td>
</tr>
<tr>
<td>1:00–5:00</td>
<td>ETS Campus and Operations Tour <em>(Princeton and Ewing)</em></td>
<td>Bus from hotel</td>
</tr>
<tr>
<td>5:10–6:10</td>
<td>LTRC Newcomers Session</td>
<td>Cambridge</td>
</tr>
<tr>
<td>6:30–8:00</td>
<td>Opening Reception</td>
<td>Crystal Garden</td>
</tr>
</tbody>
</table>
## Conference Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>TUESDAY, APRIL 3</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:15–6:00</td>
<td>Conference Registration</td>
<td>Regency Foyer</td>
</tr>
<tr>
<td>8:30–5:00</td>
<td>Exhibition</td>
<td>Austin</td>
</tr>
<tr>
<td>8:45–9:00</td>
<td>Opening Welcome</td>
<td>Columbus–Savannah</td>
</tr>
</tbody>
</table>

### SAMUEL J. MESSICK MEMORIAL LECTURE AWARD, CAROL A. CHAPELLE

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00–9:10</td>
<td>Introduction, <strong>Dan Douglas</strong>, Iowa State University</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td>9:10–9:55</td>
<td>Lecture, <strong>Carol A. Chapelle</strong>, Iowa State University</td>
<td></td>
</tr>
<tr>
<td>9:55–10:00</td>
<td>Presentation of the Samuel J. Messick Memorial Lecture Award by <strong>Scott G. Paris</strong>, Vice President, Research, Educational Testing Service</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00–10:20</td>
<td>BREAK</td>
<td></td>
</tr>
</tbody>
</table>

### Paper Session 1 (2 Parallel)

#### Session Chair: Beverly Baker, McGill University

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:20–10:50</td>
<td><strong>The Role of Exposure and Proficiency in Learning and Testing Second Language Pragmatics</strong>&lt;br&gt;Carsten Roever, The University of Melbourne&lt;br&gt;Stanley Wang, The University of Melbourne&lt;br&gt;Steph Brophy, The University of Melbourne</td>
<td>Columbus–Savannah</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:55–11:25</td>
<td><strong>Assessing the Effectiveness of a DCT Pragmatics Test</strong>&lt;br&gt;<strong>Fred S. Tsutagawa</strong>, Seikei University</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:30–12:00</td>
<td><strong>Assessing Job-Related Language Performance Capability through Situational Role-Plays</strong>&lt;br&gt;<strong>Eric Surface</strong>, SWA Consulting Inc.&lt;br&gt;<strong>Aaron Watson</strong>, SWA Consulting Inc.</td>
<td></td>
</tr>
</tbody>
</table>

#### Session Chair: Gary Ockey, Iowa State University

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:20–10:50</td>
<td><strong>Bridging Frameworks for Assessment and Learning:</strong>&lt;br&gt;<strong>The ACTFL Guidelines and the CEFR</strong>&lt;br&gt;<strong>Erwin Tschirner</strong>, University of Leipzig</td>
<td>Lexington–Cambridge</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:30–12:00</td>
<td><strong>Modeling the Structure of Passage-Based Tests:</strong>&lt;br&gt;<strong>An Application of a Two-Tier Full Information Item Factor Analysis</strong>&lt;br&gt;<strong>Ikkyu Choi</strong>, University of California, Los Angeles</td>
<td>Lexington–Cambridge</td>
</tr>
</tbody>
</table>

### Lunch Break

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00–1:15</td>
<td>LUNCH BREAK</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00–1:00</td>
<td>Professional Event 1: Summer Internship Opportunities</td>
<td>Lexington–Cambridge</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00–1:15</td>
<td>Language Assessment Quarterly Editorial Board Meeting</td>
<td>Wilson</td>
</tr>
<tr>
<td>Time</td>
<td>TUESDAY, APRIL 3</td>
<td>Location</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>1:15–2:25</td>
<td>POSTER SESSION</td>
<td>Foyer/Baltimore</td>
</tr>
<tr>
<td></td>
<td><strong>Session Chairs:</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Meg Malone, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>John Pill, The University of Melbourne</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Development of a Computer-Based Formative Assessment Instrument for Nurses</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Using ESL in Quebec, Canada</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Beverly Baker, McGill University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Michel Laurier, Universite de Montreal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Carolyn Turner, McGill University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>May Tan, McGill University, Canada</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maria Lourdes Lira Gonzales, Universite de Montreal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ainsworth Judith, Universite de Montreal</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Development of an Assessment of High Frequency English Vocabulary</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>for Young English Language Learners</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lauren Artzi, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lindsey Massoud, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diane August, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Writing Assessment Literacy: A Survey of Teachers’ Knowledge, Beliefs, and</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Practices</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deborah Crusan, Wright State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lia Plakans, The University of Iowa</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Atta Gebril, The American University in Cairo</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>In-Service Training Course in Classroom-Based Language Assessment</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lyle F. Bachman, University of California, Los Angeles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Barbara Dambock, Academy for In-service Teacher and Management Training</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Development of a Spanish-English Language Proficiency Scale (SEPS)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ekaterina Smyk, Arizona State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maria Adelaida Restrepo, Arizona State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joanna Gorin, Arizona State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shelley Gray, Arizona State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Developing a Validity Argument for Score Meaning of a Computer-Based ESL</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Collocational Ability Test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Erik Voss, Iowa State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Investigating Factors Affecting Students’ Performance on an English Achievement</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bokyoung Cho, Korea Institute for Curriculum and Evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hun Jeon, Korea Institute for Curriculum and Evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Woojin Jung, Korea University</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>The Impact of the Number of Multiple-Choice, Test-Item Options on a Test’s</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Construct: Do Five Options Versus Four or Three Really Change Things?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hyesun Lee, University of Nebraska, Lincoln</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paula Winke, Michigan State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>How Do Raters Weigh Fluency, Linguistic Accuracy, and Pronunciation When</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rating L2 Speaking?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Klaartje Duijm, University of Amsterdam</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Measuring Speaking Ability Through Elicited and Spontaneous Speech</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>With Automated Scoring</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Troy Cox, Brigham Young University</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Revising the Written Test of a University-Level Summative Assessment</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>in Hong Kong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Felicia Fang, Hong Kong Polytechnic University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alan Urmsgton, Hong Kong Polytechnic University</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>“Faking” Lower Proficiency or “Just Coping” During Legal Procedures?</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Margaret van Naerssen, Immaculata University</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Deconstructing the Construct of Teachers’ Diagnostic Competence:</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Investigating Teachers’ English Language Learning Assessments Practices</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maryam Wagner, OISE/University of Toronto</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eunice Eunhee Jang, OISE/University of Toronto</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Assessing Language Proficiency for Air Traffic Controllers:</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The Brazilian Air Force Exam</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Natalia de Andrade Raymundo, Brazilian Air Force</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Patricia Tosqui Lucks, Brazilian Air Force</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paula Ribeiro Souza, Brazilian Air Force</td>
<td></td>
</tr>
</tbody>
</table>
### CONFERENCE SCHEDULE

<table>
<thead>
<tr>
<th>Time</th>
<th>TUESDAY, APRIL 3</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:15–2:25</td>
<td>POSTER SESSION cont’d</td>
<td>Foyer/Baltimore</td>
</tr>
<tr>
<td></td>
<td><strong>The EAQUALS CEFR-Based Assessment Special Interest Project</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glyn Jones, Pearson</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Alternate ACCESS for ELLs Assessment</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tatyana Vdovina, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jennifer Christenson, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Creating a Test of British English Collocation Knowledge for Young Learners</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sara Smith, University of Oxford</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Victoria Murphy, University of Oxford</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>The Cross-National Comparability of the Internal Structure of the TOEFL Junior Test</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mikyung Kim Wolf, Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jonathan Steinberg, Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Native Speaker Responses in Language Test Development</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>John de Jong, Pearson / VU University Amsterdam</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ying Zheng, Pearson</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Investigating the Validity of a Test of Academic Productive Grammar</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yoo-Ree Chung, Iowa State University</td>
<td></td>
</tr>
<tr>
<td><strong>SYMPOSIA (2 PARALLEL)</strong></td>
<td>Session Chair: Barbara Dobson, Cambridge Michigan Language Assessments</td>
<td></td>
</tr>
<tr>
<td>2:30–4:10</td>
<td>The Interplay Among Teaching, Assessment, and Placement of International Teaching Assistants</td>
<td>Lexington–Cambridge</td>
</tr>
<tr>
<td></td>
<td>Organizer: Elvis Wagner, Temple University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussant: Barbara Hoekje, Drexel University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Using an ITA Assessment to Provide Detailed Feedback on Performance: Implications for Learners, Teachers, and Validity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jonathan Schmidgall, University of California, Los Angeles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TOEFL iBT Speaking for ITA Certification: State of Practice and Outstanding Validation Questions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Timothy Farnsworth, Hunter College, City University of New York</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Investigating the “Interactive” in the Interactive Performance Test**: A Conversation Analysis Approach</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Christos Theodoropulos, University of Pennsylvania</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assessing the Classroom Pragmatic Competence of International Teaching Assistants</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elvis Wagner, Temple University</td>
<td></td>
</tr>
</tbody>
</table>
## CONFERENCE SCHEDULE

<table>
<thead>
<tr>
<th>Time</th>
<th>TUESDAY, APRIL 3</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>SYMPOSIA (2 PARALLEL) cont’d</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Session Chair: Yeonsuk Cho, Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td>2:30–4:10</td>
<td>Considerations in the Assessment of Young Learners</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td></td>
<td>Organizer: Yuko Goto Butler, University of Pennsylvania</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussant: Elana Shohamy, Tel Aviv University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Young EFL Learners’ Scores and Their Feedback on Listening Comprehension Tasks in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a Diagnostic Test Development Project</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marianne Nikolov, University of Pécs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Young Learners’ Perceptions of Their Assessment Experiences: Achievement, Anxiety</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and Pressure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>David Carless, University of Hong Kong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interactional Development in Oral Paired-Assessments Among Young Learners of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Foreign Language</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yuko Goto Butler, University of Pennsylvania</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Developmental Considerations in the Observational Assessment of Three-Year-Olds’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language and Early Literacy Skills</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alison L. Bailey, University of California, Los Angeles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Huan Wang, CTB/McGraw-Hill</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lorena Houston, CTB/McGraw-Hill</td>
<td></td>
</tr>
<tr>
<td>4:10–4:30</td>
<td><strong>BREAK</strong></td>
<td></td>
</tr>
<tr>
<td>4:30–5:00</td>
<td><strong>PAPER SESSION 2 (PLENARY)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Session Chair: John de Jong, Pearson / VU University Amsterdam</td>
<td></td>
</tr>
<tr>
<td>4:30–5:00</td>
<td>The Diagnosis and Treatment of Weaknesses in Reading and Writing in a Foreign</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Charles Alderson, Lancaster University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ari Huhta, Jyväskylä University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lea Nieminen, Jyväskylä University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Riikka Ullakonoja, Jyväskylä University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eeva-Leena Haapakangas, Jyväskylä University</td>
<td></td>
</tr>
<tr>
<td>5:00–5:30</td>
<td>Cognitive Diagnostic Assessment of Second Language Reading Ability for Instructional Purposes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ahyoung Alicia Kim, Seoul National University</td>
<td></td>
</tr>
<tr>
<td>5:30–6:00</td>
<td>Item Writer Perceptions of Reading Task Difficulty</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marita Härmälä, Jyväskylä University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reeta Neittaanmäki, Jyväskylä University</td>
<td></td>
</tr>
<tr>
<td>6:00–7:00</td>
<td>Professional Event 2: Researchers’ Experiences in Academia and Government,</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td></td>
<td>Nonprofit, for-Profit Organizations</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>WEDNESDAY, APRIL 4</td>
<td>Location</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>8:15–6:00</td>
<td>Conference Registration</td>
<td>Regency Foyer</td>
</tr>
<tr>
<td>8:30–5:00</td>
<td>Exhibition</td>
<td>Austin</td>
</tr>
<tr>
<td></td>
<td><strong>PAPER SESSION 3 (PLENARY)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Session Chair: Lorena Llosa, New York University</td>
<td></td>
</tr>
<tr>
<td>8:45–9:15</td>
<td>A Synthesis and Evaluation of Research into the Psychometric Properties of Automated Writing Evaluation Aek Phakiti, The University of Sydney</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td>9:15–9:45</td>
<td>Re-Creating Validity Frameworks for Diagnostically-Oriented AES Yong-Won Lee, Seoul National University</td>
<td></td>
</tr>
<tr>
<td>9:45–10:15</td>
<td>Holistic Scores of Automated Writing Evaluation: Use, Reactions, and Rater Reliability Volker Hegelheimer, Iowa State University Zhi Li, Iowa State University Stephanie Link, Iowa State University Hong Ma, Iowa State University Hyejin Yang, Iowa State University</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td>10:15–10:35</td>
<td><strong>BREAK</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>PAPER SESSION 4 (3 PARALLEL)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Session Chair: Elvis Wagner, Temple University</td>
<td></td>
</tr>
<tr>
<td>10:35–11:05</td>
<td>The Impact of World Englishes on Oral Proficiency Assessment: Rater Attitude, Rating Tendency and Challenges Huei-Lien Hsu, University of Illinois at Urbana-Champaign</td>
<td>Cambridge</td>
</tr>
<tr>
<td>11:10–11:40</td>
<td>Online Speaking Examiner Training: Does Test Length Matter? Lucy Chambers, University of Cambridge ESOL Examinations Gad Lim, University of Cambridge ESOL Examinations</td>
<td>Cambridge</td>
</tr>
<tr>
<td>11:45–12:15</td>
<td>Effects of Accent Familiarity and Language Teaching Experience on Raters’ Judgments Becky Huang, Educational Testing Service</td>
<td></td>
</tr>
</tbody>
</table>
## Conference Schedule

**WEDNESDAY, APRIL 4**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
</tr>
</thead>
</table>
| 10:35–11:05 | Using Oral Reading Performance Features to Predict Reading Comprehension Ability in Second Language Learners of Dutch  
Ryan Downey, Pearson  
Jo Fond Lam, CINOP  
Alistair Van Moere, Pearson | Lexington    |
| 11:10–11:40 | Predicting Proficiency Without Direct Assessment: Can Speaking Ratings Be Inferred from Listening and Reading Ratings?  
Aaron Watson, SWA Consulting Inc.  
Reanna Harman, SWA Consulting Inc.  
Eric Surface, SWA Consulting Inc.  
Jenn McGinnis, SWA Consulting Inc. | Lexington    |
| 11:45–12:15 | Validating a Word Associates Test in EFL Contexts  
Sumi Han, Northern Arizona University  
Yong-Won Lee, Seoul National University | Lexington    |
| 10:35–11:05 | Preparing for TOEFL iBT Speaking Tasks: Test Takers’ Experiences and Expectations  
Guoxing Yu, University of Bristol | Baltimore    |
| 11:10–11:40 | A View of Teacher and Student Interactions and Learning in Two Classroom Applications of L2-AFL  
Christian Colby-Kelly, McGill University | Baltimore    |
| 11:45–12:15 | Developing a Code of Practice for EFL Testing in China: A Data-Based Approach  
Jinsong Fan, Fudan University  
Yan Jin, Shanghai Jiao Tong University | Baltimore    |
| 12:15–2:15  | LUNCH BREAK                                                            | Zion          |
| 12:15–2:15  | ILTA Business Meeting                                                  | Zion          |
| 2:15–3:30   | WORKS-IN-PROGRESS SESSION                                              | Zion          |
|             |                                                                        |              |
| 2:15–3:30   |                                                                        | Zion          |

**Teacher Assessment Practices for ELLs in the Content Areas**  
Beth Clark-Gareca, New York University

**Defining Oral Social Language in the Assessment of English Proficiency in Kindergarten-Age ELLs**  
Kimberly Woo, New York University

**Using Student-Directed Assessment as a Learning Process**  
Anna Fletcher, Charles Darwin University

**Tense and Aspect Avoidance in Testing Procedures: The Case of Elicited and Guided Measures**  
Donna West, State University of New York at Cortland
<table>
<thead>
<tr>
<th>Time</th>
<th>WEDNESDAY, APRIL 4</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:15–3:30</td>
<td>WORKS-IN-PROGRESS SESSION cont’d</td>
<td>Baltimore–Lexington–Cambridge</td>
</tr>
<tr>
<td></td>
<td>Investigating Effective Strategies For Teaching</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assessment Literacy to Instructors of Less</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Commonly Taught Languages</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aileen Bach, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anne Donovan, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Developing a Computer-Mediated Spanish Academic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language Assessment for Young Bilinguals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>David MacGregor, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stephanie Gibson, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dorry Kenyon, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Investigating Student Growth Percentiles as a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Measure of ELL Instructional Effectiveness for</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teacher Performance Evaluations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amy Yamashiro, Alexandria City Public Schools</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assessment Literacy Across Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Margaret E. Malone, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Francesca Di Silvio, Center for Applied Linguistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Megan Montee, Georgia State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fairness of a Large-Scale High-Stakes Language Test:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Investigations With the Test Itself, Test Users, and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test-Takers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Xiaomei Song, Queen’s University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empirically Verifying Douglas’ ESP Ability Hypotheses:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>An IRT-SEM Approach</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yuyang Cai, The University Of Hong Kong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interacting with Video in L2 Listening Tests:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>An Eye-Tracking Study</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ruslan Suvorov, Iowa State University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Investigating the Construct Validity of a Diagnostic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test of Range and Strength of Grammatical Knowledge</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hamed Zandi, University of Tehran</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Illinois at Urbana-Champaign</td>
<td></td>
</tr>
<tr>
<td></td>
<td>English Language Learning through TOEFL iBT Preparation:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Case Study of Chinese Test Takers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jia Ma, Queen’s University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Liying Cheng, Queen’s University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interaction Between Summative and Formative in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>English Language Classrooms in Colombia:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers’ Perceptions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alexis Lopez, Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identifying Test-Takers Characteristics in Assessing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a Language in Minority Contexts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Karen Fung, The University of Alberta</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Samira Elatia, The University of Alberta</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The Utility of Subscores for Automated Assessment of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>English Speaking Proficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Keelan Evanini, Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leil Chen, Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Working Model for Assessing Spanish Heritage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language Learners’ Language Proficiency through a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Placement Exam</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marta Fairclough, University of Houston</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Towards a Framework for Assessing English Through</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drama: A Dynamic Assessment Approach</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Michelle Raquel, Hong Kong University of Science</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Tasmania</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assessment Practices on a Special Oral Language</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Course in the Finnish Upper Secondary School</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Raili Vildén, University of Helsinki</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Outi Hakola, University of Helsinki</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test Design Considerations in Assessing Young</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adolescent English Learners on Computerized Language</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sultan Turkan, Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brent Bridgeman, Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Testing Arabic Speaking Proficiency: An Examination</td>
<td></td>
</tr>
<tr>
<td></td>
<td>of Test Language and Breakdown in Dialect, MSA, and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined Tests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stephanie Stauffer, Federal Bureau of Investigation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rachel Brooks, Federal Bureau of Investigation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Victoria Nier, Federal Bureau of Investigation</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>WEDNESDAY, APRIL 4</td>
<td>Location</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>SYMPOSIUM (PLENARY)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session Chair: Nick Saville, University of Cambridge ESOL Examinations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organizers: Elana Shohamy, Tel Aviv University</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tim McNamara, The University of Melbourne</td>
<td></td>
</tr>
<tr>
<td>Discussant: Anthony Kunnan, California State University, Los Angeles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How Do Different Language Frameworks Impact Language Assessment Practice?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyle F. Bachman, University of California, Los Angeles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity, Values and Policy: The Case of the CEFR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tim McNamara, The University of Melbourne</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Framing the Development of Advanced FL Literacies: Challenges for the CEFR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heidi Byrnes, Georgetown University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In the Name of the CEFR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monica Barni, Università per Stranieri di Siena</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5:10–5:30</td>
<td>BREAK</td>
<td></td>
</tr>
<tr>
<td>PAPER SESSION 5 (PLENARY)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session Chair: Charles Stansfield, Second Language Testing, Inc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5:30–6:00</td>
<td>Predictive Power of Linguistic Complexity of L2 Writing on Writing Performance and Topic Effect</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td>WeiWei Yang, Georgia State University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xiaofei Lu, Pennsylvania State University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sara Weigle, Georgia State University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6:00–6:30</td>
<td>Validating a Writing Assessment Scale: Insights from SLA, Corpora, and Computational Linguistics</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td>Angeliki Salamoura, University of Cambridge ESOL Examinations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gad Lim, University of Cambridge ESOL Examinations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6:30–7:15</td>
<td>Professional Event 3: Research Grants from Different Organizations</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td>Time</td>
<td>THURSDAY, APRIL 5</td>
<td>Location</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>8:15–6:00</td>
<td>Conference Registration</td>
<td>Regency Foyer</td>
</tr>
<tr>
<td>9:00–5:00</td>
<td>Exhibition</td>
<td>Austin</td>
</tr>
<tr>
<td></td>
<td><strong>PAPER SESSION 6 (PLENARY)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Session Chair: Rebecca Present-Thomas, VU University Amsterdam</strong></td>
<td></td>
</tr>
</tbody>
</table>
| 8:45–9:15   | **Assessing the Writing Ability of Diverse L2 Populations: Comparing International L2 to Generation 1.5 Learners**  
   Kristen di Gennaro, Pace University | Columbus–Savannah         |
| 9:15–9:45   | **Assessing English for Professional Purposes: How is Test Development Informed by Stakeholder Perceptions?**  
   Yan Jin, Shanghai Jiao Tong University  
   Liz Hamp-Lyons, University of Bedfordshire |
| 9:45–10:15  | **Collocations in Learner Speech: Accuracy, Complexity, and Context**  
   Jing Xu, Iowa State University |
| 10:15–10:35 | **BREAK**                                              |                           |
|             | **PAPER SESSION 7 (3 PARALLEL)**                      |                           |
|             | **Session Chair: Yan Jin, Shanghai Jiao Tong University** |
| 10:35–11:05 | **Investigating Item Variables Influencing Test Taker Performance on Multiple Choice Items**  
   Kirsten Ackermann, Pearson  
   Ying Zheng, Pearson |
| 11:05–11:40 | **An Empirically-Based Measure of Linguistic Distance**  
   Job Schepens, Radboud University Nijmegen  
   Frans van der Slik, Radboud University Nijmegen  
   Roeland van Hout, Radboud University Nijmegen |
| 11:45–12:15 | **Are Predictor Variables for Item Difficulty of L2 Sentence Repetition Consistent between English and Chinese?**  
   Masanori Suzuki, Pearson Knowledge Technologies |
|             | **Session Chair: Carsten Roever, The University of Melbourne** |
| 10:35–11:05 | **Using Health Professionals’ Feedback Commentary to Inform the Validation of an ESP Rating Scale**  
   John Pill, The University of Melbourne |
| 11:05–11:40 | **Legislation, Policies and Educational Assessment in a Linguistic Minority Context: Plus ça change plus ça reste la même chose**  
   Samira Elatia, The University of Alberta  
   Karen Fung, The University of Alberta |
| 11:45–12:15 | **Language and Learning: Relationship Between English Language Proficiency and Mathematics Achievement of English Learners**  
   Rosalie Grant, University of Wisconsin, Madison  
   H. Gary Cook, University of Wisconsin, Madison  
   Aek Phakiti, The University of Sydney |
## Conference Schedule

**Thurs., April 5**

### Paper Session 7 (3 Parallel) cont’d

**Session Chair:** Paula Winke, Michigan State University

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Presenters</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:35–11:05</td>
<td>Reviewers’ Perspectives on the Content Validation of an ELL Assessment</td>
<td>Youngshin Chi, WIDA Consortium; Meredith Alt, WIDA Consortium</td>
<td></td>
</tr>
<tr>
<td>11:10–11:40</td>
<td>Measuring Content Quality of Writing Performance: Putting Content in Its Place</td>
<td>Li Liu, The Chinese University of Hong Kong; Barley Mak, The Chinese University of Hong Kong; Tan Jin, The Chinese University of Hong Kong</td>
<td>Baltimore</td>
</tr>
</tbody>
</table>

**12:15–1:30**

**LUNCH BREAK**

### Language Testing Editorial Board Meeting

**Session Chair:** Alister Cumming, University of Toronto

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Presenters</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:30–2:00</td>
<td>Analytic and Holistic Rating of Oral Proficiency Interviews with Computational Tools</td>
<td>Luke Amoroso, Georgetown University</td>
<td></td>
</tr>
<tr>
<td>2:00–2:30</td>
<td>Applying Unidimensional and Multidimensional Item Response Theory Models in Testlet-Based Reading Assessments</td>
<td>Shangchao Min, Zhejiang University; Lianzhen He, Zhejiang University</td>
<td>Columbus–Savannah</td>
</tr>
<tr>
<td>3:00–3:30</td>
<td>BREAK/GROUP PHOTO</td>
<td></td>
<td>TBA</td>
</tr>
<tr>
<td>Time</td>
<td>THURSDAY, APRIL 5</td>
<td>Location</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------------------------------------------------------</td>
<td>---------------------------</td>
<td></td>
</tr>
</tbody>
</table>
| 3:30–5:15| **Implementing Innovation in Language Testing Theories and Practices: The Role of Consequences in a Validity Argument**  
Organizer: Carol A. Chapelle, Iowa State University  
Discussants: Tim McNamara, The University of Melbourne  
Xiaoming Xi, Educational Testing Service  
The Role of Consequences in Validation  
Michael Kane, Educational Testing Service  
Justifying the Use of Language Assessments  
Lyle F. Bachman, University of California, Los Angeles  
Including “Consequences” in the Revised Interpretive Argument for the TOEFL iBT  
Brent Bridgeman, Educational Testing Service  
Carol A. Chapelle, Iowa State University | Columbus–Savannah |
| 5:15–5:30| Closing Comments                                      | Columbus–Savannah         |
| 6:30–10:00| Banquet and Award Presentation at Mercer Oaks         | Bus from the hotel        |
Workshop 1: Assessing the Ability to Convey Semantic and Pragmatic Meanings in Language Assessments

Sunday, April 1 – Monday, April 2, 2012
Instructors: James E. Purpura & Kirby C. Grabowski, Teachers College, Columbia University

Communication typically involves a social, cultural, and cognitive activity in which two or more participants utilize linguistic resources and relevant contextual elements to construct meanings collaboratively through written or spoken interaction. These meanings include the literal meaning of utterances, the intended meaning of these utterances given some context, and the extended meaning of utterances that are derivable primarily from the context of the situation itself and from an understanding of the shared norms, assumptions, expectations, and presuppositions of the interlocutors. In other words, communication at any level of L2 proficiency involves not only the conveyance of semantic propositions (ideas, information, beliefs, intentions), but it also embodies a host of implied meanings relating to the interpersonal relationship of the interlocutors, their affective stance, and their role and identity in the social and cultural context of communication. As the primary purpose of communication is the conveyance of these meanings (rather than the linguistic precision with which these meanings are conveyed) and all language assessments, implicitly or explicitly, elicit semantic and pragmatic meanings, language testers need to address the role that meaning and the effective conveyance of these meanings play in assessments.

This workshop will introduce participants to the principles underlying the assessment of semantic and pragmatic knowledge. We will first discuss the importance of assessing for meaning and the implications this has for validity. We will then review how the assessment of semantic and pragmatic meanings has been conceptualized and operationalized in the language assessment literature. We will then define the constructs of semantic and pragmatic knowledge to be used in the workshop. Participants will then be asked to code tasks for these constructs. The afternoon will be devoted to the principles underlying the development of assessments designed to tap into semantic and pragmatic knowledge in a range of task types. Participants will then have the opportunity to develop tasks and present them to other participants. The second day will be devoted to issues revolving around scoring. Participants will have the opportunity to score language data for a range of semantic and pragmatic meanings and to discuss the challenges of assessing these constructs.

Dr. James E. Purpura is Associate Professor of Linguistics and Language Education at Teachers College, Columbia University, where he teaches courses in second language (L2) assessment and research methods. His research focuses on the assessment of grammar and pragmatics, the cognitive underpinnings of L2 performance, and the interface between L2 learning and assessment in instructed settings. He has written *Strategy use and second language test performance* (1999, CUP), *Assessing grammar* (2004, CUP), and is currently co-authoring a book on learning-oriented assessment in language classrooms (Routledge). In addition to being a senior associate editor of *Language Assessment Quarterly*, Dr. Purpura is co-editor of the series entitled, *New Perspectives in Language Assessment* (Routledge). Dr. Purpura served as the President of ILTA (2007-2009) and is a member of the TOEFL Committee of Examiners.

Dr. Kirby C. Grabowski is Lecturer of Linguistics and Language Education in the TESOL and Applied Linguistics Programs at Teachers College, Columbia University, where she teaches courses in second language assessment and language pedagogy. Her research interests include the assessment of grammatical and pragmatic knowledge, discourse analysis, and program evaluation. Dr. Grabowski served as Managing Editor for Teachers College, Columbia University, Working Papers in TESOL and Applied Linguistics (2007-2009), and she is currently on the Editorial Advisory Board of *Language Assessment Quarterly*. She was also a consultant on the development and validation of the Oxford Online Placement Test (OOPT). Dr. Grabowski was a 2007 Spaan Fellow at the English Language institute, University of Michigan, and she was a 2007-2008 Research Fellow at the Office of Policy and Research at Teachers College, Columbia University. She was the 2011 Jacqueline Ross TOEFL Dissertation Award recipient from Educational Testing Service.

**Agenda**

**Sunday, April 1**

9:00–9:30  Introduction to the assessment of semantic & pragmatic knowledge in language assessment
9:30–10:30  Defining the components of semantic & pragmatic knowledge
10:50–12:15  Hands-on task: Coding items for semantic & pragmatic knowledge, review and discussion
1:15–3:00  Developing tasks to measure semantic & pragmatic meanings; hand-on task development activity
3:20–5:00  Review of tasks; discussion

**Monday, April 2**

9:00–10:30  Scoring considerations: task types, rubrics, rater agreement
10:50–12:00  Application of scoring rubrics to data; discussion of scores
12:00–12:15  Wrap-up & evaluation
Workshop 2: Latent Growth Modeling for Language Testing Research

Sunday, April 1, 2012
Instructor: Gregory R. Hancock, University of Maryland

This is an introductory workshop on modeling longitudinal data, with an emphasis on latent growth modeling, and is aimed at language testing professionals and graduate students in the field. Participants are expected to have a basic understanding of multiple regression and introductory structural equation modeling (SEM). The workshop will use the SIMPLIS language within LISREL (Scientific Software International), which is the first and still one of the most widely used SEM programs. The workshop will use the free student version of LISREL installed on all computers.

This workshop will start by reviewing the basic principles of SEM with measured and latent variables, illustrating the use of SIMPLIS for such models. Next, latent means models, which add a mean structure to typical covariance-based structural models, will be introduced and illustrated within SIMPLIS. Having laid the foundations, we will then cover the basic aspects of linear latent growth models, including different time centering, uneven and varied time points, and time-independent covariates. We will then proceed to more complex topics, including time-dependent covariates and models for assessing treatments and interventions. As time allows, we may also overview nonlinear models, multidomain and cohort-sequential models, second-order growth models, latent-difference score models, and the principles of growth mixtures and models with categorical data. The workshop will have hands-on exercises to allow participants to practice running and interpreting latent growth models using SIMPLIS.

Participants are encouraged to bring a laptop computer. If you do not already have a recent version of LISREL (8.8) on your laptop, then please download the free student version at the following link: http://www.ssicentral.com/lisrel/student.html

Dr. Gregory R. Hancock is Professor and Chair of the Measurement, Statistics and Evaluation program in the Department of Human Development and Quantitative Methodology at the University of Maryland, College Park, and Director of the Center for Integrated Latent Variable Research (CILVR). His research interests include structural equation modeling (SEM) and latent growth models, and the use of latent variables in (quasi)experimental design. His research has appeared in such journals as Psychometrika, Multivariate Behavioral Research, Structural Equation Modeling: A Multidisciplinary Journal, Psychological Bulletin, British Journal of Mathematical and Statistical Psychology, Journal of Educational and Behavioral Statistics, Educational and Psychological Measurement, Review of Educational Research, and Communications in Statistics: Simulation and Computation. He also co-edited with Ralph O. Mueller the volumes Structural Equation Modeling: A Second Course (2006) and The Reviewer's Guide to Quantitative Methods in the Social Sciences (2010), and with Karen M. Samuelsen the volume Advances in Latent Variable Mixture Models (2008). He is past chair (three terms) of the SEM special interest group of the American Educational Research Association (AERA), serves on the editorial board of a number of journals, is Associate Editor of Structural Equation Modeling: A Multidisciplinary Journal has taught dozens of methodological workshops in the United States and abroad, and is a Fellow of AERA and the 2011 recipient of the Jacob Cohen Award for Distinguished Contributions to Teaching and Mentoring by Division 5 of the American Psychological Association. His former students are employed in academic positions around the U.S. and internationally, as well as at leading research agencies. Dr. Hancock holds a Ph.D. from the University of Washington.

Agenda

Sunday, April 1

9:00–10:30  Structural Equation Modeling, Foundations
- Review of structural equation modeling and introduction to the SIMPLIS language

10:45–11:30  Structural Equation Modeling, Foundations (Cont’d)
- Introduction to measured and latent mean structure models

11:30–12:15  Structural Equation Models for Longitudinal Designs: Group Level Inference
- Measured and latent variable panel designs
- Repeated measure latent mean designs
- Centering, uneven and varied time points
- Time-independent covariates
- Hands-on activities

1:00–3:00  Introduction to Latent Growth Modeling
- Basics of covariance and mean structures for linear growth
- HANDS-ON ACTIVITIES
- Nonlinear models
- Overview of growth mixtures and models with categorical data
- Time-dependent covariates
- Hands-on activities

3:15–5:00  Advanced Topics (as time allows)
- Models for assessing treatments and interventions
- Nonlinear models
- Time-dependent covariates
- Multidomain and cohort-sequential models
- Hands-on activities
Workshop 3: Scaling and Equating Test Scores

Monday, April 2, 2012
Instructors: Samuel A. (“Skip”) Livingston & Shuhong Li, Educational Testing Service

Why are raw scores from a test converted to scaled scores? Why do we need to equate scores across different forms of a test, and how can we do it? What are the issues that need to be taken into account in implementing these procedures? The answers to these questions can be found in this workshop.

Testing organizations generally report test takers’ scores as “scaled scores” or “converted scores” that are not simply the number or percentage of questions answered correctly. Instead, the scores are computed by a procedure that compensates for the changes in the difficulty of the test that can occur as the test questions are replaced with new questions. The procedure is called “equating” the scores. This workshop will introduce the participants to the basic concepts underlying the choice of a score scale and the equating of the scores. Participants will learn why it is possible for a score scale to have too much precision and why it is important not to have too little. They will learn how equating differs from statistical prediction and why an equating adjustment can be determined for a group of test takers but not for each individual test taker. They will learn the advantages and limitations of several different designs for collecting the data needed to equate test scores.

The morning session will consist of a series of illustrated lectures, interspersed with written exercises to help the participants determine how well they have understood the concepts presented. The afternoon session will include a small-group exercise in which the participants will choose a score scale and an equating design for a hypothetical new test. It will conclude with a hands-on exercise in which participants determine an equating conversion from a set of data.

Instructors:
Dr. Samuel A. (“Skip”) Livingston is a principal psychometrician in the Research & Development Division at Educational Testing Service in Princeton, NJ. Since coming to ETS in 1974, he has developed performance tests in the health professions, conducted research on methods of equating test scores, and coordinated statistical work for various types of tests, including college placement tests, college outcomes assessments, and teacher and school administrator licensing tests. He has served as a consultant to test users including professional associations, government agencies, and private corporations. He has served on the board of advisory editors for the Journal of Educational Measurement and Applied Measurement in Education and has reviewed manuscripts for many other journals, including the Modern Language Journal. He is the author of Equating Test Scores (without IRT) and a co-author (with Michael Zieky and Marianne Perie) of Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.

Dr. Shuhong Li is a psychometrician on the English Language Assessment team in the Research and Development Division of Educational Testing Service. She holds a doctoral degree in educational measurement from the University of Massachusetts Amherst. Since joining ETS in 2005, she has worked on the Internet-based Test of English of Foreign Language (TOEFL iBT) and the paper-based TOEFL (TOEFL PBT), and has recently assumed the role of Statistical Coordinator for TOEFL Junior PBT. Dr. Li's current research interests include the application of statistical/psychometric models to testing problems, with a particular emphasis on the applications of item response theory and psychometric models and principles in the context of language assessment.

Agenda
Monday, April 2
9:00–10:40 Raw scores and scaled scores
Choosing a score scale
Basic concepts of test score equating
Linear equating and equipercentile equating

10:50–12:30 Data collection designs for equating
Elevating through common items
Elevating constructed-response tests and performance assessments

1:15–3:00 Small-group exercise: Choosing a score scale and an equating design

3:20–5:00 Hands-on task: equating scores on a new test form
Carsten Roever  
The University of Melbourne  
The Role of Exposure and Proficiency in Learning and Testing Second Language Pragmatics  
This study is located at the interface of computer-based second language testing and second language learning. It investigates the relative impact of L2 proficiency and exposure to the target language environment on scores on a test of L2 pragmatics. Previous research has shown that second language proficiency impacts appropriateness of speech acts (e.g., Rose, 2000, 2009) whereas exposure affects knowledge of routine formulae (e.g., House, 1996; Roever, 2005). The case of implicature is less clear, with Bouton (1999) claiming an exposure effect and Roever (2005) making a case for an almost exclusive role of proficiency. This study used Roever’s web-based test of L2 pragmalinguistics (Roever, 2005), which consists of three sections with 12 items each, assessing interpretation of idiosyncratic and formulaic implicature, recognition of situational routine formulae, and productive knowledge of the speech acts request, apology and refusal. The implicature and routines section are in multiple-choice format and the speech act section requires brief, written responses to. The test was administered through a standard web browser, and the reliability of the total test was α = .91. 242 ESL and EFL learners in Germany and the United States participated. The learners ranged in proficiency from 3rd year of EFL in Germany (approximately CEFR level A2) to international graduate students in the United States (approximately CEFR level C2). Their average exposure to an English-speaking environment was 6.5 months, with a large number of the EFL learners reporting no exposure at all. To investigate the relative effect of proficiency and exposure for the individual sections, we employed a Poisson regression since the range of section scores (from 0 to 12) was small, scores were not normally distributed and showed excessive zeroes. We also included gender and multilingualism beyond L1 and English as predictors, but multilingualism was not a significant predictor. Gender was significant only for speech acts and total scores, with women outscoring men. Proficiency proved the strongest predictor variable, showing significance for all three sections and the total score. Residence was only significant in the Routines section, where every month of residence led to 5% increase in section scores. The Poisson model fit the sections well, with the chi-square goodness of fit test being insignificant for all sections and the total score. Adjusted R-square indicated that the model accounted for 56% of the variance in total scores and between 35% and 47% of variance in section scores. Findings from this study suggest that proficiency has the strongest impact on pragmatic knowledge with residence affecting it much less, but several caveats are in order. For one thing, the test was in the written mode, which probably advantaged test takers with high reading proficiency. Also, “residence” as a variable does not account well for the degree of contact with target community members. We will argue that the relationship between proficiency, exposure, and pragmatic ability needs further investigation employing a range of measures. Implications for testing and teaching will be discussed.

Erwin Tschirner  
University of Leipzig  
Bridging Frameworks for Assessment and Learning: The ACTFL Guidelines and the CEFR  
There are two major frameworks for learning, teaching, and assessing foreign language skills: the ACTFL Guidelines and the Common European Framework of Reference for Languages (CEFR). Both frameworks form the basis of major testing and certification systems. In addition, these frameworks are used for curriculum and textbook development and educational standard setting. Despite the fact that both systems have co-existed for 10 years, there have been very few empirical studies to establish correspondences between them. The fact that there are no official correspondences leads to the need for double certification, impedes student mobility, and may lead to a misfit in curriculum planning and textbook selection when textbooks developed for one system are used with the other one. The aim of this study is to make a contribution towards establishing correspondences for speaking proficiency. According to Kenyon (in press), several types of evidence are needed to do so: (1) internal evidence from content analyses of assessments; (2) standard setting to create links between verbally-defined levels of student performance and performances on specific tests; (3) empirical studies in which common sets of examinees take assessments linked to each framework; and (4) studies on the use and social context of the two approaches to defining and describing and language proficiency levels. The present study focuses on the second category of evidence identified by Kenyon. Six expert CEFR tester trainers used their CEFR-based evaluation criteria to evaluate 54 samples of the ACTFL OPI and the ACTFL OPIc. In September 2011, six expert CEFR raters rated speech samples of 54 participants over the course of two days. The study followed the benchmarking procedure suggested by the CEFR Manual by reacquainting participants with the CEFR speaking scale and rating procedures using benchmark samples for all CEFR levels (familiarization); by testing raters individually on the reliability of their ratings using twelve precalibrated samples (two each at CEFR levels A1 to C2); and by asking them individually to rate 54 OPI and OPIc samples that had ACTFL certified OPI ratings (six samples each at ACTFL levels Novice Mid to Superior). In addition, raters filled out a background questionnaire, answered standardized questions on each rating sample to provide insights into their rating confidence, and participated in a focus group discussion to shed additional light on their rating experience. There was a strong correlation between the judges’ ratings and the calibrated CEFR ratings (Spearman’s Rho between .879 and .988). The interrater reliability of judges rating the OPI and OPIc samples was very strong (Cronbach’s Alpha = .987) as was the correlation between ACTFL and CEFR ratings (Spearman’s Rho = .980). The paper will focus on the correspondences between ACTFL and CEFR ratings for the same set of samples and will provide additional data (model fit, rater consistency) as well as a discussion of the qualitative data obtained. Finally, the paper will discuss some of the implications for the use of tests, textbooks, and the development of curricula and educational standards across both systems for oral proficiency.
A Meta-Analytic Review of L2 Reading Component Tests

Investigation of second language (L2) reading comprehension has been largely componential (e.g., Haynes & Carr, 1990; Shiotsu, 2010). This approach, which views reading as comprised of multiple mental processes or components, enabled researchers to identify areas of individual reading ability differences (Koda, 2005) and to examine contributions of different components to reading comprehension (e.g., Droop & Verhoeven, 2003; Jeon, 2011; Van Gelderen et al., 2004). Due to these methodological advantages, L2 reading componential analysis research is steadily growing. Despite such trends, however, no meta-analysis has been conducted to render a systemic overview of findings until present. This meta-analysis in progress synthesizes findings from 63 independent samples involving 9,460 participants representing thirteen first language (L1) backgrounds and six L2s. The study quantitatively synthesizes the correlations between reading comprehension and ten variables which are frequently included in L2 reading componential analysis studies: decoding (real word and pseudoword), phonological awareness, orthographic knowledge, morphological knowledge, grammar knowledge, L2 listening comprehension, metacognition, working memory, and L1 literacy. In order to investigate whether the relative contributions of difference reading components vary as a function of moderator variables (i.e., age, L1-L2 script distance, L1-L2 language difference, measurement type) the study also reports on a series of moderator analyses. As part of the moderator analyses, the study examines how different reading components have been theoretically and operationally defined in different studies. The study also investigates whether measurement type (e.g., standardized test vs. in-house test) and test task type (e.g., multiple-choice test vs. open answer items, error correction task vs. grammaticality judgment task) cause systematic variability of the size of correlations between reading comprehension and its components across studies. Although there have been efforts to develop L2-specific tests for some reading components (e.g., Schoonen et al., 1997), L2 reading component research has been heavily relying on L1 tests, which suggests potential problems with validity and measurement practices. The present research will provide useful guidance for the improvement of measurement practices in L2 reading research.
Assessing Job-Related Language Performance Capability through Situational Role-Plays

The ability to communicate using a foreign language is an important determinant of job performance in many jobs. Organizations that hire or provide foreign language training for these jobs often require individuals to complete a standardized foreign language proficiency assessment (e.g., Oral Proficiency Interview, OPI). While such proficiency assessments are useful for documenting functional ability across a range of everyday situations, these assessments do not always indicate ability to perform the specific tasks required for successful job performance. Using McCloy, Campbell, and Cudeck’s (1994) job performance determinants model, two behavioral assessments were developed and administered to assess job-related language capability in two different training samples completing different training courses. In Study 1, a job-specific role-play assessment was developed and administered telephonically (immediately following an official OPI) to a sample of U.S. military personnel (n = 23) completing 40 hours of classroom training in Modern Standard Arabic at the Defense Language Institute. This assessment included job-specific role-plays in which test takers were required to achieve a series of objectives using only Arabic to communicate with the interviewer. Assessments were administered and rated by ILR-certified testers. All but one test taker received an ILR O+ rating on the OPI; however, there was substantial variability in test takers’ achievement of task objectives across the role-play scenarios. Some test takers achieved as few as 25% of the objectives while others achieved 100% of the objectives. Following the role-plays, there was also considerable variability in raters’ overall evaluations of test takers’ ability to perform basic communication tasks in the Arabic. These results suggest the broad OPI proficiency rating failed to capture variability in test takers’ job-specific language skills. In Study 2, a similar job-specific role-play assessment was designed and administered via a computer-mediated videoconferencing platform to a sample of U.S. Marines (n = 12) completing 92 hours of pre-deployment training in Pashto. This assessment included constructed response items designed to assess declarative knowledge, procedural knowledge, and skill, followed by simulated job scenarios in which test takers were required to achieve a series of objectives using only Pashto. Certified testers administered and rated all assessments. Findings in Study 2 were similar to Study 1. All test takers received the same unofficial ILR speaking rating. However, there was considerable variability in test takers’ job-specific declarative knowledge and achievement of objectives across the role-play scenarios (range 55% to 100%). Taken together, these findings highlight potential deficiencies in evaluating job-specific language capability using broad, standardized proficiency assessments. Even if individuals receive the same overall proficiency rating, there may be variability in those individuals’ performances within a specific job context. Thus, when using assessment outcomes to infer context-specific capability, it is advantageous to incorporate salient aspects of the target context into the assessment. Our paper will discuss performance testing in relation to the job performance determinants model and our empirical findings, with a focus on providing guidance for future research and practice related to language-related performance testing.

Modeling the Structure of Passage-Based Tests:
An Application of a Two-Tier Full Information Item Factor Analysis

Tests that consist of passages are prevalent in language testing, especially in receptive skills testing. Passages usually contain multiple items, and such items are prone to be locally dependent on one another by virtue of sharing the same passage. Traditional measurement models typically assume independence among items conditional on a construct, and passage-based tests with locally dependent items are subject to a number of problems when investigated with the local independence assumption such as overestimated reliability (e.g., Gesserolli & Folske, 2002; Lee, 2002; Lee & Frisbie, 1999; Sireci, Thissen, & Wainer, 1991) and biased model parameter estimates (e.g., So, 2010; Steinberg & Thissen, 1996). Researchers have suggested several solutions to properly capture the inherent local dependency in passage-based tests, mostly introducing additional factors or parameters to address the local dependency to existing item factor analysis models (e.g., Bradlow, Wainer, & Wang, 1999; Hoskens & De Boeck, 1997; Steinberg & Thissen, 1996; Thissen, Steinberg, & Mooney, 1989). Most such solutions, however, tend to involve excessive computational burden due to the additional factors and parameters, especially when there are multiple passages. One promising approach in this regard is the use of a bi-factor model, which consists of a general factor corresponding to the main construct and a set of specific factors for capturing passage effects. Utilizing orthogonality among factors, Gibbons and Hedecker (1992) propose a clever computing algorithm that can estimate all parameters for a bi-factor model within practical computing time. However, a bi-factor model allows only one general factor, which renders itself inapplicable to certain contexts. In particular, such a restriction limits the applicability of a bi-factor model for language tests, considering the well-known multidimensional structure of language ability (e.g., Bachman, 1990; Canale & Swain, 1980). A recent generalization of the bi-factor model, namely a two-tier item factor analysis model (Cai, 2010), does not impose the sole general factor restriction, while maintaining the theoretical and computational advantage of a bi-factor model. Furthermore, multiple general factors in a two-tier model can be correlated with one another. Therefore, a two-tier model has a large potential as a modeling tool for passage-based language tests in which one typically finds several highly correlated factors (e.g., Sawaki, Stricker, & Oranje, 2008). The current study shows an application of two-tier models in modeling passage-based language tests. To illustrate flexibility of a two-tier model, conceptual model path diagrams reflecting various possible test structures involving passages are presented, followed by the corresponding mathematical models and software to estimate model parameters. Using both simulated and actual datasets, parameter estimates of a two-tier model are compared to the corresponding estimates from a unidimensional model and a typical bi-factor model. In this presentation, I will discuss how the result of such comparison indeed supports the two-tier model's capability to flexibly model-passage-based structures in language tests.
**The Diagnosis and Treatment of Weaknesses in Reading and Writing in a Foreign Language**

In this study, a test of academic productive grammar was developed and its test validity was examined under Bachman’s Assessment Use Argument framework. This test was developed with an intention to supplement an ESL writing placement test administered at a large Midwestern research university, the decision of which is currently made on the basis of human raters’ interpretations of examinee performance on a single essay item. Fifteen constructed-response items that tap into various grammar features were developed drawing upon findings in a general developmental order of second language acquisition. The test was then administered to approximately 500 university-level ESL students in order to assess their productive grammatical ability within academic settings. Preliminary analysis results reveal that the Cronbach’s alpha reliability coefficient of the test is .75, and the item discrimination indices (represented by point-biserial correlation coefficients) are all in acceptable ranges (.30-.58). The test-takers’ performance on the academic grammar test has a relatively weak, but statistically significant, correlation with their performance on the ESL writing placement test (r=.37, p=.000). The item difficulty indices, however, do not precisely correspond to the predicted developmental stages of the target grammar features. The correlation between the examinee performance on the academic grammar test and that on the TOEFL iBT will also be reported in the study. A moderate, significant correlation is expected between the grammar test and TOEFL iBT scores, given the results of the previous research in a test of productive grammatical ability in general settings. After examining the validity of the academic productive grammar test in terms of domain definition, items and rubric developments, and warrants and backing evidence for subsequent interpretive arguments, a discussion will continue to look into the validity of the test use in a hypothetical, computer-delivered medium-stakes testing situation. To this end, examinees’ performances on the academic grammar test and on the ESL writing placement test will be analyzed at individual student and rater levels. The test use validity issues will particularly address possible consequences that the use of the academic grammar test in ESL decision making may effect on (a) individual students’ academic plans in relation to their ESL placement as well as their L2 development, (b) rater training and rubric refinement for the ESL writing placement test, and (c) the revision of the ESL curriculum at program level.

**Cognitive Diagnostic Assessment of Second Language Reading Ability for Instructional Purposes**

Cognitive diagnostic assessment (CDA) was developed to obtain detailed information regarding individual examinee’s strengths and weaknesses in a cognitive domain. Thus, CDA can be used for providing instructional implications to stake-holders (Alderson, 2005). In CDA, it is crucial to first identify the various components or attributes (e.g., knowledge, strategies) that are measured on the test based on theoretical and empirical evidence. While a handful of studies (e.g., Buck, Tatsuoka, & Kostin, 1995; Jang, 2005; Sawaki, Kim, & Gentile, 2009) have examined L2 reading ability for CDA, few have identified L2 reading attributes in relation to language ability models and reading theories. To address this issue, the current study analyzed second language (L2) reading test data with the ultimate goal of providing diagnostic feedback to administrators and teachers at the Community English Program (CEP) at Teachers College, Columbia University. In the study, first, five content experts identified the attributes required for successfully completing each item on the reading section of the CEP placement test by referring to a list of L2 reading attributes. This list was developed based on language ability models (e.g., Bachman & Palmer, 1996; Purpura, 2004) and reading research (e.g., Grabe, 2009; Pressley & Afflerbach, 1995; Weir, Hawkey, & Devi, 2009). Specifically, since reading ability is an instance of language use in which language ability is used for understanding the written text, reading ability was defined to broadly consist of language knowledge and reading strategies. Experts’ codings of the reading test were analyzed and developed into an item-by-attribute Q-matrix. Then, 1,982 CEP students’ reading scores, in conjunction with the Q-matrix, were analyzed using the Fusion model, an IRT-based cognitive diagnostic model. Using the Fusion model, it was possible to investigate individual examinees’ performance on various attributes at the item-level. Findings suggested that ten major L2 reading attributes were involved in the reading test: five knowledge-related attributes (i.e., lexical, cohesive, sentence, paragraph/text, and pragmatic meanings) and five reading strategies (i.e., identifying word meaning, finding information, skimming, summarizing, and inferencing). Examinees’ strengths (mastery of L2 reading attributes) and weaknesses (non-mastery of L2 reading attributes) on these attributes were discussed in terms of the overall group, three reading proficiency groups (i.e., beginner, intermediate, and advanced), and individual examinees. The overall group and the three reading proficiency groups showed varying attribute mastery probabilities, suggesting a hierarchy of difficulty among the ten L2 reading attributes. Similarly, individual examinees demonstrated varying attribute mastery probabilities, indicating the different needs of the students. Such diagnostic information could be provided to CEP program administrators and teachers for instructional purposes.
Item Writer Perceptions of Reading Task Difficulty

Fair assessment is based on relevant and accurate information about test takers’ abilities. This, in turn, requires a systematic definition of the abilities the band scores stand for (Cizek & Bunch 2007). The present study reports on a tentative standard setting exercise, which was carried out in the Finnish National Certificates of Language Proficiency (NC). The aim was to find out on what grounds the item writers assigned a reading comprehension task to a certain proficiency level, what factors, according to them, contributed to the overall difficulty of the task, and finally, how item writer ratings compared with authentic test scores. The NC is a national language test for adults consisting of five subtests. The NC can be taken in nine languages and on three levels (A1-A2; B1-B2; C1-C2). For the purposes of the study, three tasks of the reading comprehension subtest in Finnish were chosen to be analyzed in detail. In the first phase of the study, 26 experienced item writers in seven different test languages were asked to rate the difficulty of the text passage and each item separately. In their ratings, the item writers used a 6-point scale based on CEFR. They also briefly motivedated their ratings. After the item writers had given their ratings, Facets analysis was used to define task difficulty so that item writer’s lenience/strictness and consistency could be taken into account. The motivations written down by the item writers were analyzed qualitatively. In the second phase, the item writer data were compared with authentic test taker data (N=2300), analyzed by the Rasch-model. The results show that item writers’ ratings of item difficulty corresponded quite well to test taker scores. Nor did item writers’ native language affect the ratings. Three most important factors contributing to task difficulty were, according to the item writers, vocabulary, reading strategies needed for answering, and language structures. The results of the study can be further made use of in planning future standard setting procedures as well as in item writer training.

A Synthesis and Evaluation of Research into the Psychometric Properties of Automated Writing Evaluation

This study synthesizes and evaluates previous research into the psychometric properties of automated writing evaluation (AWE) programs, namely E-rater, Intelligent Essay Assessor [IEA], Project Essay Grade [PEG] and IntelliMetric, through the use of a meta-analysis approach. Three research questions were asked: (RQ1) What is the agreement between AWE scores and human scores, compared to the agreement between human scores?; (RQ2) What is the correlation between AWE scores and human scores, compared to the correlation between human scores?; and (RQ3) What is the correlation between AWE scores and other writing scores or measures of academic writing skills, compared to the correlation between human scores and these external measures?. In total, 24 studies were included in the meta-analysis sample, which yielded 25 agreement effect sizes for RQ1 (N = 8), 43 correlation effect sizes for RQ2 (N = 22) and 8-13 effect sizes for RQ3 (N = 2-6). SPSS was used to answer RQ1 and Comprehensive Meta-analysis Program was used to answer RQs 2 and 3. Moderator variables such as AWE programs, groups of test-takers and publication types were considered. Publication bias was considered throughout the analyses. To answer RQ1, the difference between the average agreement scores (exact and adjacent) between a human score and another human score was statistically non-significant (~53% for exact agreements and ~93% for adjacent agreements). To answer RQ2, the average human-human correlation for the random effects model was 0.727 (95% CI = 0.687-0.763), while the average human-computer correlation for the random effects model was 0.706 (95% CI = 0.668-0.741). There was no difference between the average correlation between two human scores and that between AWE scores and human scores. To answer RQ3, generally, there was a moderate relationship between AWE scores and other external measures and human scores and computer scores did not differ statistically in terms of their strength of the relationships to external objective measures. The implications, limitations and recommendations for future research into the psychometric properties of AWE programs are discussed.
Automated essay scoring (AES) is drawing a great deal of attention nowadays in writing assessment. Given its consistency and objectivity in scoring and efficiency in cost and time, it is an attractive scoring option for performance-based writing assessments. One type of positive impact expected of AES is that its adoption can facilitate wider use of constructed response writing tasks in language assessment. More recently, AES is also promoted as a means of generating diagnostic feedback for test-takers and language learners. In line with such trends, AES systems often provide not only holistic, trait, and essay feature scores but also descriptive feedback to the test-takers. This implies that there are layers of AES-produced outcomes that need to be validated and evaluated in such diagnostic application of the technology. Although a variety of validity frameworks have been used to evaluate the AES produced outcomes, however, the major focus of validation works has been on examining the relationship of the automated holistic scores with corresponding human rater scores or with other independent indicators of writing proficiency. In relation to this, some argue that the AES scores should not be evaluated in isolation from other components of a writing assessment but examined as part of a larger process of assessment development, administration, and use (Bennett & Bejar, 1988). At the moment, however, there seems to be a lack of coherent frameworks for AES validation, which is acceptable to various stakeholder communities across different contexts of writing assessments. The two major goals of the study were: (a) to critically re-examine the existing score validity/test evaluation frameworks in terms of their relevance for validating the automated scores and (b) to create a coherent evaluation framework for the AES-generated outcomes that can be used to make decisions or inform instruction in various settings. In the current study, several major existing validation frameworks were analyzed and compared in terms of their relevance and adequacy for evaluating AES-produced outcomes. Included in the analysis are traditional concepts of validity (content, criterion, construct), Messick’s (1989) unified framework, Kane’s (1992, 2004) argument-based validation approach, Chapelle et al.’s (2008) TOEFL validation framework, Bachman & Palmer’s’ (1996) six test usefulness qualities, Bachman & Palmer’s (2010) assessment use argument, and Yang et al.’s (2002) validation framework for automated scores. In the paper, I argue that the scope of validation and evaluation should not be confined to the automated holistic scores alone in the diagnostic use of AES but rather expanded to full layers of AES-produced outcomes that are going to be used, including both automated scores and feedback. This also requires that the validation process should include examining not only the relationships among the key scoring variables in AES, the rating criteria in the scoring rubrics, and the underlying models of writing proficiency, but also the impact of using the AES-produced outcomes on writing instruction. Both score interpretative and validity argument framework are proposed and illustrated in the presentation that can show links among different components of a writing assessment system employing AES.

While research in second language writing suggests that immediate feedback can have a positive influence on students’ written work (Ferris & Roberts, 2001; Stern & Solomon, 2006; Ferris, Pezone, Tade, & Tinti, 1997), the advancements of automated writing evaluation (AWE) have taken the provision of instructor commentary to a new level. Nowadays, writing analysis tools provide students not only with qualitative feedback but also with numeric scores to inform the level of their writing proficiency. Research has claimed that computerized applications provide automated scores that correlate strongly with human raters (Attali & Burstein, 2006), however little is known about the use of scores, and the reliability of these scores in comparison to instructors’ numeric grades in the language classroom. One of the AWE programs is Criterion®, an online writing evaluation service developed by Educational Testing Service (ETS). As an instructional tool, Criterion® can provide learners with a series of trait feedback and an overall holistic score for writings on the prompts and the perspectives, attitudes and linguistic development of 279 ESL students at a large Midwestern university in the United States. In this presentation, we will focus on two classroom sections and report findings of three main research questions: (1) How did learners perceive the holistic scores from Criterion®, (2) How did students use the holistic scores to improve their writing?, and (3) How well are Criterion holistic scores correlated with the instructors’ grades? To investigate these questions we administered student questionnaires, conducted student and teacher interviews, and evaluated the AWE holistic score reports and instructors’ grades that are based on a pre-developed grading rubric. Survey and interview data showed that students’ perception of score trustworthiness was moderate, yet students relied more on the feedback from Criterion®, in order to get a better score than on searching for other sources (e.g. sample essays, instructors’ help). A Spearman rho correlation coefficient was computed to assess the relationship between holistic scores and instructors’ grades. Findings indicated a positive correlation (Spearman rho=0.553, p=0.001), which raises some issues for integrating the scores for pedagogical purposes. These findings including more of our research outcomes not only add to previous work on the reliability of AWE programs but also provide justification for use or nonuse of automated feedback for formative assessment purposes.
The Impact of World Englishes on Oral Proficiency Assessment: Rater Attitude, Rating Tendency and Challenges

Listeners’ attitude of non-standard variety has been a research focus in the fields of SLA and social psychology, suggesting that listeners tend to hold negative attitudes towards speakers of non-standard English. More importantly, the relationship between attitude and behavioral tendency has been established by psychologists. It indicates a potential threat to inferences from test scores in that listeners (i.e. raters) may be biased of or in preference to particular varieties of English. As an attempt to address this issue, the present study investigated the scoring validity of the IELTS speaking section by examining its relationship to a criterion measure designed to measure rater attitudes towards World Englishes. The design of the criterion measure, named Rater Attitude Instrument (RAI), included three phases. It began with elicitation of raters’ perspectives of World Englishes through in-depth interviews with raters of a commercial oral proficiency test and a varietal speaker evaluation study. The results informed the item construction of the RAI that consisted of 22 semantic differential scale and 32 Likert scale items, which represent triplet attitude dimensions as proposed by psychologists: feeling, cognition and behavior tendency. Guided by Social Process Model (Cargile et al, 1994) that claims language attitude formation is affected by multi-factors, the items specifically reflected raters’ expectation of World Englishes, interpersonal history, rating tendency, perceived cultural factors and standard(s) against which the test should be based on.

The RAI and the six IELTS speaking samples produced by the Indian examinees were administered on-line to a group of ESL teachers in the US and India. Data from 96 participants were analyzed. The confirmatory factor analysis supported a 48-item 2-factor model with acceptable model fit indices (χ²=20.052, p=.094, RMSEA=0.076, CFI=.946, TLI=0.965). Moderate to strong correlations were observed between the RAI scores and the IELTS speaking section total and sub scores, ranging from .272 to .560. A multiple regression analysis revealed that rater perception accounted for a significant amount of variance on rating performance. Moreover, the RAI scores were analyzed by FACTS, supporting the rater variability and classification of raters into positive, neutral and negative perception according to measurement logit. MANOVA was performed which suggested ratings given by the positive and negative groups were significantly different, with positive group consistently gave higher ratings on all of the four rating criteria: fluency, sentence structure, vocabulary and pronunciation. Furthermore, rater background characteristics (e.g. nationality, gender, level of highest education and length of ESL/EFL teaching experience) were not significantly related to raters’ attitude towards World Englishes. Nevertheless, the verbal protocol study revealed insightful information on raters’ scoring judgment in relation to distinctive features of Indian English, suggesting attitude is a complex construct and may or may not be activated to influence rating decision.

This study is timely in view of the growing interest in research on World Englishes principally in relation to language assessment. The results of this study are of particular importance for a variety of reasons. It is among the first in language testing research engaged in instrument development that attempts to investigate rater belief, effect, and intention. It is expected to bring researchers’ attention to rater psychological traits that have been rarely examined. Furthermore, the PRI with further modifications may provide testing professionals a common tool and unified approach to evaluate rater attitude toward varieties of English. Lastly, the impact of the findings on validity argument, test fairness and rater trainings will be discussed.

Using Oral Reading Performance Features to Predict Reading Comprehension Ability in Second Language Learners of Dutch

This paper presents a study examining the extent to which performance-based characteristics of candidates’ oral reading, such as reading accuracy, rate, and expression, can be used to predict the reading comprehension ability of candidates reading in their non-native language. Previous research suggests that oral reading features such as reading rate generally correlate with reading comprehension ability in developing readers, on the order of r = 0.67 – 0.80. The relationship is not as well established, however, in second language, adult learners (e.g., Samuels, 2006). The Dutch Reading Test scores candidates’ oral reading performance in Dutch and returns estimates of reading comprehension ability, placing candidates at “Below A1” or “A1 or above” relative to the reading comprehension scales on the CEFR. Candidates perform a variety of oral reading tasks, reading their responses aloud into a telephone. Candidates are scored along five independently measured subscores: Accuracy, Expression, Reading Rate, Phonics, and Content. Subscores are produced on an internal scale from 5 to 40. The Content subscore is drawn from performance on two “traditional” reading comprehension tasks. In Fill in the Gaps, a modified cloze task, candidates select the best option out of three which correctly completes a sentence they read aloud. In the Passage Comprehension task, candidates read a passage and then answer open-ended questions about it (e.g., “How old did Jaap turn on his birthday?”). Scores on these tasks have been aligned with the CEFR subskills related to reading comprehension (Lam et al., 2011). The other performance-based features which do not appear in the CEFR scales, however, are used to provide further evidence of reading comprehension ability. Multiple regression was used on data from Dutch Reading Test candidates (n = 2623) to examine how Accuracy, Expression, Phonics, and Reading Rate could be used to predict Content scores. Only Accuracy and Reading Rate made significant contributions to the models (p < .05): Accuracy alone accounted for 72% of the variance with Content (r= .846), while a model with Accuracy and Reading Rate accounted for 73% of the variance (r=.851). As expected, the correlation between Phonics and Content scores was low (r=.53); candidates with poor Dutch pronunciation may comprehend what they are reading, and vice versa. Reading Rate and Expression subscores were correlated (r = .90), likely due to the fact that most rapid readers are also able to read smoothly and with appropriate phrasing. Those two traits correlate moderately with Content (r=.66 and .61, respectively), suggesting that in this sample of second language readers of Dutch, Reading Rate and Expression by themselves have comparable usefulness in predicting reading comprehension ability as similar traits when measured in developing readers reading in their first language. This study concludes with an examination of the extent to which these results hold when scores are limited to a specific test task – reading short passages out loud – and whether using scores from this task alone could provide sufficiently reliable predictions of a candidate’s reading comprehension ability.
Examiner training is a vital element of performance testing, essential for the consistent provision of valid and reliable test marks. Training needs to be efficient and practical for participants whilst still maintaining quality. One way to improve training is to give raters more practice by giving access to more and varied training samples; though this places an increased resource burden on both raters and exam providers. A possible solution is to have raters assess shorter test chunks, or test parts. This however raises the question of whether rating shorter performances affects rater tendencies and therefore affects training quality. The introduction of an online system of speaking examiner training and standardisation allows the use of test parts in addition to traditional full tests. In addition, we will report the test takers’ views on the usefulness of (c), (d) and (e) above in preparing them for the TOEFL iBT speaking tasks. The findings have important implications, not only for test developers but also providers of test preparation programmes that may influence test performance. They contribute to the on-going validity argument for the TOEFL iBT speaking tasks and further, to enhancing our understandings of the pedagogic practices of the test preparation programmes and the extent to which such programmes offer opportunities for developing speaking proficiency, and hence addressing the delicate relationship between testing, learning and ethical coaching.
Predicting Proficiency Without Direct Assessment: Can Speaking Ratings be Inferred from Listening and Reading Ratings?

Language proficiency testing can be a costly investment. Organizations have a strong motivation to reduce testing costs, particularly during difficult economic times. One proposed solution is using test results from one language modality to predict scores in other modalities, thus reducing costs. Despite limited empirical evidence documenting the relationships between proficiency in second language modalities (or skill domains), such as listening, reading, and speaking proficiency (Hartley, 2007), theoretical perspectives provide compelling evidence for the multifaceted and interrelated nature of language skill modalities in general (e.g., Bachman & Palmer, 1981; Eckes & Grotjahn, 2006). The current research takes an evidence-based approach (e.g., Briner & Rousseau, 2011) and consists of four separate studies with the objective of exploring the interchangeability of scores obtained from two tests of language proficiency, the Defense Language Proficiency Test (DLPT), which tests reading and non-participatory listening, and the Oral Proficiency Interview (OPI), which tests speaking. Study 1 was a meta-analytic review which explored the relationships between speaking, listening, and reading proficiencies. A total of 8,343 studies were identified in our literature review and 86 studies included sufficient information for our study. Meta-analytic correlations between speaking and non-participatory listening (k=191; r=.67) and speaking and reading (k=184; r=.58) proficiency suggest moderate relationships. Furthermore, examining specific tests shows only moderate relationships between the OPI and the DLPT (k=35, r=.59 for listening and k=38, r=.59 for reading). These findings suggest speaking proficiency estimates based solely on non-participatory listening and reading proficiency ratings would be inaccurate for a substantial proportion of cases. Studies 2 and 3 included correlational and regression analyses conducted with two different groups of U.S. military personnel. Study 2 included data from 58 language trainees from Air Force Special Operations Forces (SOF), while Study 3 included data from 3,305 language trainees from the United States Army SOF. Study 2 showed ratings on the DLPT-L and DLPT-R were related to OPI-S proficiency ratings (r = .66 and r = .49, respectively), but the absolute agreement between DLPT and OPI ratings was poor, ranging from 12% agreement between DLPT-R and OPI-S to 30% agreement between DLPT-L and OPI-S. Study 3 findings were consistent with Study 2. In addition, Study 3 findings showed that using an optimally weighted combination of DLPT-L and DLPT-R ratings provided only a partially accurate (41%) estimate of OPI-S. Therefore, the preponderance of evidence suggests that while the scores are related, non-participatory listening and reading proficiency ratings cannot produce sufficiently reliable and accurate estimates of speaking proficiency to indicate interchangeability. Finally, Study 4 reports user reactions gathered at two different time periods from SOF operators to assess reactions to the DLPT and the OPI. Studies from both time periods indicated that operators believe the OPI is more related to job performance than the DLPT. These findings suggest that even if ratings for different language modalities could be used interchangeably from a psychometric perspective, using a test that is not clearly related to job requirements can have negative consequences, resulting in misalignment between training, testing, and job requirements.

A View of Teacher and Student Interactions and Learning in Two Classroom Applications of L2-AFL

Descriptions of teacher-student and student-peer interactions in classrooms using the Assessment for Learning (AFL) approach have been frequently reported in general education research (Anderson, 2004; Black & Wiliam, 1998; Hodgen & Marshall, 2005). Conversely, interactions of this kind have remained relatively infrequently reported in studies of second language classroom AFL (L2-AFL) applications, since those applications have only recently occurred (Colby-Kelly & Turner, 2007; Davison, 2007). It is important to examine how teachers and students interact in L2-AFL contexts to gain insight into how AFL could best succeed there. This paper reports on teacher-student and student-peer interactions as part of a larger, quasi-experimental, exploratory study using mixed methods to investigate an application of L2-AFL. The study also sought evidence of the assessment bridge (AB), the area where classroom assessment, teaching, and learning intertwine. The research question asked is, What is the nature of teacher and student interactions in two second language classrooms taking an AFL approach? The participants were 2 teachers and 28 students in two classes of pre-university English for Academic Purposes (EAP). An L2-AFL methodology was developed for the study in the form of teacher training. Additionally, L2-AFL pedagogical materials were developed for the study. The learners were taught the usage of English would and will forms, using CALL and concept mapping modules to support learner conceptualization of grammatical usage, and with group and teacher-class activities. The data collection instruments included transcribed classroom discourse, teacher and student questionnaires, observation field notes, and the concept maps produced. The data were analyzed by mixed methods and the results triangulated. From a qualitative analysis of classroom observations and discourse transcriptions of student group activities, 15 categories of student interaction types emerged. These were further qualified as either co-construction focused interactions (those behaviors which have as their focal point an interactive contribution to the group experience), or individually-focused interactions (those serving as an expression of realization of the self). In addition, evidence of the AB, the interweaving of assessment, teaching (both from teachers and peers) and learning, was found in eight of the interaction types. From a qualitative analysis of classroom observation data and discourse transcriptions of teacher and class activities, 12 categories of teacher interactional types emerged. Of these, seven reflected an L2-AFL approach. Qualitative and quantitative analysis of teacher and student questionnaire data suggested that learner metacognition had been enhanced, and demonstrated that the participants believed learning had taken place as a result of the L2-AFL methodology. The students reported that the group activity (91%) and the teacher-class activity (100%) had helped them learn, and the two teachers both reported that they believed their students had learned from the L2-AFL activities, thus, showing strong teacher and student support for L2-AFL. The qualitative evidence here presented informs on the nature of the teacher and student interactions in these L2-AFL settings. Based on the results, a call is made for a research agenda into further study of the nature of teacher and student interactions in L2-AFL settings in support of L2-AFL best practices.
Effects of Accent Familiarity and Language Teaching Experience on Raters’ Judgments

In the rater cognition literature, rater effects are defined as the construct-irrelevant variation associated with rater characteristics, and are critical to the reliability and validity of the assessment (Bachman, Lynch, & Mason, 1995; Kunnan, 2000). Potential sources of rater variation, such as native language background and rater training (e.g., Kim, 2009; Xi & Mollaun, 2009) have been investigated in the context of speaking tests. However, little is known about the effects of raters’ familiarity with the speakers’ accents and their experience in teaching English as a second/foreign language (ESL/EFL). The current study thus seeks to contribute to the scarce literature by investigating the two under-explored background characteristics. This study examined the effects of the two variables through a quasi-experimental design, including three groups of raters (n=22 in each group) who varied on the two characteristics: Naïve Non-Teachers (no familiarity with foreign accents and no ESL/EFL teaching experience), Experienced Non-Teachers (some familiarity with foreign accents but no ESL/EFL teaching experience), and Experienced Teachers (some familiarity with foreign accents and some ESL/EFL teaching experience). All raters participated in an accent identification task with 26 speech samples taken from an international English speaking test, and assigned holistic and analytical ratings (i.e., foreign accents, grammar and vocabulary, content) to each speech file. Participants also filled out a survey reporting their demographic information and perceived impacts of their background on their rating. The ANOVA results suggested that the three groups did not differ significantly in their ratings of L2 proficiency or in the evaluative features they attended to while rating. However, the survey results indicated that experienced raters (Experienced Non-Teachers and Experienced Teachers) believed that their accent familiarity and teaching experience did affect their rating decisions. They also commented that their familiarity and teaching experiences enhanced listening comprehension and error detection. In particular, Experienced Teachers believed that their teaching experience made them more lenient toward judging the English proficiency of non-native speakers. Compared to non-teachers, Experienced Teachers were also less biased by the foreign accents when judging the overall proficiency or the content of the non-native speech. The results have practical implications for rater selection and training in speaking assessments. Despite the absence of significant group differences in ratings, which may be due to the sample size constraints, Experienced Teachers believed that their background led them to be more lenient in rating. Caution should thus be taken to guard against the potential leniency effect from raters with ESL/EFL training. Additionally, training materials that include speech from a variety of native languages would also help familiarize raters with foreign accents and prevent biases toward a particular language background. The study also has broader implications for ESL/EFL instruction. ESL/EFL professionals appear to have higher tolerance for thick accents than non-teachers, but it is the non-teachers that students mostly come into contact with in daily life. ESL/EFL professionals may thus need to pay more attention to pronunciation instruction, particularly for students with strong accents that could impede effective communication.

Validating a Word Associates Test in EFL Contexts

While Read’s (1998) Word Associates Test (WAT) has been widely used in word association research, little has been done to validate the test or its format as a depth test in various settings. As the WAT was developed targeting ESL learners, it is questionable if the test or format would work with EFL learners. How to score the WAT is another issue to address to obtain reliable and useful test results (Schmitt et al., 2011). The present study investigated if the WAT is a valid and reliable word-associates measure for Korean ESL students, with six scoring schemes. In Study 1, the WAT was evaluated as a word-associates test for Korean ESL learners at varying proficiency levels and modified into a 30-item test, MWAT, following the results of a pilot test on 10 Koreans of proficient English speakers. The MWAT was constructed to have two subtests, measuring either paradigmatic knowledge (synonyms) or syntagmatic knowledge (collocations). In Study 2, the MWAT and its concurrent measure, a collocation test, were conducted with 195 Korean EFL learners at various proficiency levels. Six scoring schemes based on the All-or-Nothing (0 or 1) were used to score the MWAT, and two statistical analyses—generalization theory (in a crossed two-facet design) and correlations—were performed on the test data. The analyses yielded a quite high level of dependability of the MWAT and correlational coefficients when the strictest scoring scheme was used: one point per item of each subtest with all correct associates. These results provided a considerable amount of reliability and validity evidence for the MWAT. Findings from the present study point to the importance of various reliable and valid word associates tests considering different learning settings. This paper ended with a discussion of some major limitations of the present study, along with a review on avenues for further investigation.
Developing a Code of Practice for EFL Testing in China: A Data-Based Approach

Jinsong Fan
Fudan University

Yan Jin
Shanghai Jiao Tong University

Wednesday, April 4
11:45–12:15
Baltimore

The pursuit for professionalism, quality and fairness has been an apparent trend in language testing and assessment on the international front, as is evidenced by a number of standards or codes of practices which have been developed, implemented or enforced by language testing or research organisations from all over the world (see e.g., AERA/APA/NCME, 1999; ALTE, 1994; EALTA, 2006; ETS, 2002; ILTA, 2000, 2007; JLTA, 2003). Concomitantly, EFL testing has been developing with great momentum in China in the past 20 years or so. These EFL tests are normally large-scale and standardised, and the results on these tests are often used to make high-stakes decisions. Therefore, it is essential that the development, administration and use of these EFL tests follow a set of well-developed code of practice, which can cater to the specific needs and circumstances of the Chinese testing context. This study reports a data-based approach to developing such a code (hereinafter the Code) and to promoting good testing practice among test developers and other stakeholders. The data in this study were collected through a systematic and detailed review of the language testing standards and codes of practice which were developed or enforced in other testing contexts (see Fan & Jin, 2010), and more important, an empirical investigation of the current EFL testing practices, involving both test developers and the primary stakeholder groups (see Fan & Jin, 2011). The review and critique of the language testing standards made it possible to construct a preliminary model of code creation and validation. The empirical investigation involved 6 examination boards, 490 students and 166 teachers from different parts of China, adopting documentary analysis, interview and questionnaire as its primary research instruments. While identifying the areas which were in need of improvement in test development, administration and use, the investigation also concluded that stakeholder groups, such as examinations authorities, educational administrators, students and teachers played a vital role in ensuring testing quality and fairness in the Chinese testing context. Then based on the review of the language testing standards and the empirical investigation of language testing practice, the structure of the Code was determined, stipulating the ethical and professional responsibilities of the major stakeholders in the testing process, including test developers, examinations authorities, test takers and test users. In addition, the empirical investigation also helped us to draft the guidelines which catered more specifically to the Chinese testing context. It is expected that this study, the first research effort ever in China to develop a code of practice in language testing, can help to raise the awareness of professionalism and quality among language test developers, and at the same time, help to promote the communications between test developers and the stakeholder groups in China. Future studies will be focusing on the a posteriori validation of the Code, including its applicability, usefulness and impact.

WeiWei Yang
Georgia State University

Xiaofei Lu
Pennsylvania State University

Sara Weigle
Georgia State University

Wednesday, April 4
5:30–6:00
Columbus–Savannah

Predictive Power of Linguistic Complexity of L2 Writing on Writing Performance and Topic Effect

Jinsong Fan
Fudan University

Yan Jin
Shanghai Jiao Tong University

Wednesday, April 4
11:45–12:15
Baltimore

In second language (L2) writing assessment, linguistic complexity has been used as a criterion in judging the quality of L2 writing, as evidenced in the explicit criteria of range of vocabulary used, lexical variation, and complex constructions in many writing rating rubrics (e.g., Gentile, Riazanzeva, & Cline, 2002; Jacobs, et al., 1981). Yet, the exact relationships between linguistic complexity and writing performance await more empirical explorations, particularly with computational tools which allow analysis of large sets of data in an efficient manner and enable precise and diverse linguistic measures to be automated. Further, the role of writing topic on linguistic complexity of L2 writing has not been given due attention. The current study addressed the above research needs. The linguistic complexity measures the study examined included lexical diversity, lexical sophistication, and syntactic complexity. A total of 382 essays written by 191 ESL graduate students studying in the U.S. were collected and analyzed, with each participant writing on two different topics. Each essay was rated on a five-point holistic scale by two experienced and trained raters, and the average was taken as the writing score for each. For each essay, D index, a lexical diversity measure, was obtained with the vocd program (Malvern, Richards, Chipere, & Duran, 2004) in the Computerized Language Analysis (CLAN) programs (MacWhinney, 2000) of the CHILDES project, normalized frequency for the number of academic and lower-frequency words, a lexical sophistication measure, was obtained using the Range program (Heatley, Nation, & Coxhead, 2002), and 14 syntactic complexity indices were obtained from the L2 Syntactic Complexity Analyzer (Lu, 2010). 337 essays with their length falling into the range of 200-500 words were selected for analysis of all the linguistic complexity measures, due to the requirement for reliability of vocd D measure and lexical sophistication index. Syntactic complexity of all 382 essays was analyzed to facilitate predictor selection in the regression analysis. Statistical analyses included correlations, multiple regressions, and dependent samples t tests. Correlation analyses for essays with 200-500 words showed that lexical diversity, lexical sophistication, global sentence complexity, and amount of use of complex nominals significantly correlated with writing scores for both topics. Three other syntactic complexity indices also significantly correlated with scores on one topic. Multiple regressions revealed lexical sophistication and amount of subordination as significant predictors of writing scores for one topic, accounting for 24 percent of variance. For the other topic, lexical sophistication itself was found to be a significant predictor of writing scores, explaining 23 percent of variance. Regression analysis for all the 382 essays with the 14 syntactic complexity measures as predictors also revealed mean length of sentence, a global sentence complexity measure, as a significant predictor of writing scores for both topics. Elaboration at the finite-clause level was also a significant predictor for scores on one topic. Finally, dependent samples t tests showed that topic significantly affected linguistic complexity of L2 writing, including lexical diversity, lexical sophistication, and syntactic complexity. The study has implications for assessing second language writing and measuring linguistic complexity.
Comparing International L2 to Generation 1.5 Learners

Assessing the Writing Ability of Diverse L2 Populations: Comparing International L2 to Generation 1.5 Learners

Validating a Writing Assessment Scale: Insights from SLA, Corpora, and Computational Linguistics

The paper considers the interface between a number of sub-fields within applied linguistics by describing how data driven analyses of learner output enabled by the latest developments in corpus and computational linguistics can aid the description of key features of learner writing at the B2 level of the Common European Framework of Reference (CEFR; Council of Europe, 2001) and of validating the assessment criteria for an examination at that level. For the examination in question, aimed at the B2 level, a new analytic assessment scale covering different aspects of language ability was recently developed, with descriptors more explicitly tied to those in the CEFR. Multiple methods were used to verify that the descriptors were indeed at the right level; this paper presents one method, which considered pass-fail decision consistency between the old and new rating schemes. A corpus of writing from the exam was assembled consisting of approximately 3,000 scripts. Using machine learning techniques, an engine was built to discriminate between passing and failing scripts on the old rating scheme. A user interface was also developed to support exploration of features that highly discriminate between passing and failing scripts; this calculated feature co-occurrences within sentences, correlations with errors and L1s, and the like. Using this interface, the top 40 positive and top 40 negative discriminative features identified by the engine was analyzed. This data-driven analysis provided insights into key features of writing performance at B2 level in comparison to performance below B2, e.g. the successful use of complex tense, nominal reference and word order, a variety of linking and cohesive words, formulaic expressions. These features were then compared with the new assessment scale descriptors to see the extent to which they matched each other. The insights gained helped to provide supporting criterion/concurrent validity evidence about the appropriateness of the new descriptors. The paper therefore addresses the application of innovations and practices in other disciplines to language assessment and the benefits that these bring.

Reference

Assessment of second-language (L2) learners for placement into college writing programs poses a challenge given the heterogeneous nature of the L2 population in US post-secondary institutions. Traditionally, the majority of L2 learners attending US colleges have been international students, having arrived in the US after having completed high school in their home countries. Currently, however, an increasing number of L2 learners are entering US colleges upon completion of high school in the US. These so-called Generation 1.5 learners are said to share characteristics of both first- and second-generation learners. Research in second-language writing suggests that the writing ability of Generation 1.5 learners differs noticeably from that of international L2 learners, yet very little empirical evidence exists confirming such differences. The present study provides empirical evidence concerning the nature of similarities and differences in the writing ability of these two groups of learners. A mixed-methods research design was adopted to examine International L2 and Generation 1.5 learners’ writing ability from both quantitative and qualitative perspectives. Many-facet Rasch measurement (MFRM) procedures were used to analyze 134 learners’ writing scores from 3 raters in 5 components intended to represent the construct of writing ability. Specifically, learners’ writing was rated for grammatical, cohesive, rhetorical, sociopragmatic, and content control. A whole-group MFRM analysis indicated that the International L2 learners, as a group, outperformed the Generation 1.5 learners. Results also revealed an interaction effect for the component of grammatical control only. Separate-group MFRM analyses revealed that the two groups had opposing strengths and weaknesses in two components of writing ability. Specifically, the International L2 learners performed best in grammatical control yet poorly in sociopragmatic control, and the Generation 1.5 learners performed best in sociopragmatic control yet poorly in grammatical control. Subsequent qualitative analyses included an in-depth examination of a subset of International L2 and Generation 1.5 learners’ writing. Specifically, ten learners (five from each group) with a similar degree of writing ability according to the MFRM analyses were analyzed according to the types of grammatical errors and sociopragmatic markers exhibited in their writing. Findings from these analyses revealed that the Generation 1.5 group’s grammatical errors reflected a lack of awareness of certain grammatical features typical of academic writing. Likewise, the International L2 group’s use of sociopragmatic markers reflected a tendency to draw on personal opinions and other non-academic sources in their writing. Considering both the quantitative and qualitative findings in light of one another, the results suggest that both groups’ writing included a lack of adherence to different aspects of academic writing. The findings from this study offer much-needed empirical evidence as to observable differences concerning the writing ability of International L2 and Generation 1.5 learners. As such, the findings can assist ESL program coordinators, placement test designers, and writing instructors in making placement and pedagogical decisions for post-secondary L2 learners, especially in contexts including both International L2 and Generation 1.5 learners.

Comparing International L2 to Generation 1.5 Learners

Assessing the Writing Ability of Diverse L2 Populations: Comparing International L2 to Generation 1.5 Learners

Validating a Writing Assessment Scale: Insights from SLA, Corpora, and Computational Linguistics

This paper considers the interface between a number of sub-fields within applied linguistics by describing how data driven analyses of learner output enabled by the latest developments in corpus and computational linguistics can aid the description of key features of learner writing at the B2 level of the Common European Framework of Reference (CEFR; Council of Europe, 2001) and of validating the assessment criteria for an examination at that level. For the examination in question, aimed at the B2 level, a new analytic assessment scale covering different aspects of language ability was recently developed, with descriptors more explicitly tied to those in the CEFR. Multiple methods were used to verify that the descriptors were indeed at the right level; this paper presents one method, which considered pass-fail decision consistency between the old and new rating schemes. A corpus of writing from the exam was assembled consisting of approximately 3,000 scripts. Using machine learning techniques, an engine was built to discriminate between passing and failing scripts on the old rating scheme. A user interface was also developed to support exploration of features that highly discriminate between passing and failing scripts; this calculated feature co-occurrences within sentences, correlations with errors and L1s, and the like. Using this interface, the top 40 positive and top 40 negative discriminative features identified by the engine was analyzed. This data-driven analysis provided insights into key features of writing performance at B2 level in comparison to performance below B2, e.g. the successful use of complex tense, nominal reference and word order, a variety of linking and cohesive words, formulaic expressions. These features were then compared with the new assessment scale descriptors to see the extent to which they matched each other. The insights gained helped to provide supporting criterion/concurrent validity evidence about the appropriateness of the new descriptors. The paper therefore addresses the application of innovations and practices in other disciplines to language assessment and the benefits that these bring.

Reference

Assessing the Writing Ability of Diverse L2 Populations: Comparing International L2 to Generation 1.5 Learners

Assessment of second-language (L2) learners for placement into college writing programs poses a challenge given the heterogeneous nature of the L2 population in US post-secondary institutions. Traditionally, the majority of L2 learners attending US colleges have been international students, having arrived in the US after having completed high school in their home countries. Currently, however, an increasing number of L2 learners are entering US colleges upon completion of high school in the US. These so-called Generation 1.5 learners are said to share characteristics of both first- and second-generation learners. Research in second-language writing suggests that the writing ability of Generation 1.5 learners differs noticeably from that of international L2 learners, yet very little empirical evidence exists confirming such differences. The present study provides empirical evidence concerning the nature of similarities and differences in the writing ability of these two groups of learners. A mixed-methods research design was adopted to examine International L2 and Generation 1.5 learners’ writing ability from both quantitative and qualitative perspectives. Many-facet Rasch measurement (MFRM) procedures were used to analyze 134 learners’ writing scores from 3 raters in 5 components intended to represent the construct of writing ability. Specifically, learners’ writing was rated for grammatical, cohesive, rhetorical, sociopragmatic, and content control. A whole-group MFRM analysis indicated that the International L2 learners, as a group, outperformed the Generation 1.5 learners. Results also revealed an interaction effect for the component of grammatical control only. Separate-group MFRM analyses revealed that the two groups had opposing strengths and weaknesses in two components of writing ability. Specifically, the International L2 learners performed best in grammatical control yet poorly in sociopragmatic control, and the Generation 1.5 learners performed best in sociopragmatic control yet poorly in grammatical control. Subsequent qualitative analyses included an in-depth examination of a subset of International L2 and Generation 1.5 learners’ writing. Specifically, ten learners (five from each group) with a similar degree of writing ability according to the MFRM analyses were analyzed according to the types of grammatical errors and sociopragmatic markers exhibited in their writing. Findings from these analyses revealed that the Generation 1.5 group’s grammatical errors reflected a lack of awareness of certain grammatical features typical of academic writing. Likewise, the International L2 group’s use of sociopragmatic markers reflected a tendency to draw on personal opinions and other non-academic sources in their writing. Considering both the quantitative and qualitative findings in light of one another, the results suggest that both groups’ writing included a lack of adherence to different aspects of academic writing. The findings from this study offer much-needed empirical evidence as to observable differences concerning the writing ability of International L2 and Generation 1.5 learners. As such, the findings can assist ESL program coordinators, placement test designers, and writing instructors in making placement and pedagogical decisions for post-secondary L2 learners, especially in contexts including both International L2 and Generation 1.5 learners.

Validating a Writing Assessment Scale: Insights from SLA, Corpora, and Computational Linguistics

This paper considers the interface between a number of sub-fields within applied linguistics by describing how data driven analyses of learner output enabled by the latest developments in corpus and computational linguistics can aid the description of key features of learner writing at the B2 level of the Common European Framework of Reference (CEFR; Council of Europe, 2001) and of validating the assessment criteria for an examination at that level. For the examination in question, aimed at the B2 level, a new analytic assessment scale covering different aspects of language ability was recently developed, with descriptors more explicitly tied to those in the CEFR. Multiple methods were used to verify that the descriptors were indeed at the right level; this paper presents one method, which considered pass-fail decision consistency between the old and new rating schemes. A corpus of writing from the exam was assembled consisting of approximately 3,000 scripts. Using machine learning techniques, an engine was built to discriminate between passing and failing scripts on the old rating scheme. A user interface was also developed to support exploration of features that highly discriminate between passing and failing scripts; this calculated feature co-occurrences within sentences, correlations with errors and L1s, and the like. Using this interface, the top 40 positive and top 40 negative discriminative features identified by the engine was analyzed. This data-driven analysis provided insights into key features of writing performance at B2 level in comparison to performance below B2, e.g. the successful use of complex tense, nominal reference and word order, a variety of linking and cohesive words, formulaic expressions. These features were then compared with the new assessment scale descriptors to see the extent to which they matched each other. The insights gained helped to provide supporting criterion/concurrent validity evidence about the appropriateness of the new descriptors. The paper therefore addresses the application of innovations and practices in other disciplines to language assessment and the benefits that these bring.

Reference
Assessing English for Professional Purposes: How is Test Development Informed by Stakeholder Perceptions?

A new computer-based assessment of English at an advanced level is being proposed in China in response to the needs of society for professionals proficient in English for specific purposes. This assessment is expected to provide a diagnostic evaluation of the English language proficiency of potential employees at a level and in areas of relevance to multinationals and professional fields operating in China and working with clients internationally. This advanced test of English for professional purposes (EPP), provisionally named PACE—Profile of Advanced Competence in English, is designed to have multiple profession-related strands in addition to a general component in order to produce a “profile matrix” that reports test takers’ performances on major language skills as well as profession-linked skills. An online survey drawing on the structure of the Common European Framework of Reference for Languages (CEFR) was conducted among the major stakeholder groups for the assessment: the survey questions elicited stakeholders’ views on 1) the types of communicative situations employees are involved in, 2) the types of language tasks employees are engaged in, 3) the difficulty level of the tasks, 4) competence requirements to fulfill the tasks, and 5) the types of oral and written texts produced (see Chapters 3, 4 and 5 in the CEFR Handbook, Council of Europe 2001). Adaptation to the overall questionnaire was made for each broad professional area and stakeholder group based on the results of pilot surveys. Data for the main study were collected from carefully targeted groups: employers of different types of companies (Ne1=60), employees in these companies (Ne2=89), university teachers providing content-based instruction in English (Nt1=39), university English language teachers (Nt2=59), and university students majoring in different disciplinary areas (Ns=104). Descriptive and ANOVA analyses of these stakeholder groups’ expectations of performance presented a complex picture of language skill needs and disciplinary areas, and convinced us that the English language tests currently used in practice in China such as CET, PETS, or TEM are neither specific nor advanced enough for the purpose of assessing English proficiency for workplaces. Based on the results of this study, and to support a validity argument for the design of the test, a “model workplace”—hypothetical, but based in a real case—for those who would be consumers of score reports from each test domain was constructed. Following these models, detailed test specifications were drafted and prototype tests developed for the five professional contexts of interest in the study. After reporting the findings of the survey, and illustrating the models that were constructed, the paper will end with a discussion of the challenges facing an EPP assessment with reference to the experiences of the original ELTS test (British Council 1980). Some of the complexities of the delivery, marking and score reporting of such a complex assessment, one with different modules and a speaking component in each module, over the huge territory of China will also be discussed, as will some of the security issues involved.

Collocations in Learner Speech: Accuracy, Complexity, and Context

The state-of-the-art automated speech evaluation (ASE) systems such as SpeechRaterSM and VersantTM use a small subset of the criteria evaluated by human raters to predict human scores on second language (L2) oral proficiency. However, for more accurate and trustworthy score prediction and to meet the needs of high-stakes speaking assessment, the construct coverage of the current ASE systems needs to be expanded to include additional high-level speech features that are both construct-relevant and measurable (Xi, 2010, p. 294). Collocations play a scaffolding role in building up language but these formulaic expressions, which are used by native speakers to facilitate speaking, were found to pose serious problems for L2 speakers (e.g., Wang & Shih, 2011). Many researchers argue that L2 learners’ collocational competence is a good indicator of their L2 oral proficiency (e.g., Handl, 2008). Hence, the collocational production in learner speech, if measured appropriately, might be useful in improving machine graders’ predicting power. The present study explored several ways of evaluating collocations produced by L2 speakers in spontaneous speech and the potential for using these collocational features to predict overall L2 oral proficiency. Sixty Chinese learners of English representative of three oral proficiency levels were randomly selected from a spoken corpus comprising both monologic/academic and dialogic/non-academic speech samples of an institutional oral English test, namely Oral English Certification Test (OECT). From the transcriptions of these speech samples, collocational strings of 10 syntactic patterns were manually extracted. These strings were then double-coded by trained native-English-speaking linguists for acceptability and grammaticality. Additionally, the sophistication of each string was measured via “modifier combinability”: the number of acceptable combinational partners the modifier in a collocation can take according to the Corpus of Contemporary American English, a 425 million word collection of American English. Based on a series of variables derived from the above human-coded features, the collocations produced by L2 speakers of three different oral proficiency levels and across the two distinctive speaking contexts were compared. Further, these variables were included in a regression model to predict the OECT holistic scores given by human raters. The study revealed a developmental pattern of L2 collocational competence across L2 oral proficiency levels as well as an interaction between oral collocational performance and speaking contexts. Finally, the suitability and feasibility of applying the collocational features investigated in this study to the ‘collocation tool’, an automated collocation evaluation program currently being developed for SpeechRaterSM, will be addressed.
Investigating Item Variables Influencing Test Taker Performance on Multiple Choice Items

This paper combines quantitative and qualitative analysis in order to investigate the effect that specific item variables of multiple choice item types have on item difficulty and test taker performance. The study looks at two item types that assess academic listening and reading comprehension respectively in a high-stakes English proficiency test. Test takers read a text or listen to a recording and either choose one correct option (Multiple Choice, choose single answer) or more than one correct option (Multiple choice, choose multiple answer). Each reading item is tagged for the following features: academic discipline, CEFR level, length (number of words), number of answer options and question type. Each listening item is tagged for the following features: academic discipline, CEFR level, audio length, word count, accent, number of answer options and question types. The question types for both, reading and listening items fall into seven categories: 1) Main idea or gist, 2) Detailed information, 3) Writer / Speaker’s purpose, 4) Writer / Speaker’s style, tone or attitude, 5) Organization, 6) Inference and 7) Textual value. These variables are analyzed using 100 items per item type. Multiple regression analysis (stepwise method) is employed to determine which of the variables best predicts item difficulty. Preliminary results indicate that item variables affect test takers’ performance differently. For example, the analysis shows that the listening score was significantly influenced by accent and academic discipline. In addition, subsets of items from each item type are analyzed for their lexical complexity looking at the use of academic words, type-token ratio, and lexical density. Furthermore, a qualitative analysis of 50 reading items (Multiple Choice, choose single answer) provides insight into how lexical cohesion contributes to item difficulty. Although lexical cohesion is commonly thought of as existing between and within the sentences of connected discourse, MacMillan (2007) showed that it can also be seen as existing between the reading comprehension task (i.e., question and options) and specific extracts of the reading passage to which they refer. Our findings reveal that the number and nature of lexical cohesive links between the reading passage and the task influences the difficulty value of an item. It was found that the prevalence of literal lexical repetition links (in which the same lexical item occurs in reading passage and task) has very little effect on item difficulty. On the other hand, it was found that prevalence of links that depend on the recognition of semantic relations - such as synonymy, antonymy and hyponymy - does have an effect on item difficulty: the more such links are present the easier the item. The findings of the study will provide insights for test developers in terms of how item variables influence test taker performance on these multiple choice item types.


**PAPER ABSTRACTS**

Youngshin Chi  
**WIDA Consortium**  
**Thursday, April 5**  
10:35–11:05  
Baltimore

**Reviewers’ Perspectives on the Content Validation of an ELL Assessment**

Reviewers understanding of the expectations inherent in language standards is an important consideration for the alignment of assessments to standards. The present study investigated the alignment of the ACCESS for ELLs test and the associated language proficiency standards, and how reviewers understand and transform language functions in the language proficiency standards while they are aligning with assessment items. ACCESS for ELLs is an English proficiency test for English Language Learners (ELL) developed by the World-Class Instructional Design and Assessment (WIDA) Consortium. Various studies have been conducted to determine the accuracy and consistency of the test (Bauman, Boals and Cranley et al., 2007). The validity of the assessment was examined in various ways, including expert review, determining concurrent validity, and calculating correlations between scale scores across items (Porter and Vega, 2007). Alignment between assessment items and WIDA's English proficiency (ELP) standards was conducted to collect the content-related evidence that is part of an overarching validation argument of the test. The methodology of the ACCESS test to ELP standards alignment study is based on Cook's adaptation of Webb's (1997) alignment framework. In this process, groups of trained reviewers examined the relationship between standards and assessment in three dimensions: match, depth, and breadth. The WIDA ELP standards include multiple dimensions: standards, domains, and levels. The standards describe the language of the content areas by providing examples of the types of language activities. Model performance indicators in the standard are functional, measurable indices of the language domains (listening, speaking, reading, and writing) aimed at the targeted age/developmental levels of English language learners. In this study, we evaluated participants’ review process by closely examining their notes and comments on the alignment process. During the alignment, reviewers were encouraged to offer their opinions on the quality of the standards or items by writing comments. After the study, they were asked to provide their opinions on topics on each alignment criterion. The fifty-five reviewers who participated in the alignment study were highly qualified in this process in terms of both content knowledge and prior knowledge of ELLs. They had experience in teaching and are familiar with the ELP standards; however, the alignment process was challenging because they needed to review standards and items in multiple ways. Findings from their debriefing comments on the alignment study show that some reviewers experienced difficulty transforming model performance indicators in the ELP standards when they reviewed each assessment item. As a consequence, some reviewers focused on a simple relationship while others focused on complex relationships between standards and assessments. Though the reviewers had experience with the assessment and the standards, they could experience difficulty identifying the meaning and intention of each indicator and the language difficulty level to assessment items. As an important outcome of this study, we could provide reviewers with the assistance they need to understand what is expected and to share common expectations on the nature of the standards.

Job Schepens  
**Radboud University Nijmegen**  
**Thursday, April 5**  
11:10–11:40  
Cambridge

**An Empirically-Based Measure of Linguistic Distance**

Van der Slik (2010) applied two linguistic distance measures, i.e. the cognate linguistic distance measure (McMahon and McMahon, 2005), and the genetic linguistic distance measure (Cavalli-Sforza, Menozzi, and Piazza, 1994) to explain the variation in oral and written skills in Dutch among 6,000 West-European immigrants in the Netherlands. All immigrants were adult L2 learners of Dutch participating in an official exam on their Dutch language skills. The prediction was that the linguistic distance between their mother tongue (11 West European languages L1s) and Dutch (L2) was an important predictor of their achievements (controlling for other learner and country of origin variables). The explanatory power of the cognate linguistic distance measure exceeded that of the genetic measure substantially. The present study aims at extending our knowledge on the relationship between linguistic distance and language proficiency. We will do so by explaining variance in L2 proficiency scores on the basis of two different linguistic distance techniques: the expert-based, historical-comparative method as used in Gray and Atkinson (2003), and the automatic distance based method as applied in the ASJP project (Brown et al., 2005). We want to argue that the learnability of an L2 is not only a measure to validate linguistic distance measures, but that it can be seen as a way to measure (structural) linguistic distance, as well. We will test our claim that “learnability” is another manifestation of linguistic distance by analyzing a large-scale longitudinal database that contains language proficiency test scores of more than 30,000 immigrant L2 examinees, originating from 89 countries, speaking 35 different Indo-European mother tongues, and combined in 119 mother tongue country of origin combinations. The average proficiency scores of these examinees, broken down by mother tongue, will be used as a yardstick for empirically validating the structural distance between Dutch and the immigrant background languages. The methodological challenge of such an approach is evident. L2 proficiency depends on numerous factors; the mother tongue being just one of them. Many confounding variables such as education, age on arrival, proficiency of an L3, or educational quality of the country of origin can be controlled for and we will do so by applying cross-classified multi-level analyses.
Legislation, Policies and Educational Assessment in a Linguistic Minority Context: Plus ça change plus ça reste la même chose

In 1969, the Canadian Parliament passed The Official Language Act whereby French is recognized as the other official language of Canada. Since then, there is a political mandate that all teaching and assessment programs in elementary and secondary education be offered in both languages across provinces. But in reality, except in Quebec, French is in a minority context and in most cases spoken in scattered and/or isolated communities across Canada. In practice, school achievement tests are developed first in English (the dominant language) then translated and/or adapted into French, assuming that a standardized form in both languages is equally used throughout the country. In this process, the premise is that the validity and reliability of both the original and the translated tests hold the same. However, this is decidedly not the case when we look at the diverse dialectical variations of French spoken in these communities.

On a test administered in multiple languages, examinees perform differently depending on the content and the language of the test item because of the strong association between language systems, their uses, the social status and ethnic background of the test takers (Wolfram 1983 & 2006). Consequently, sociolinguistic factors are a source of differential item functioning (DIF) and produce group differences that are a threat to test validity and fairness. Tests are often written and translated into a single standard French variety which does not take into consideration sociolinguistic variations among these students from the minority closed French communities across Canada. In such cases, the validity and fairness of the measure should be of concern to decision makers and stakeholders alike.

This research examines sociolinguistic biases resulting from test translation and adaptation. Data from the 2007 Pan-Canadian Assessment Program (PCAP) in four academic subjects are used to conduct a series of substantive and statistical analyses to identify any sociolinguistic DIF. The PCAP was chosen because it provides data from all Canadian provinces. Using the English version as a reference, we first studied the linguistic equivalence of the French versions by conducting a meticulous discourse analysis of test items, including translation analyses and identification of language variations within the exams. Second, using a taxonomy developed by — (2007), we classified potentially problematic items into three categories of either syntactic, semantic or pragmatic DIF. Results of this study demonstrate that students’ competence in and mastery of a certain variety of French play a big part in the test outcomes. The International Test Commission (ITC) emphasized the fact that “adaptation needs to consider the whole cultural context within which a test is to be used. Indeed, the adaptation guidelines apply wherever tests are moved from one cultural setting to another” (ITC, 2009). Legally, educational tests in French and English must be accessible to all Canadian children. The process that ensures that this is the case raises fairness issues since the general assumption is a uni-dialectical model for the language used. Between the politics and the legal statuses, the children learning in French are at a disadvantage.

Measuring Content Quality of Writing Performance: Putting Content in Its Place

Measuring Content Quality of Writing Performance: Putting Content in Its Place In assessing the writing proficiency of English as a second language (L2), focus has long been on linguistic aspects of writing performance. However, though the issue of content quality is of fundamental concern in assessing L2 writing, few attempts (e.g., Byrnes, 2008; Ruegg & Sugiyama, 2010) have been made at measuring content quality. In the context of large-scale L2 writing tests, content quality has been conceptualized as two components: topic relevance and topic development, for example, TOEFL iBT and IELTS. Specifically, the two aspects usually involve a logical connection between topic and statement/argument, and an inclusion of supporting details to develop ideas. To measure the content quality of L2 writing performance, the design of the current study incorporated a boundary choice approach based on a range of discourse features (see Fulcher, Davidson & Kemp, 2011). The current study involved two experienced raters applying a proposed rating approach to the rating of writing scripts. 58 students in an English language teaching program in one comprehensive university in mainland China participated in a writing assessment administered by the research team. The study comprised two parts. First, a Binary-choice Boundary definition (EBB) scale development procedure was adopted to develop a rating model, delivered through binary questions in order to group scripts into five levels. Two components were included in the rating model: Topic relevance (Related ideas; Requirements/Expectations of the task) and Topic development (Hierarchy of ideas with facts and supporting information; Introductory framing and concluding phrases; Paragraphing). Second, the resulting scores of the proposed rating approach were compared to scores awarded on the basis of a traditional five-point rating scale developed by the research team with reference to TOEFL iBT writing rubrics and IELTS Writing Band Descriptors (Public version). Raters were then invited to participate in an open interview to canvass their opinions about the new approach. Results showed that, first, the proposed approach, empirically grounded in candidate performance data, was feasible in measuring content quality of writing performance; second, the proposed approach was as effective as the traditional approach in differentiating candidates in terms of content quality, with only four out of 58 candidates being mismatched. The contribution of the study rests on two aspects: on the one hand, it provides empirical evidence for validating content quality for L2 writing proficiency; on the other hand, rich diagnostic information on content quality will be obtained for the learning and teaching of L2 writing.

References
Are Predictor Variables for Item Difficulty of L2 Sentence Repetition Consistent Between English and Chinese?

In developing sentence repetition items for a test of a new language, an accurate model that predicts the expected difficulty of an authentic utterance would be useful. Nair analyzed 18 linguistic and acoustic attributes of item prompts to predict the difficulty of items in an L2-English Sentence Repetition task. The variables included, log number of phonemes, log duration of the spoken prompt, log number of words, and number of syllables. The number of phonemes (r=0.77) and the number of words (r=0.71) had the highest correlations with the observed Rasch item difficulty values. Multiple regression over several variables yielded a slightly higher correlation of 0.83. The present study investigated how well the stronger predictor variables from English would predict item difficulty in L2-Chinese Sentence Repetitions. A total of 465 Sentence Repeat prompts were studied. These items had Rasch item difficulty values established from a sample of about 900 tests administered to learners of Chinese in a field test. Four item variables were examined: 1) Number of Words, 2) Number of Chinese Characters, 3) Total Number of Phonemes, and 4) Average number of Phonemes per Chinese Character. Number of Chinese Characters and Number of Phonemes had strong correlations (r=0.823 and r=0.817 respectively), each explaining about the same amount of variance (~67%) as in Nair’s (2005) study. ‘Number of Words’ was also moderately highly correlated at 0.77. ‘Number of Phonemes per Characters’ correlated only at 0.20. Although the Rasch item difficulty was derived from the learner data, it was hypothesized that native speakers might speak more slowly in response to more difficult items. We analyzed a sample of about 500 tests taken by native Chinese speakers who responded to the same items. Means and medians of four variables were calculated from the native response recordings: Leading Silence (Response Latency), Word Duration, Inter-word Silence, and Total Speaking Time (the sum of the first three variables). Mean Word Duration and Total Speaking Time correlated with the item difficulty values at 0.75-0.76. Inter-word Silence correlated at 0.19 (mean) or 0.30. A stepwise regression process found that including Number of Words, Number of Phonemes and Mean Inter-word Silence produced the most reliable model (r=0.829, Adjusted R = 0.684). However, interestingly enough, including two variables, Number of Words and Number of Phonemes, alone yielded practically the same result (r=0.828, Adjusted R = 0.684). This result suggests that the native performance data added little useful information to predicting the Rasch item difficulty values which were estimated on the learner data. Despite linguistic differences between English and Chinese, the results suggest that the best predictor variables for estimating the difficulty of Sentence Repeat items are the number of words and phonemes in the prompt for both languages. This result appears to lend an empirical support for the common claims about elicited imitation that the amount of the material that a L2 speaker can repeat in real time is a good measure of differing levels of spoken language proficiency. The presentation concludes with implications for writing and piloting sentence repeat items.

Language and Learning: Relationship between English Language Proficiency and Mathematics Achievement of English Learners

In his recent review of English learner policy and practice, Kenji Hakuta (2011) notes that content teachers should be engaged in the development of academic language and he expects “the coordination of English language proficiency standards and content will continue to be an issue ... since there is a strong agreement that all teachers need to think of themselves as teachers of academic English” (p. 171-172). The overall purpose of the research project is to expand our understanding of the relationships in different contexts between academic language proficiency and mathematics content knowledge of English learners - students whose home language is not English and whose English language proficiency hinders their ability to meet expectations for students at their grade level. Structural equation modeling is used to analyze linear relationships across the assessment of English language proficiency in five content areas (mathematics, science, social studies, language arts, and social and instructional language) and four language domains (speaking, writing, listening and reading), and performance on state mathematics tests. The research project aims to compare the stability of these relationships across different grade levels and tiers (i.e. each tier assesses a different range of language proficiency and each student is placed into the tier that best matches his or her proficiency level). Preliminary structural equation models, based on 1,173 English learners from one state in grades 3 to 8 who were all placed in the same tier suggest that English language proficiency latent factors, incorporating all five content areas and all four language domains, influence achievement on a state’s mathematics test. Importantly, the models contain a second higher-order construct which directly influences language proficiency in the productive skills in general content areas, indirectly influences writing skills in technical content areas, and subsequently influences receptive skills and mathematics achievement. The stability of the relationships with data from English learners in grades 3 to 8, but who were placed in different tiers, will be reported, together with implications for assessing English language proficiency and for mathematics teachers.
Jamie L. Schissel  
*University of Pennsylvania*

**Thursday, April 5**  
11:45–12:15  
*Baltimore*

---

**“Test Makers Can’t See Thoughts in Your Head”: Pedagogical Practices of Accommodations with Dynamic Multilinguals**

Commonly known as tools that “level the playing field,” test accommodations in the U.S. present a case of policy-enforced washback. Test accommodations (e.g., interpreters, small group administration, extra time) are designed for dynamic multilingual students to reduce construct irrelevant variance due to language. Most states policies mandate that the accommodations used during testing must also be used during classroom teaching and assessment situations. How this policy informs and is realized in practice is key to understanding the interplay of classroom teaching, formative assessments, and standardized testing situations. In this paper, I argue that—despite earnest efforts to support dynamic multilinguals—attempting to match practices across these situations through the use of test accommodations conflates, misrepresents, and overlooks multilingual competencies. I draw from data collected from an ethnographic case study in grades 3 and 5 in two urban Northeast schools in Pennsylvania. I look at how accommodations are practiced during classroom instruction, formative assessments, and standardized testing situations. The methods used in the study are participant observation field notes, interviews, and classroom artifacts. Informed by literature on classroom ethnographic case studies and experimental design investigations of test accommodations, I developed a recursive, inductive coding schema to identify evidence about how different administrators, teachers, and students make decisions. In addition, and more relevant here, these data demonstrate how accommodations are realized in practice and how these decisions and practices function as a hidden curriculum that promote—implicitly and explicitly—(language) ideologies of idealized standard American English monolingualism. In this analysis, I compare how accommodations are defined in federal and state policies, how teachers and administrators interpret these definitions, and how they are taken-up in practice. By conceptualizing accommodations from policy- and classroom-based perspectives, I aim to broaden the definition and understanding of how policy-enforced washback operates in these classrooms. Further, I focus on how accommodations interact with multilingualism. This complex process includes attempts to recognize multilingualism that ultimately conflate, misrepresent, or overlook these competencies. In both schools, teachers and administrators view their accommodations options, although limited, as useful tools to help dynamic multilingual students. One teacher, for example, employs a classroom-based accommodation: integrated use of Spanish. She tells her students during a 3rd grade standardized testing preparation course that they can think in Spanish. She emphasizes that test makers cannot see the thoughts in their heads, only the answers they write. The teacher presents Spanish as a tool students can use but one that will not be “counted” as a response. Despite the teacher and students’ earnest attempts to support multilingualism, accommodations in this setting fall short of this goal. I argue that this positioning of test accommodations as beneficial legitimizes accommodation practices. This legalization is realized when accommodations, in fact, level the playing field by way of facilitating assimilation to assumed idealized standard American English monolingualism, thereby erasing multilingual competencies. I conclude by discussing how test accommodations can be re-positioned within these contexts to provide learning opportunities for dynamic multilingual students such as teaching and assessing through a multilingual, content-language curriculum.

---

Luke Amoroso  
*Georgetown University*

**Thursday, April 5**  
1:30–2:00  
*Columbus–Savannah*

---

**Analytic and Holistic Rating of Oral Proficiency Interviews with Computational Tools**

The oral proficiency construct underlying the Oral Proficiency Interview (OPI) and the ACTFL Speaking Guidelines used to rate them (Breiner-Sanders, et al., 2000) has proven difficult to identify because of concerns related to the conflation of rating traits and elicitation methods (Bachman, 1988) and also the fact that the OPI is rated holistically. Holistic ratings are not composed of constituent parts, and do not allow the source of rater disagreement to be easily identified, thus making replication of ratings and explication of the construct difficult. The first part of the paper is an attempt to define the OPI proficiency construct through an analysis of the ACTFL Speaking Guidelines and to reliably rate proficiency interviews with the proposed construct. A close reading identified five traits (tense, vocabulary, abstractness, fluency, grammaticality) present at all proficiency levels of the Guidelines. Computer programs were then written to evaluate the degree to which all five traits were present in transcribed Simulated Oral Proficiency Interviews (SOPIs). The SOPIs were all previously rated by two ACTFL-trained raters who agreed on the proficiency level of the interviewee. Parameters for each of the five traits at the different proficiency levels were obtained by taking a subset of the previously rated SOPIs and measuring them for the presence of the five traits. In this way a range for each of the traits was established for each proficiency level. The remaining SOPI transcripts (those not used in the construction of the trait ranges) were then rated automatically according to the established ranges. The automatic ratings based on the five traits agreed with ACTFL-trained human ratings 84% of the time, providing preliminary evidence that the five traits identified are operational in a SOPI rating and are objectively measurable. Complementing the analytic rating described above, a holistic computational rater was also constructed. For the holistic rater, small corpora were constructed for each ACTFL proficiency level, with each corpus composed of OPI transcripts from a single proficiency level. A novel SOPI or OPI transcript was then measured against each of these corpora to determine which it was closest to. This was done by comparing bigram and trigram frequency distributions of a novel OPI transcript to those of each proficiency level corpus, with a rating assigned based on the degree of similarity between the two. The degree of agreement between holistic and analytic ratings, both done automatically, sheds light on the construct of language proficiency that is operational in the OPI rating process and also provides opportunities for feedback to test-takers. Information about which of the five areas were relatively lower than the others will allow examinees to focus on improvement in certain areas as they seek to improve their oral proficiency.
This paper examines the extent to which performance-based characteristics of candidates’ oral reading, such as reading accuracy, rate, and expression, can be used to predict reading comprehension ability in a second language. Previous research in L1 suggests that oral reading features such as reading rate generally correlate with reading comprehension ability in developing readers, in the order of r=0.67–0.80. This relationship is not well established, however, in L2 adult learners (e.g., Samuels, 2006). The Dutch Reading Test assesses oral reading performance in Dutch and provides estimates of reading comprehension ability, placing candidates at “Below A1” or “A1 or above” relative to the reading comprehension scales on the CEFR. Candidates perform a variety of oral reading tasks, reading their responses aloud into a telephone. Candidates are scored on five independently measured subscores: Accuracy, Expression, Reading Rate, Phonics, and Content. The Content subscore is drawn from performance on two reading comprehension tasks. In Fill in the Gaps, a modified cloze task, candidates select the best option out of three which correctly completes a sentence that they read aloud. In the Passage Comprehension task, candidates read a passage and then answer open-ended questions about it (e.g., “How old did Jaap turn on his birthday?”). Scores on these tasks have been aligned with the CEFR subskills related to reading comprehension (Lam et al., 2011). The other orally performed aspects of reading ability (accuracy, expression, rate, phonics) are not described in the CEFR scales. However, they are used to provide further evidence of reading comprehension ability. Multiple regression analysis was used on score data from Dutch Reading Test candidates (n = 2,623) to examine how oral Accuracy, Expression, Phonics, and Reading Rate could be used to predict Content scores. Accuracy and Reading Rate made significant contributions to the models (p < .05); Accuracy alone accounted for 72% of the variance with Content (r=.846), while a model with Accuracy and Reading Rate accounted for 73% of the variance (r=.851). Moreover, Reading Rate and Expression subscores were correlated (r =.90), likely due to the fact that most rapid readers are also able to read smoothly and with appropriate phrasing. Those two traits correlated moderately with Content (r=.66 and .61, respectively), suggesting that in this sample of second language readers of Dutch, Reading Rate and Expression by themselves have comparable usefulness in predicting reading comprehension ability as similar traits when measured in developing readers reading in their first language. As expected, the correlation between Phonics performance and Content scores was somewhat low (r=.53); it is unclear to what extent oral phonics is a threshold ability or whether it might be distinct from reading comprehension ability in adult second language learners. This study concludes by examining the extent to which these results hold when scores are limited to a specific test task — reading short passages out loud — and whether using oral performance scores from this task alone could provide sufficiently reliable predictions of a candidate’s reading comprehension ability.
Overview of the Symposium

Each year, International Teaching Assistants (ITAs) teach tens of thousands of undergraduate courses at North American universities. Assessing the oral proficiency of ITAs to determine if they are qualified to be instructors is an ongoing concern. Many university ITA programs use standardized test scores like the TOEFL iBT to make advanced placement decisions regarding ITAs, and/or local screening exams. Often, cut scores are used to place the ITAs into three categories: able to teach; able to teach while simultaneously taking an ITA preparation course; or not able to teach until successfully completing the ITA preparation course. The ITA preparation course is often a credit-bearing course that involves both ESL instruction and pedagogical training. The issue of ITA assessment has received research attention (e.g., Plough, Briggs, & Van Boon, 2010; Xi, 2008) that has mostly focused on issues of placement decisions and measurement attributes. At the same time, the language and pedagogical training of ITAs has been the subject of extensive research (e.g., Gorsuch, 2006). Much of this research has focused on pedagogical training and language learning for ITA training courses. Considering that many ITA training classes and programs are assessment driven, in that the ITAs in training usually need to pass a teaching simulation performance assessment, one would expect to find a broad range of research on the interplay between assessment and ITA pedagogical practices. However, while valuable research in the area has been conducted (e.g., Hoekje & Linnell, 1994) considering the high stakes nature of ITA assessment, there is a relative paucity of recent research in the area. In an effort to address this gap in the research, the purpose of this symposium is to bring together empirical studies that have investigated: (1) how ITA assessment practices can provide useful and meaningful summative and formative feedback to test stakeholders, and (2) the authenticity of test tasks meant to simulate the real world domain of teaching at the university level. The first paper discusses how a particular ITA assessment utilizes performance data to enhance a secondary use of the test: providing test-takers and their ESL instructors detailed, targeted feedback. The second paper examines the advantages of and challenges involved in using TOEFL iBT speaking scores for ITA certification and placement purposes, including consideration of test taker feedback and ITA ESL coursework. The third paper explores the nature and type of interaction observed in an ITA performance test, questions the test’s interactional authenticity, and discusses the implications for ITA training and placement decisions. The fourth paper examines how different components of the analytic rating scale used in an ITA performance assessment correlate with and predict broader measures of oral proficiency and teaching effectiveness. The symposium concludes with a discussant summarizing and critiquing the four papers. While the papers in the symposium focus on the North American context of ITA assessment, the broader issue of the interplay among classroom teaching and formative and summative assessments (one of the themes of the conference) is relevant for all audiences.
TOEFL iBT Speaking for ITA Certification: State of Practice and Outstanding Validation Questions

The TOEFL iBT Speaking subtest score has been considered for its potential usefulness in ITA certification decision making (Axe & Farnsworth 2007, Farnsworth 2007, Wylie & Tannenbaum 2006, Xi 2007). Until now, however, studies have focused on the measurement attributes of TOEFL iBT Speaking, primarily its dependability (Lee, 2006), concurrent validity (Farnsworth, in progress, Xi, 2007), and exploration of standard setting (Wylie & Tannenbaum 2006). While many institutions have expressed interest in or explored the use of TOEFL iBT Speaking for ITA certification because of the many potential practical advantages of such a policy, many questions remain unanswered about the use of TOEFL iBT Speaking for this new purpose, including critical questions of construct representation and impact. The current paper presents results of a review of current practice in TOEFL iBT Speaking for ITA certification in U.S. and Canadian universities. A survey was administered to ITA program coordinators and administrators that asked about their current practices regarding ITA certification, and specifically about current and future policy regarding TOEFL iBT Speaking use. Results indicate substantial interest in TOEFL iBT Speaking as a component of ITA assessment practices. The paper then discusses the question of authenticity, following the model of Hoekje and Linnell (1994) who examined the SPEAK exam for ITA certification. The TOEFL iBT Speaking task characteristics were compared against real world ITA duties, such as conducting lectures, holding office hours, and supervising laboratory work. A notable area of difference between the characteristics of the test tasks and the target language use tasks included the lack of interactivity of the TOEFL iBT Speaking task. The study then lays out a logical framework, based on a validation and assessment use argument (Bachman, 2010), to highlight needed validation research for TOEFL iBT Speaking in this area, and to illustrate some decision making guidelines to inform best practice for ITA assessment. In particular, this validation and assessment use argument approach will address the promise and inherent challenges in tying TOEFL iBT Speaking scores to a coherent ITA language instructional program and for making local placement decisions.
Investigating the “Interactive” in the “Interactive Performance Test”: A Conversation Analysis Approach

This paper reports on a study that used the methodological framework of Conversation Analysis (CA) to examine the sequential development of talk-in-interaction in the context of the Interactive Performance Test (IPT), which is used by the University of Pennsylvania to assess the speaking abilities of international teaching assistants (ITAs) for placement purposes and which is similar to the micro-teaching activities prevalent in the ITA training classroom. The IPT is a type of “simulation” technique in which an ITA is asked to give a 15-minute prepared lecture on a topic or process specific to his or her field of study to three undergraduates who listen and ask questions throughout. The data set used for this study came from approximately thirty 15-minute IPT performances that were videotaped during the Summer 2008 testing sessions. Although the IPT was designed to engage ITAs and undergraduates in interactional activities that approximate the interaction found in TA instructional contexts, the findings suggest that the participants were orienting to a two part “undergraduate-question ITA-answer” structure rather than the three part “teacher-initiation student-response teacher-evaluation” structure typically found in classroom or lab settings (Tanner, 1991). Furthermore, the analysis revealed an asymmetry of participation: the undergraduates could be viewed as the institutional representative who determined through the composition of their questions how a topic would be attended to and handled and when it should be concluded. Finally, the roles and relationship established through the interaction were that of interviewer/interviewee. Overall, the findings provide a counterargument to the IPT’s assumed “interactional authenticity” and provide evidence that the test tasks, as well as the related classroom training activities, do not engage the ITAs’ language knowledge and ability in the same way that the target language use tasks do. The paper concludes with a discussion of how these findings necessitate: (a) a reexamination of not only placement decisions but also the diagnostic information ITAs are given in their IPT reports, (b) a reconceptualization of the nature of instruction the ITA trainees receive, and (c) a restructuring of the IPT and the classroom training activities that lead to and are influenced by the IPT.

Assessing the Classroom Pragmatic Competence of International Teaching Assistants

It is important for teachers not native to a particular culture to be knowledgeable of the “classroom culture” where the teaching takes place (e.g., Gorsuch, 2003, 2006; Trentin, 2009). This cultural knowledge can be seen as a vital component of pragmatic competence for teachers, and this study investigates how the pragmatic competence of prospective International Teaching Assistants (ITAs) is measured in a performance assessment that serves as both a formative and summative assessment. At Temple University, prospective ITAs who do not reach a prescribed level of oral proficiency in English are required to take a 3-credit course designed to improve their spoken English and their teaching effectiveness. The culminating assignment for this course requires ITAs to successfully complete a performance assessment in which they teach a simulated mini-lesson on a topic in their field of study. The prospective ITAs engaging in this performance assessment are evaluated by raters using a scoring rubric designed to assess the major curricular goals of the class: oral proficiency and teaching effectiveness. Two of the subscales (“Nonverbal Communication” and “Interaction with Students”) in the “Teaching Effectiveness” component of the rubric were designed to assess two components of the ITAs’ pragmatic competence: non-verbal pragmatic competence, and classroom pragmatic competence. This study investigates the extent to which raters’ assessment of a speaker’s non-verbal pragmatic competence and classroom pragmatic competence correlate with and predict broader measures of oral proficiency and teaching effectiveness. The results from 70 of these performance assessments were examined. Correlational analyses and multiple regression were used to examine these issues, and the results suggest that the operationalized measures of pragmatic competence do correlate moderately with other measures of oral proficiency and teaching effectiveness, and while these two subscales are successful at predicting a test-taker’s overall oral proficiency and teaching effectiveness, they contribute little unique variance to the overall scores. The discussion will illustrate how these results can influence the teaching and assessment of ITAs’ teaching pragmatic competence, as well as how the results of the assessment can provide summative and formative feedback for test stakeholders.
Considerations in the Assessment of Young Learners
Tuesday, April 3, 2:30–4:10, Columbus–Savannah

Overview of the Symposium

In part due to increasing globalization, the number of young learners of second/foreign languages (L2/FL) is rapidly increasing worldwide. Despite the urgent need to employ assessments that are “appropriate” for young learners of L2/FL, one cannot underestimate the lack of theoretical and empirical information regarding how best to assess young learners’ L2/FL as well as the effects of assessment on their learning and emotions. It has been suggested that existing models of language assessment for adult learners cannot be applied to children without significant adjustments (Donato, 1998; Oller, 1998). However, the nature of such “adjustments” remains unclear. Cognitive-social developmental and contextual considerations specific to young learners must be seriously accounted for in developing, administering and interpreting assessments (Shohamy, 1998). This symposium sheds light on factors that need to be considered for assessing young learners of L2/FL (defined as L2/FL-learning children up to the elementary school level). Such factors include the learners’ cognitive, social and affective developments, their first language/literacy development, their familiarity with assessment formats and so forth. The four presentations in this symposium deal with different types of assessments (a large-scale proficiency test, task-based assessment, and an observational assessment of early language/literacy skills), cover a wide range of contexts and regions (Europe, Asia, and the U.S.), and employ various data collection and analysis methods (Rasch analysis, surveys, interviews, a picture-drawing technique, teacher observations, discourse analysis, etc). All the presentations, however, address the importance of assessment for learning for young learners and all are concerned with how best to achieve such assessment in their respective contexts. The first author presents an initial validation result of a newly implemented large-scale diagnostic test of foreign language in Hungary. She employs multiple methods to collect validation information from the students and their teachers. The second author focuses on the effect of assessment on young learners’ affective factors, such as motivation and anxiety, in Hong Kong, where young learners are frequented tested. Using interviews and a “draw-a-picture technique,” he uncovers a number of tensions around assessment for young learners in what he characterizes as “a heritage of exam culture”. The third author addresses tasks used as classroom-based assessments and focuses on the developmental differences in young learners’ interactional patterns and skills during paired-task-based assessment. The young learners’ interactions are examined both in their first language (Chinese) and their foreign language (English). Finally, the last paper presents a case of designing a new observational assessment of early language and literacy skills for 3-year-old native and immigrant students in the U.S. (creating both Spanish and English versions). This presentation addresses a number of cognitive and contextual considerations that are needed to develop a valid assessment for early monolingual and bilingual students based in part on survey results from teachers who field-tested the assessment. This symposium not only addresses considerations for developing, administering, and interpreting assessments for an understudied group of learners (i.e., young learners), but also addresses how to conceptualize and implement assessment for learning, which is one of the central topics in language assessment.
Considerations in the Assessment of Young Learners
Individual Abstracts

Marianne Nikolov
University of Pécs
Tuesday, April 3
2:30–4:10
Columbus–Savannah

Young EFL Learners’ Scores and Their Feedback on Listening Comprehension Tasks in a Diagnostic Test Development Project

EFL programs have become typical in grade one of primary school in most European countries, although age appropriate methodology is often an issue. In many contexts teachers are not well trained or unmotivated to teach children. Classroom practice is not always in line with how children learn languages, expectations are often not realistic (Nikolov, 2009a,b; Nikolov & Mihaljevic Djigunovic, 2006; 2011), and testing is a challenge for teachers (Butler, 2009). Therefore, a shift from testing of learning to testing for learning is necessary (e.g., Black & Wiliam, 1998; McKay, 2006; Nikolov & Szabó, 2011; Teasdale & Leung, 2000). The key issue is for teachers to know how to scaffold their pupils’ development with the help of diagnostic tests. The talk presents findings of a large-scale study involving a convenience sample of 2,173 students (ages 6-13) learning English in 161 groups at 26 public schools in Hungary. The study aimed to develop, pilot, and validate new diagnostic tests for young learners in the four basic skills and to place the tests on a scale of difficulty corresponding to the A1 and A2- levels of CEFR by using Rasch analysis. Data were collected with the help of various instruments in the fall of 2010: (1) listening, speaking, reading and writing tasks arranged in 21 booklets each comprising 20 tests (including anchor tests); (2) short questionnaires on each test for pupils and (3) their teachers; (4) a questionnaire filled in by the teacher on pupils’ background data; and (5) teachers were also invited to comment on each task. The talk discusses how the listening comprehension tests worked, how students performed on them, and how these data compare to three criteria on which the students gave feedback: (1) how easy or difficult, (2) how familiar or unfamiliar they were, and (3) to what extent they liked or disliked them. The strongest relationships were found between actual and perceived task difficulty (.637) and ability logit and perceived task difficulty (.629) on the anchor task, whereas other data indicate weaker relationships. The role of various task types will be explored in some detail.

David Carless
University of Hong Kong
Tuesday, April 3
2:30–4:10
Columbus–Savannah

Young Learners’ Perceptions of Their Assessment Experiences: Achievement, Anxiety and Pressure

In a Confucian-heritage setting, such as Hong Kong, young learners are exposed to a diet of frequent testing. How do they respond to the tests and assessments they undertake? To what extent does assessment motivate them to study or accentuate the anxiety involved in learning a language? To what extent does summative assessment dominate at the expense of a more formative role? The aim of the paper is to analyze how a sample of young learners of English in Hong Kong responds to internal school assessment. Data collection from focus group interviews involved 22 groups of 4-5 students in 9 schools and explored principally their experiences of testing and its impact on their study practices and feelings. Complementary data were also collected through ‘draw-a-picture technique’ - an alternative methodology which invites students to express their viewpoints pictorially. 76 drawings from 9 schools were collected via this method. Inductive analysis of the data indicated a number of tensions: the role of parents and teachers in motivating and rewarding students versus pressure and punishment; a sense of achievement versus anxiety or fear; the performative nature of examinations being emphasized more than the long-term mastery of taught content; and assessment as a summative judgment rather than a formative learning process. The main significance of the paper is twofold. First, it lies in documenting the powerful impact of testing on young learners’ experience in an examination-oriented Confucian-heritage setting. Second, it discusses the strengths and limitations of the use of pictorial images as an alternative mode of student expression of their feelings and emotions. Implications for pedagogy include, firstly, the need for assessment of young learners to be handled differently to that of older children. Secondly, it is recommended to emphasize more the formative role of assessment and the development of productive synergies between formative and summative assessment. Thirdly, there is a need for enhanced communication about purposes and roles of assessment; and in particular how assessment might be re-engineered to focus less on performance, and more on learning progress and positive affect.
Considerations in the Assessment of Young Learners
Individual Abstracts

Yuko Goto Butler
University of Pennsylvania
Tuesday, April 3
2:30–4:10
Columbus–Savannah

Developmental Considerations in the Observational Assessment of Three-Year-Olds’ Language and Early Literacy Skills

Increasingly, young children are part of the accountability systems of district, state and federal governments who fund their earliest encounters with formal education. With growing numbers of young students arriving at preschool in the U.S. with a language other than or in addition to English there is also pressure to know more about the language and literacy experiences of these children in their Early Childhood Care and Education (ECCE) settings in order to best meet their linguistic needs. However, the assessment of such young children poses unique challenges that must be addressed in order to provide reliable, valid, and fair information (National Association for the Education of Young Children and National Association of Early Childhood Specialists in State Departments of Education [NAEYC & NAECS/SDE], 2003). This presentation describes how the design of a new observational assessment of the language and early literacy skills of three year olds was driven by developmental considerations. The assessment is intended for use with both monolinguals and dual-language learners. Separate English and Spanish language versions were created. The specifications were guided by the child development literature and best practices guidelines (e.g., NAEYC & NAECSS/NAECS, 2003; Shepard, Kagan, & Wurtz, 1998). Development at this young age is rapid, uneven, and highly influenced by experience and environment. Young children’s performance can be strongly influenced by anxiety, fatigue or poor health, as well as by familiarity with test-taking situations. Cognitive abilities are still developing and constrain memory and processing speeds. Young children also experience short-term motivation issues (Hasselgreen, 2005). ECCE assessment should address not only these processes by which children develop and learn, but also the individual and cultural differences among children (e.g., Bailey, 2008; McKay, 2006). The new assessment has a number of features that take account of these developmental constraints: 1) naturally occurring activities throughout the day are observed; 2) observations of each skill/behavior are made on multiple occasions; 3) observations are made across different contexts; 4) ratings reflect the degree of assistance a child requires to demonstrate a skill/behavior. Results of surveys administered to field-test users provided feedback on the developmental appropriateness and usefulness of the assessment.

Interactional Development in Oral Paired-Assessments Among Young Learners of Foreign Language

Despite the growing popularity of task-based language teaching (TBLT) in foreign language (FL) education at the elementary school level, it remains unclear how best to evaluate young learners’ FL abilities using tasks. The notion of assessment for learning has been promoted among teachers, and the role of classroom-based assessment is increasingly important in practice. Some have argued that the assessment format and procedures need to match young learners’ social-cognitive and linguistic developmental levels. The present study therefore seeks to understand the developmental differences in interactional patterns and features among elementary school students during task-based language assessments. By doing so, this study aims to address the possibilities and limitations associated with task-based paired assessments (paired-TBLA) among young learners of FL. The participants were 70 students (24 4th, 22 6th and 24 8th grade students) and their teachers in a FL program in China where TBLT was mandated. The students were engaged in a set of decision-making tasks (used to do assessments) with their peers, both in their first language (L1, Chinese) and their second language (L2, English). After each task, the teacher evaluated each student’s performance and the students self-assessed their performance. All the interactions were videotaped and transcribed. The study found that the overall interactional patterns were different across grade levels; the older they became, more collaborative interactional patterns were observed during paired-TBLA in both languages. Compared with younger grade levels, the older students also showed more elaborate mutual topic development and more variability in their turn-taking patterns. In contrast, younger grade students appeared to still be developing their interactional skills and strategies in both their L1 as well as their L2. Variability among dyads with respect to the differences in interactional patterns between L1 and L2 were also generally more evident among the younger students. The potential for eliciting a wide range of interactional functions, which is the primary advantage of introducing paired-assessment, may not apply well to younger learners. At conclusion of the study, the implications of its results for understanding interactional competence among young learners of FL are discussed.

Alison L. Bailey
University of California, Los Angeles

Huan Wang
CTB/McGraw-Hill

Lorena Houston
CTB/McGraw-Hill
Tuesday, April 3
2:30–4:10
Columbus–Savannah

Yuko Goto Butler
University of Pennsylvania
Tuesday, April 3
2:30–4:10
Columbus–Savannah
Overview of the Symposium

It is now ten years since the publication of the Common European Framework Reference for Languages (CEFR) (Council of Europe, 2001), and its spread has been remarkable. It is used in so many diverse contexts and domains in Asia, South America, Australia and beyond that the term ‘European’ is no longer adequate. However, despite the influence of the CEFR on the development of teaching materials, classroom teaching methods, and curricula for teacher training around the world, there are serious issues that need to be addressed if the CEFR is to be used as a basis for developing language assessments, as a standard to which a wide variety of language assessments may be linked, or as a criterion for the recognition and accreditation of language programs and language tests. While the intended beneficial consequences of basing assessments on the CEFR are laudable, claims that it can provide a “common metric” for certification of language ability across different languages, countries, groups of test takers, language learning contexts and intended uses need to be supported by more than rhetoric and policy statements. In short, there is a need to provide a rationale, supported by evidence, that the desired beneficial consequences can be and are being achieved by the assessments in development or in use that are based on and geared to the framework.

The four papers in this symposium will challenge the assumptions of the widespread use of the CEFR in assessments in several ways, as follows. The linking of tests to the CEFR assumes that they are compatible in their constructs and the decisions that are to be made from their use. The CEFR defines its construct in performance terms, which raises issues of sampling adequacy and task representativeness, and the exact nature of the underlying knowledge which is the basis for the performance. Its particular conceptualization of ‘communication’ builds on structuralist notions of language, and ignores recent discussion of language and language use in terms of multilingual competence and trans-languaging. Its instrumentalist orientation assumes a set of goals for language education that represent a narrowing of the possible educational meanings of language learning for students, particularly at school. Despite this, assessments are subject to overwhelming pressure to demonstrate their interpretability in terms of CEFR levels. Policies in education, immigration and citizenship specify CEFR levels which must be reached or targeted. Market forces also dictate that achievement on popular tests should be easily interpretable in terms of the CEFR. The CEFR promises the advantages of the Euro in terms of a currency that is easily convertible, but the contexts and goals of language learning and the meaning of language assessments are unfortunately not like currencies. Finally, the papers explore the claims and impacts of the CEFR in the areas of both immigration and education, considering the compatibility of the CEFR with the goals of education in differing specific contexts. The implications of the papers for the validity of the use of the CEFR in assessment will be drawn together by the discussant.
Validity, Values and Policy: The Case of the CEFR

Increasing attention has been given to test consequences in discussions of test validity, but less attention has yet been paid to the way in which test constructs and the uses of tests involve questions of social and political values. Messick, in his famous discussion of validity, was clear about the way in which test constructs embody values, but naively assumed that it was up to the test developer to articulate and defend the value implications of test constructs. The case of CEFR raises difficulties in relation to this view. First, tests which are linked to or based on the CEFR must share their construct with the CEFR, and thus will necessarily reflect the values of the CEFR. What are they? The paper argues that the values implicit in the CEFR reflect the instrumentalist view of language, language learning and language use characteristic of the communicative movement in language teaching which is in turn itself a reflection of the demands on education associated with globalization. They also reflect the assumed privilege of the native speaker. The paper considers alternative views of the purpose of language learning which are not reflected in the CEFR construct. Second, the paper also examines the way in which CEFR levels are specified as part of policies and legislation, especially in education, but also in policies on immigration and access to citizenship. This requires that any test used to certify achievement in these terms must itself share the construct of the CEFR. The CEFR thus has the power to impose a set of values which are potentially not best suited to the diversity of language learning and language use contexts in which the need for tests arises. Examples are given from education, workplace, immigration and citizenship contexts in several countries.

How Do Different Language Frameworks Impact Language Assessment Practice?

Developers of large-scale language proficiency tests have drawn on a variety of theoretical frameworks of language to define what they want to assess. However, “linking” tests that have been informed by different views of the construct to be measured may be problematic. This is because such linking requires that the two tests measure the same construct, and that the decisions to be made and the consequences of these decisions are similar. Many current linking activities do not satisfy these basic requisites, so it is imperative for language testers to address these.

The different language frameworks that have informed large-scale language tests in the past half-century fall into two general types: “can do/performance” and “ability”. An assessment use argument (Bachman & Palmer 2010) is used to analyze how these two approaches to defining language differ in terms of the claims they make about interpretations, decisions, and consequences.
Framing the Development of Advanced FL Literacies:
Challenges for the CEFR

Like most current frameworks guiding language policy, including language education policy, the CEFR asserts a ‘communicative’, meaning orientation. However, the CEFR’s conceptualization of ‘communication’ remains remarkably additive and presents a notion of language use that builds on entirely structuralist notions of language. That conceptual mismatch affects (1) the development of a principled way of linking language use to formal features of language that have been functionally described; (2) the possibility of orienting instruction on the basis of a user and use orientation – structuralist approaches being the default position; and (3) the conceptualization of instructed language development that would assure a language use orientation and language learning in relation to each other. All three concerns have grave consequences when one considers the varied educational settings, levels, and preferred educational goals for language learning, including the nature of language learning (e.g., FL vs. SL, most especially, lingua franca English language learning); and the learners’ likely characteristics and interests. The paper discusses these matters from the standpoint of higher education FL learning in the U.S. that must espouse a type of language education that is humanistically oriented in its interest in oral and written texts; learner- and learning-oriented in its focus on language learning as learning to make meaning-oriented choices; and content-oriented in its understanding of emergent multilingual languaging as a semiotic tool for situated multilingual and multicultural knowing. While the CEFR is not at this point a dominant force in U.S. higher education thinking, its undeniable conceptual proximity to dominant ‘communicative’ frameworks in the U.S. (Proficiency, Standards) further complicates much-needed educational change. The paper discusses these matters from five interrelated perspectives: (1) countering additive, formalist thinking by conceptualizing meaning-making in multiple languages as choice; (2) overcoming the constraints of communicative language teaching through a textual focus; (3) rethinking emerging language development as meaning-making in diverse content areas; (4) conceptualizing the social and cultural through diverse genres; and (5) selecting and sequencing instruction with a view to long-term development toward advanced forms of literacy.

In the Name of the CEFR

In Europe, the main role of the CEFR is to act as a standard against which to assess people’s competence. This use is very problematic especially in countries such as Italy, where there is a tremendous lack of competence in and awareness of the theory and practice of language testing and assessment. In surveys conducted over a number of years, we have found not only that teachers have a distinct lack of professionalism in the area of language testing and assessment, but that terms such as validity, reliability, impact, and ethicality in relation to language assessment are simply unfamiliar. This lack of competence has a strong impact in the education sector, where tests are increasingly used, and also in the broader social context. “In the name of the CEFR” every decision is permitted, but the reality is that the CEFR levels are simply used as a relatively meaningless label, with little consideration of the impact and consequences of their use. For example, in schools reference to the CEFR levels is made in the testing of both native and non-native students; and CEFR levels are used as a benchmark as part of policies governing the right of immigrants to gain long term residence and in the recent “integration agreement”. In neither case, however, is attention given to how the CEFR is being interpreted, how tests are being implemented and how students are being assessed, and to the consequences of the use of test results. Examples will be given from tests in Italy, both those used at school and at university level, and those used for immigrants.
Implementing Innovation in Language Testing Theories and Practices: The Role of Consequences in a Validity Argument
Thursday, April 5, 3:30–5:15, Columbus–Savannah

Overview of the Symposium

With respect to linking assessment and learning, the inclusion of consequences within the scope of validity inquiry is perhaps the most significant innovation in language testing theory over the past twenty-five years. The link is made in Messick’s (1989) definition of validity as “an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores” (1989, p. 13). The judgment is to be made on the basis of an appraisal of 1) evidence concerning test interpretation and use, as well as 2) consequences associated with test interpretation and use. The significance of Messick’s perspective on validity, which did not fundamentally change but rather elaborated on previous work, is evident in language testing, and in professional standards in educational and psychological measurement (e.g., AERA, APA & NCME, 1999), although it is not without critics. The consequential basis refers to “the appraisal of the value implications of the construct label, of the theory underlying test interpretation, and of the ideologies in which the theory is embedded” as well as “the appraisal of both potential and actual social consequences of the applied testing” (Messick, 1989, p. 20). Both value implications and social consequences, in the large majority of language testing, entail links between assessment and learning. The challenge set by Messick was for test developers, and researchers to analyze the values and social consequences underlying test interpretation and use in part through analysis of these links. The complete validation program implied by Messick’s definition raises challenges for test developers and researchers who need to carry out validation research and build validity arguments for real tests. In response to these challenges, researchers have attempted to frame a process of validation that takes into account Messick’s definition, but which is specified with sufficient precision to guide an operational research program. Central to these proposed solutions is the use of a principled means of specifying arguments pertaining to test interpretation and use (e.g., Kane, 2006; Bachman & Palmer, 2010); a validation program follows from the manner in which the argument is specified. This approach has been used in language assessment (e.g., Chapelle, Enright & Jamieson, 2008), but not extensively. Moreover, as McNamara and Roever (2006) point out, it is not entirely clear how the relevant value implications and social consequences are to be incorporated into a validity argument. The proposed symposium will address the question of how the consequential basis for validity inquiry can be appropriately considered in a validity argument. The first paper will review the history of this dispute and will propose an argument-based framework for the role of consequences in validation. The second paper will describe specific claims and warrants used to include consequences in arguments about language tests, and the third will describe a test framework revision aimed at including consequences in the validity argument. Two discussants will provide commentary from the perspectives of analysis of values and effects of such assessment arguments on learning, respectively.
Implementing Innovation in Language Testing Theories and Practices:
The Role of Consequences in a Validity Argument

Michael Kane
Educational Testing Service
Thursday
April 5, 3:30–5:15
Columbus–Savannah

The Role of Consequences in Validation

The role of consequences in validity has been quite contentious over the last 30 years, but the two sides in the argument often speak past each other, because they emphasize different kinds of consequences and focus on the validation of different parts of score-based decision procedures. This paper will attempt to clarify what’s at stake in these disputes by delimiting the kinds of consequences that are most relevant to validity, and will assign consequences different roles in validating score interpretations per se and score-based decision procedures. The consequences that are most relevant to validity are determined by society, through legal developments, legislation, and institutional policies and practices. At least three kinds of consequences are currently recognized as being relevant to the evaluation score-based decision procedures: the intended outcomes (e.g., increased productivity), fairness, and student/client outcomes. Decision procedures, and in particular score-based decision procedures, are evaluated mainly in terms of their consequences, and without an analysis of consequences, a score-based decision procedure cannot be fully evaluated, or validated. An underlying interpretation of test scores in terms of traits, competencies, performance domains, etc. can be evaluated to a large extent, without giving much attention to consequences; in evaluating such interpretations, negative consequences count against the validity of the interpretation only if they are found to be due to some problem with the specification of the interpretation (the construct definition) or with the assessment (e.g., construct underrepresentation or construct-irrelevant variance). In discussing consequences, Cronbach tended to focus on test uses and saw negative consequences as playing a major role in the validation of the score-based decision procedure, even if the negative consequences were not due to defects in the construct definition or the assessment procedure. Messick focused on score interpretation and saw negative consequences as playing a role in the validation of the construct interpretation only if the negative consequences were not due to defects in the construct or the assessment. They were both right.

Lyle F. Bachman
University of California, Los Angeles
Thursday
April 5, 3:30–5:15
Columbus–Savannah

Justifying the Use of Language Assessments

We generally give a language assessment because we have some decisions that need to be made, such as selecting students for a program, placing them at appropriate levels for instructional purposes, determining their progress in the program, or providing formative feedback on instruction and learning. The use of the assessment and the decisions that are made will have consequences for stakeholders—the various individuals, programs, or institutions who will be affected in one way or another by the use of the assessment and by the decisions that are made. We need to be able to justify the decisions we make on the basis of test scores, as well as the consequences of these so that we can be accountable to the stakeholders. In order to justify using the results of a language assessment for making decisions, we need to provide a rationale for linking students’ performance on the assessment with the intended use, or decision. This rationale is provided by an assessment use argument (AUA). Through the process of assessment justification—articulating an AUA and collecting evidence in support of this—we provide the information we need to be accountable to the individuals who are affected by the assessment and the way it is used.
Implementing Innovation in Language Testing Theories and Practices: The Role of Consequences in a Validity Argument

Individual Abstracts

Brent Bridgeman
Educational Testing Service

Carol A. Chapelle
Iowa State University

Thursday
April 5, 3:30–5:15
Columbus–Savannah

Including “Consequences” in the Revised Interpretive Argument for the TOEFL iBT®

The current TOEFL interpretive argument was developed to guide validation research that would support the following conclusion: “The test score reflects the ability of the test taker to use and understand English as it is spoken, written and heard in college and university settings. The score is useful for aiding in admissions and placement decisions and for guiding English-language instruction” (Chapelle, Enright & Jamieson, 2008, p. 321). The second part of this conclusion refers to test use and consequences, both of which are associated with a “utilization” inference in the interpretive argument. When this argument was formulated, the TOEFL iBT was new and so opportunities for gathering data on consequences of test use had not yet been available. However, several years later as the interpretive argument is being revised, the need exists to specify more precisely the intended consequences of TOEFL iBT use in a manner that will guide research on consequences. This paper reports on this aspect of the TOEFL iBT framework updating project. We will explain the problem encountered in an interpretive argument that subsumes multiple facets of use and consequences within one inference. We will explain how we moved beyond this problem by describing how guidelines from scholarship in this area (e.g., Kane, 2006; Bachman and Palmer, 2010) contribute to the new formulation of an inference concerning “consequences” in the new interpretive argument. We will outline the draft claims, assumptions, and types of backing that are suggested as well as comment on the issues raised in adding dimension to the interpretive argument in this context. Two discussants will each offer commentary on the three papers. One of the discussants will examine the extent to which Messick’s conception of values is taken into account within a validity argument framework, and a second will comment on the implications of such assessment use arguments, interpretive arguments and validity arguments for teaching and learning when consequences are explicitly included.
Development of an Assessment of High Frequency English Vocabulary for Young English Language Learners

This poster reports on findings from a National Institute for Child Health and Human Development (NICHD)-funded study on the assessment of vocabulary in young Spanish-speaking English-language learners (ELLs). The goal of the study was to develop an assessment that can measure a variety of word types in young children. Because teachers cannot possibly teach all the high frequency vocabulary that students need to learn, it is crucial to learn more about the types of vocabulary (e.g. words distinguished by cognate status, part of speech, conceptual complexity) that ELLs have not easily acquired through exposure to English. Most existing standardized vocabulary measures (e.g. the Peabody Picture Vocabulary Test) entail matching target vocabulary to pictures. While these assessments may be adequate for assessing student’s knowledge of concrete vocabulary words, they are problematic for assessing vocabulary that is more conceptually complex because it is difficult to create images for these words (e.g., responsibility, conscious). The goal of the study was to develop a valid and reliable measure of vocabulary that could assess different types of words, including conceptually complex and concrete words as well as cognates and non-cognates. Additionally the goals were to develop an assessment that could be group administered and used to assess vocabulary knowledge in both dual language learners and English proficient students. The assessment consisted of four sub-tests of 74 target words total, selected because they were highly frequent in grade level text as indicated by the Educator’s Word Frequency Guide (Zeno, Millard, Ivans, & Duuvari, 1995) and had been instructed as part of a vocabulary intervention. They were coded for Spanish/English cognate status and conceptual complexity, and stratified on both variables prior to random selection, allowing for even distribution of cognates and non cognates, and of conceptually simple and complex words. The assessment was group administered to 187 second grade ELLs in transitional bilingual education programs in a large urban district in the Southwest. On their assessment students saw a picture that demonstrated the meaning of the target word but also listened to the teacher read a one-sentence child-friendly definition of the word. They were asked to bubble in one of three words that appeared in large print under the picture. One word was the target and the other two were distracters matched to the target on part of speech, semantic congruence, morphology, word length and appearance. An example of one item is: Daniel notices that his dog is eating his ice cream. When someone notices that something is happening, are they conscious, confused, or current? Listen again and bubble in the word that means notices that something is happening: conscious, confused, or current. To assess the concurrent validity of the assessments we examined its correlation with two standardized measures of vocabulary—Word Knowledge Subtest of the Gates-MacGinitie and the Oral Vocabulary Subtest of the Test of Language Development. Preliminary findings are encouraging. There were significant correlations between our researcher-developed assessment and the Gates-MacGinitie Word Knowledge test (r=.60, p<.01); and the TOLD Oral Vocabulary subtest (r=.66, p<.01).

In-Service Training Course in Classroom-Based Language Assessment

Assessment in general, and language assessment in particular, is typically not included as part of teacher training at universities or in pre-service training. This is a common phenomena around the world and also in Bavaria, Germany. This lack of any pre-service training in language assessment makes it necessary to offer in-service teacher training on this topic. In Bavaria a five day course on classroom language assessment has been offered for teacher trainers of English at the Academy for In-service Teacher and Management Training in Dillingen. The Academy offers courses for teachers of all school types and subjects and this course was part of a larger program for English teacher trainers for primary and secondary schools. The course objective was to prepare the teacher trainers to provide in-service training on classroom language assessment to English teachers in their school districts. The course content was based on an application of an assessment use argument to the development and use of classroom assessments. Activities included presentations by the course leaders, discussions, and hands-on small group activities. One of the outcomes of the course was a set of prototype assessment specifications with example assessment tasks. The poster session describes the planning, preparation, and delivery of the course, along with focus on in-service training in classroom language assessment for English teachers in Bavaria.
Beverly Baker  
McGill University

Michel Laurier  
Universite de Montreal

Carolyn Turner  
McGill University

May Tan  
McGill University

Maria Lourdes Lira Gonzales  
Universite de Montreal

Ainsworth Judith  
Universite de Montreal

Development of a Computer-Based Formative Assessment Instrument for Nurses Using ESL in Quebec, Canada

Due to globalization and other factors, healthcare access for linguistic minorities is increasingly becoming an important issue. Research in the medico-linguistic arena has shown that communication gaps can occur when patients and caregivers do not share the same first language. This can consequently compromise the quality of healthcare that is provided (Candlin & Candlin, 2003). As members of an inter-university research group funded by Health Canada, our current research addresses problems of healthcare delivery in the face of language and cultural barriers, especially with regard to the English-speaking minority population in Quebec, Canada. Phase 1 of this research project empirically identified 19 relevant speech events (nurse-specific oral interactions) through a literature review, a focus group with expert Quebec nurses and a survey with an additional 133 Quebec nurses (Turner & Isaacs, 2010; Isaacs et al., 2011). Two categories of speech acts emerged that could lead to the design of assessment tasks for formative purposes: “emotional aspects of caregiving” and “factual aspects related to the patient’s medical condition.” These findings provided an empirical basis for deciding which speech events to focus on in L2 training. Our poster presentation describes Phase 2 of this project: the development of a computer-based formative self-assessment instrument focusing on nurse interational competence. This phase comprises the operationalization stage of the project—a mixed methods approach to test development (i.e., a sequential exploratory model—Creswell & Plano Clark, 2011). The stages of instrument development are the following: a) initial task writing processes based on the 19 speech events which were identified in Phase 1; b) verification of task content by expert nurse informants; c) creation of computer-based pilot tasks; d) piloting of tasks with a participant group of senior nursing students; e) analyses of pilot tasks; f) fine tuning and creation of formative assessment; and g) implementation for wider use by Quebec nurses and nurses-in-training. The implementation of the instrument will provide data for continued validation and re-examination for its usefulness. This instrument is not intended for certification purposes but as a focused tool for nurses’ language training efforts. We believe that this type of instrument will help nurses to improve their language skills to meet their professional skills, either in a self-study context or as a complementary resource for teachers in a classroom setting. This poster presentation relates directly to this year’s theme at LTRC, specifically in the area related to “Key research issues and practices in language assessments across diverse contexts and domains.” Expertise in language assessment is applied here to the goal of providing free and user-friendly access to formative assessment tools that encourage language learning success for this professional domain. Future research could expand these tools to other health practitioner domains.

Bokyung Cho  
Korea Institute for Curriculum and Evaluation

Investigating Factors Affecting Students’ Performance on an English Achievement Test

The purpose of this study is to analyze factors that can influence Korean middle students’ English achievement tests in the longitudinal aspect. Since English is the most important foreign language in Korea, the society and parents have demanded the educational system to develop students’ English abilities effectively. Researchers and teachers have tried to find the most successful ways of how to improve students’ English abilities and help them to achieve their goals on the tests. Researchers have investigated factors which can influence students’ improvement in English, such as teaching methods, educational materials, motivations, students’ learning styles, strategies, etc. In fact, researchers have obtained desirable output from their research by employing the cross sectional analysis. This research method certainly provides powerful information at a specific point of time. However, it does not present appropriate information on changes of variables on the time continuum. For instance, a cross sectional study cannot provide information on participants’ emotional changes affecting their performance on English tests. During middle school years, adolescences can experience emotional changes which can impact their learning styles and academic achievement (Newby et al., 2001; Applebaum, 1995). Based upon this premise, this research began to investigate how changes of students’ attitudes toward their English teachers’ teaching approaches affect their learning behaviors and achievement on English tests on the chronicle continuum. This study analyzed a panel data which Korea Educational Development Institute has collected since 2005. Over 6,000 middle school students participated in the data collection and have still responded to identical survey items and taken achievement tests every year. This study analyzed the accumulated data for three years from 2005 to 2007 because this data is more stable than other years and maintains appropriate sample size. Researchers used SPSS 17 to do data mining and AMOS 7.0 to analyze each variable’s changes on the time continuum. The used model for the investigation was Multivariate Latent Growth Model. According to the analysis, variables that this study analyzed have linearly changed. Students’ attitudes towards their English teachers’ teaching approaches powerfully influence their learning behaviors, self-regulated study, and English achievement.
POSTER ABSTRACTS

Sun Joo Chung
University of Illinois
at Urbana-Champaign

Validiy Issues in the English Placement Test: Future Direction of Online Testing

The UIUC English Placement Test (EPT) is administered to incoming international undergraduate and graduate students who do not meet the campus English proficiency standard. The test is administered to students once they arrive on campus prior to the start of each semester. The test is anchored to the test specifications that have evolved over several years to meet changing mandates and agenda settings. For several years, the test has changed its format in order to improve consistency and accuracy. In 2008, the undergraduate population demographics changed drastically due to a sudden increase in the number of international students. The surge of international undergraduate students brought about new challenges to test administrators and test score users which consequently led to a review of the validity of the test. The surge, in addition to questioning the test validity, led to the development of a new test delivery mode and administration methods. The new mandates from the surge, mainly from undergraduate advisors and admission officers, barely existed in the past. Before the surge, the test functioned as a peripheral assessment tool to provide additional support to the international undergraduate population. Since the surge, the test is playing a more crucial role in the undergraduate academic setting. Because of the test’s increasing influence on the international undergraduate population, the test administrators are faced with the demand of developing a pre-arrival test delivery format, an online version of the English Placement Test. The goal of the pre-arrival online test is to have newly admitted students take the test during the summer prior to their enrollment at the university. This external mandate motivated a research project to investigate ways to create online assessment tools so that students can take the EPT before coming to campus. In this study, we investigated validity issues of the EPT by examining layers of test specification structures based on Fulcher and Davidson (2009). We also looked into the history and policy changes of the test that was reflected in the previous test specifications and test revision. Moreover, we closely examined the needs and advantages of adopting online assessment tools for international students by modified use of online instructional tools. Findings from the test specifications and policy review show that the external mandate was a crucial factor that affected the various format changes of the test. Also, the current policy redirects teachers and advisors to use test scores differently than the previous test setting. Online assessment tool has merit on test administrations but there is a possible threat to test validity due to technical difficulties and issues related to test security.

Youngshin Chi
University of Illinois
at Urbana-Champaign

Measuring Speaking Ability through Elicited and Spontaneous Speech with Automated Scoring

As technology and travel make the world smaller, the need to assess speaking ability becomes increasingly important. Unfortunately, traditional speaking assessments like oral interviews are often impractical due to the expense and time needed to for administration and scoring. The challenge is how to assess spoken language in an efficient and cost effective manner. One possibility to reduce costs is to administer speaking tests with alternative formats. For example, with elicited oral responses or elicited imitation, examinees hear a sentence as the prompt, and then repeat what is heard. When the sentence is sufficiently long, the examinee needs to comprehend the utterance and store it in working memory and then reconstruct and repeat it (Erlam, 2006). Since the examinee response is defined a priori, a grader can simply determine if the words uttered are correct or incorrect which can result in higher rates of scoring reliability. Another way to reduce the cost associated with rating speech is the implementation of Automatic Speech Recognition (ASR) software. ASR is based on pattern recognition and as such requires predictable speech patterns in order to recognize what is spoken (O’Shaughnessy, 2008). With dictation software, the patterns become restricted as the ASR is trained to an individual’s voice. When it is not possible to restrict recognition to a specific individual’s voice in this manner, restricting the content being processed can improve performance. Since elicited imitation uses pre-defined sentences, ASR can reliably score these utterances (Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008; Bernstein, Van Moere, & Cheng, 2010) Other types of alternative speaking prompts result in spontaneous speech that is difficult for ASR to recognize because of the challenging of defining all possible vocabulary domains a priori. However, an ASR can also recognize fluency features such as pauses, rate of speech, duration of bursts, etc., and these fluency features can easily be used in combination with the accuracy measures from elicited imitation to produce a more accurate measure of speaking ability (Higgins, Xi, Zechner & Williamson, 2010). This poster will examine how speaking ability was determined by using a conjoint measure of 1) ASR-scored elicited oral response test to determine accuracy and 2) ASR-scored spontaneous speech to determine fluency features. It will use data in existence from a battery of tests that were administered to 180 students in an intensive English Program. There seems to be great promise in using ASR to score a two-pronged assessment that includes fluency and accuracy.

Troy Cox
Brigham Young University

Validiy Issues in the English Placement Test: Future Direction of Online Testing

The UIUC English Placement Test (EPT) is administered to incoming international undergraduate and graduate students who do not meet the campus English proficiency standard. The test is administered to students once they arrive on campus prior to the start of each semester. The test is anchored to the test specifications that have evolved over several years to meet changing mandates and agenda settings. For several years, the test has changed its format in order to improve consistency and accuracy. In 2008, the undergraduate population demographics changed drastically due to a sudden increase in the number of international students. The surge of international undergraduate students brought about new challenges to test administrators and test score users which consequently led to a review of the validity of the test. The surge, in addition to questioning the test validity, led to the development of a new test delivery mode and administration methods. The new mandates from the surge, mainly from undergraduate advisors and admission officers, barely existed in the past. Before the surge, the test functioned as a peripheral assessment tool to provide additional support to the international undergraduate population. Since the surge, the test is playing a more crucial role in the undergraduate academic setting. Because of the test’s increasing influence on the international undergraduate population, the test administrators are faced with the demand of developing a pre-arrival test delivery format, an online version of the English Placement Test. The goal of the pre-arrival online test is to have newly admitted students take the test during the summer prior to their enrollment at the university. This external mandate motivated a research project to investigate ways to create online assessment tools so that students can take the EPT before coming to campus. In this study, we investigated validity issues of the EPT by examining layers of test specification structures based on Fulcher and Davidson (2009). We also looked into the history and policy changes of the test that was reflected in the previous test specifications and test revision. Moreover, we closely examined the needs and advantages of adopting online assessment tools for international students by modified use of online instructional tools. Findings from the test specifications and policy review show that the external mandate was a crucial factor that affected the various format changes of the test. Also, the current policy redirects teachers and advisors to use test scores differently than the previous test setting. Online assessment tool has merit on test administrations but there is a possible threat to test validity due to technical difficulties and issues related to test security.

Measuring Speaking Ability through Elicited and Spontaneous Speech with Automated Scoring

As technology and travel make the world smaller, the need to assess speaking ability becomes increasingly important. Unfortunately, traditional speaking assessments like oral interviews are often impractical due to the expense and time needed to for administration and scoring. The challenge is how to assess spoken language in an efficient and cost effective manner. One possibility to reduce costs is to administer speaking tests with alternative formats. For example, with elicited oral responses or elicited imitation, examinees hear a sentence as the prompt, and then repeat what is heard. When the sentence is sufficiently long, the examinee needs to comprehend the utterance and store it in working memory and then reconstruct and repeat it (Erlam, 2006). Since the examinee response is defined a priori, a grader can simply determine if the words uttered are correct or incorrect which can result in higher rates of scoring reliability. Another way to reduce the cost associated with rating speech is the implementation of Automatic Speech Recognition (ASR) software. ASR is based on pattern recognition and as such requires predictable speech patterns in order to recognize what is spoken (O’Shaughnessy, 2008). With dictation software, the patterns become restricted as the ASR is trained to an individual’s voice. When it is not possible to restrict recognition to a specific individual’s voice in this manner, restricting the content being processed can improve performance. Since elicited imitation uses pre-defined sentences, ASR can reliably score these utterances (Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008; Bernstein, Van Moere, & Cheng, 2010) Other types of alternative speaking prompts result in spontaneous speech that is difficult for ASR to recognize because of the challenging of defining all possible vocabulary domains a priori. However, an ASR can also recognize fluency features such as pauses, rate of speech, duration of bursts, etc., and these fluency features can easily be used in combination with the accuracy measures from elicited imitation to produce a more accurate measure of speaking ability (Higgins, Xi, Zechner & Williamson, 2010). This poster will examine how speaking ability was determined by using a conjoint measure of 1) ASR-scored elicited oral response test to determine accuracy and 2) ASR-scored spontaneous speech to determine fluency features. It will use data in existence from a battery of tests that were administered to 180 students in an intensive English Program. There seems to be great promise in using ASR to score a two-pronged assessment that includes fluency and accuracy.
Investigating the Validity of a Test of Academic Productive Grammar

In this study, a test of academic productive grammar was developed and its test validity was examined under Bachman's Assessment Use Argument framework. This test was developed with an intention to supplement an ESL writing placement test administered at a large Midwestern research university, the decision of which is currently made on the basis of human raters' interpretations of examinee performance on a single essay item.

Fifteen constructed-response items that tap into various grammar features were developed drawing upon findings in a general developmental order of second language acquisition. The test was then administered to approximately 500 university-level ESL students in order to assess their productive grammatical ability within academic settings. Preliminary analysis results reveal that the Cronbach's alpha reliability coefficient of the test is .75, and the item discrimination indices (represented by point-biserial correlation coefficients) are all in acceptable ranges (.30-.58). The test-takers' performance on the academic grammar test has a relatively weak, but statistically significant, correlation with their performance on the ESL writing placement test (r=.37, p=.000). The item difficulty indices, however, do not precisely correspond to the predicted developmental stages of the target grammar features. The correlation between the examinee performance on the academic grammar test and that on the TOEFL iBT will also be reported in the study. A moderate, significant correlation is expected between the grammar test and TOEFL iBT scores, given the results of the previous research in a test of productive grammatical ability in general settings.

After examining the validity of the academic productive grammar test in terms of domain definition, item and rubric developments, and warrants and backing evidence for subsequent interpretive arguments, a discussion will continue to look into the validity of the test use in a hypothetical, computer-delivered medium-stakes testing situation. To this end, examinees' performances on the academic grammar test and on the ESL writing placement test will be analyzed at individual student and rater levels. The test use validity issues will particularly address possible consequences that the use of the academic grammar test in ESL decision making may effect on (a) individual students' academic plans in relation to their ESL placement as well as their L2 development, (b) rater training and rubric refinement for the ESL writing placement test, and (c) the revision of the ESL curriculum at program level.

Writing Assessment Literacy: A Survey of Teachers’ Knowledge, Beliefs, and Practices

Good writing assessment practices are essential to the teaching of second language writing. Teachers may receive formal instruction in giving feedback to students, but guidance in assessment involving scoring, grading, or making judgments about student work is often absent. In addition to classroom assessment, writing is often assessed in general tests of English language proficiency and used in decision making, such as placement of students into classes. In order for teachers to interpret and use scores from these larger scale measures, they need to understand fundamentals of writing assessment. Contributing to the assessment literacy of practitioners is the responsibility of language testers (Taylor, 2009), and writing assessment professionals have identified teacher development areas on this topic (Weigle, 2007). However, we have yet to determine what teachers currently know, believe, and practice, which should guide us in how to provide support for further learning on this topic. This study investigates what teachers currently know, believe, and practice regarding writing assessment. A survey was designed to elicit teachers’ backgrounds and perspectives on assessment using multiple choice, Likert-scale, and open-ended response items. The questions were designed to elicit information on L2 writing/ ESL teachers’ knowledge about writing assessment and the sources of this knowledge, teachers’ beliefs about writing assessment, and common practices used in classroom writing assessment. The survey was sent to ESL/EFL teachers in two- and four-year colleges as well as K-12 settings. Quantitative analysis of the multiple choice and Likert scale items included individual item descriptive statistics and analysis of variance (ANOVA). Correlation analysis was also used to investigate connections across the three areas of interest (i.e. assessment knowledge, beliefs, and practices). In addition, the open-ended responses were read for thematic patterns. Our results showed infrequent formal training in rubric design, scale calibration, and score interpretation. Additionally, the results from the survey showed a moderate relationship between the teachers’ writing assessment beliefs and their practices. Furthermore, teaching context was found to affect writing assessment literacy. The poster will report on this study, providing implications for better preparation of new teachers in the field of second language writing.

References
When Rating L2 Speaking?

When raters rate oral L2 productions, which features of the performances affect their perception? Several studies have found that components like fluency, linguistic accuracy, pronunciation and vocabulary are important facets of L2 speaking proficiency. In these studies holistic ratings of speaking performances are related to objective measurements of specific features of these performances, or to analytic rating. As Lumley (2005) showed that raters’ knowledge and experiences play a central role in the process of language proficiency rating, it is not surprising that various studies - most recently Hsieh (2011) - demonstrated that trained raters pay attention to different features of speech performances than naive raters do. What we know about components of speaking proficiency is largely based upon studies in which the ratings were performed by expert raters who were trained to pay attention to certain aspects. However, for valid testing of speaking proficiency the use of the rubrics for rating should be relevant to the purpose of the test and concur with what for stakeholders turns out to be important in the speaking performance. Therefore it is important to investigate which components of speaking proficiency are involved in the judgment of not only those who are expert raters but also those who are stakeholder of the test. To gain more understanding in what relative weight raters give to different components of L2 speaking performances in holistic rating tasks, I have applied an experimental approach to address the following research questions: RQ 1: How does improvement of (a) the morphosyntactic accuracy, (b) the pronunciation, and/or (c) the fluency of original speaking performances affect holistic ratings of speaking proficiency of Dutch L2 speakers? RQ 2: Do professional and non-professional raters weigh these factors similarly? In this study holistic ratings are collected for 17 original speech performances with adequate content, and also for four other versions of these speech performances, that is with improved pronunciation, and morphosyntactical errors corrected and/or fluency improved. Magnitude estimation is used as a means to obtain holistic ratings. Hundred raters are employed to rate the 85 (17*5) performances: 50 professional expert Dutch L2 raters and 50 raters who can be considered stakeholders. All raters will be asked to recall the rationale behind their choices, and additional information about their exposure to other languages and other background variables will be collected. More in-depth post-hoc interviews will be held with a subsample of raters. As data collection has started in October 2011, at the LTREC the first results will be available and reported.

References


**Latent Classes in Reading Assessment: A Mixture-Distribution Rasch Analysis**

Thomas Eckes  
*TestDaF Institute*

Analysis of examinee responses to test items typically rests on the assumption that item parameters are homogeneous across examinees; that is, the items are assumed to behave in the same way for all examinees. In a conventional Rasch analysis, a single difficulty parameter is estimated for each item, and all item difficulty estimates are located on a single dimension along with a single ability parameter for each examinee.

However, when examinees systematically differ in the ways they understand or solve the items, this assumption may no longer hold. Rather, there may be subgroups of examinees, requiring more than one set of item difficulty estimates.

The usual approach to checking the assumption of homogeneous item parameters is to study differential item functioning (DIF) based on manifest grouping variables. In language testing, these variables most frequently refer to language background, gender, or ethnicity. Yet, this approach may be inadequate because manifest groups often are highly heterogeneous regarding a number of factors possibly causing DIF.

In the present research, a fundamentally different approach was adopted. Examinee responses in a large-scale reading assessment were analyzed building on a mixture-distribution (mixed) Rasch model (Rost, 1990). Specifically, it was asked whether the group of examinees was composed of non-overlapping subgroups, or latent classes, where a Rasch model holds within each latent class, and item parameter values differ across classes. By definition, the latent classes are not specified in advance, but are recovered as part of the model estimation.

The reading assessment comprised 30 dichotomous items included in the reading section of the Test of German as a Foreign Language (TestDaF). This test was administered to 2,214 examinees (1,429 females, 785 males). All examinees were foreign students applying for entry to an institution for higher education in Germany. Data were analyzed by means of the dichotomous mixed Rasch model using the computer program WINMIRA (von Davier, 2001). Solutions for 1 to 7 latent classes were considered. In order to determine the appropriate number of latent classes, the relative fit of each solution was assessed based on information theory statistics.

Results showed that (a) the 1-class solution had the lowest fit and had to be rejected, (b) the 2-, 3-, and 4-class solutions were similarly rejected due to relatively poor fit, (c) the best-fitting solution consisted of five classes, with class proportions of .27, .26, .21, .19, and .07, respectively. Examinees belonging to Class 1 had substantially higher reading ability than examinees from any of the other classes. Moreover, within Class 1, difficulty estimates for Items 1 through 30 increased in line with the basic design of the reading section. Each of the four other classes deviated from the expected difficulty progression in specific ways. Between-class differences in difficulty estimates were particularly pronounced for the first 10 items, which were matching items referring to the easiest of three reading texts. Taken together, the results provided evidence supporting a mixed Rasch approach to the analysis of language testing data.

**Revising the Written Test of a University-level Summative Assessment in Hong Kong**

Felicia Fang  
*Hong Kong Polytechnic University*

Alan Urmston  
*Hong Kong Polytechnic University*

The written test of the Graduating Students Language Proficiency Assessment (GSLPA) has been in operation as a university English exit assessment in Hong Kong for five years since 2006/7. During this period of time, the assessment content, format and rating scales have changed little. Feedback from stakeholders including raters, teachers and consultants has indicated that there is a necessity to review and possibly revise the test in order to maintain its credibility and the meaningfulness of the interpretation of its scores by stakeholders. There is also a broader context to consider with the introduction into Hong Kong of a new university curriculum, extending the current three years of undergraduate study to four, from 2012 onwards. In addition to the informal feedback mentioned above, previous research has indicated specific areas for consideration of modifications to the GSLPA. There are two tasks in the written test and whether this would be enough to provide reliable and valid estimates of candidates’ writing ability expected in the dynamic workplace is worthy of investigation. Post-test surveys of students and raters carried out by the test administrators (2009, 2011) also suggest the need for revision in terms of the types of topics, the amount of writing and the difficulty level of the tasks. This study reports the findings from a revision project of the written test including, what changes to make to the test in terms of the nature, topics and number of tasks, the task difficulty level and the reasoning behind any changes that might be made, based on survey results from the stakeholders (test takers, host departments, employers, graduates/professionals, raters). The revision project carried out a review of the existing task specifications and how the tasks currently work based on evidence from test data of approximately 14,000 previous candidates. The project also explored the possibility of a technical upgrade of the written test, i.e. changing the delivery mode from paper-based to computer-based. It is hoped that through the revision process, the test will become a more effective reflection of candidates’ writing ability and improve the potential washback to the discipline-specific writing courses planned under the new 4-year curriculum.
Glyn Jones  
Pearson

The EAQUALS CEFR-Based Assessment Special Interest Project

This poster presents the objectives, rationale and methodology of the EAQUALS Special Interest Project (SIP) on CEFR-based assessment, as well as a sample of its end product. EAQUALS (Evaluation & Accreditation of Quality in Language Services) is an international association of institutions and organizations involved in language education. Its aim is to promote and guarantee high quality in language teaching and learning. EAQUALS has a strong commitment to the application of the Common European Framework of Reference for Languages (CEFR). EAQUALS has developed its own version of the European Language Portfolio as well as CEFR based certificates of achievement which member schools can issue to students. In November 2010 EAQUALS launched a special interest project on assessment to support these activities. The project brings together EAQUALS members and associate members from the fields of language training, language policy, language assessment and publishing. Its overall objective is to help language teachers to produce effective assessments that are linked to the CEFR.

Its detailed objective is to:
- formulate a principled methodology for developing language assessment tasks that are linked to CEFR levels
- produce a set of “scenarios” consisting of thematically linked assessment tasks that exemplify the methodology and serve as models for further task development.
- develop a users’ guide to enable teachers to apply the methodology in order to develop their own CEFR-based assessments.

These deliverables should enable teachers to produce assessments that are:
- Practical – relatively easy and cheap to produce with readily available resources
- Capable of yielding diagnostic information - and appropriate feedback – where possible
- Compatible with good classroom practice - so likely to have a positive washback effect.

The poster will include a sample scenario for assessment of reading (or a substantial extract from such a scenario) and will show how this has been derived from CEFR descriptors in accordance with the rationale and methodology of the project.

John de Jong  
Pearson / VU University Amsterdam

Ying Zheng  
Pearson

Native Speaker Responses in Language Test Development

There’s recent debate in the literature on whether language testing needs the native speakers (Davis, 2011; Hulstijn, 2011). Researchers have indicated that in theory native speakers’ responses are invaluable. In practice, however, how native speaker trials should be carried out and how their responses can be used remains largely unexplored. This study examines the use of native speakers as a quality control measure in developing an English test targeted at test takers who are non-native speakers of English and who intend to study in institutions where English is the principle language of instruction.

This study reports on the use of native speakers’ responses in two separate stages of the test development. Native speakers were defined as a group who reported that they 1) speak English at home; 2) hold an English-speaking country citizenship; 3) were born in an English-speaking country. The first stage involves field tests when items were trialled with university students for whom English is not their first language (n >10,000). English native speakers (n > 1,000) were recruited as a control group for item quality analyses. Native speakers’ responses were analyzed with regard to their response times and their response correctness. Based on classical test theory, items were removed when:
- items had lower proportion of correct answers from native speakers than from non-native speakers;
- item difficulties were out of range for non-native speakers;
- item-total correlations were too low (<.02);
- item-total correlation for one or more of the distractors was greater than that of the item-total correlation of the keyed option.

Based on the results obtained from the above analysis, further item scaling and dimensionality analyses were conducted.

The second stage involves an ongoing item seeding process. Native speakers’ responses are collected on seeded items within live operational test forms. The patterns and characteristics of native speakers’ responses were analyzed in depth on their specific psychometric features to ensure item quality. Further, test takers’ performance on those who only indicated their home language as English and those who are categorized as native speakers were compared. The results indicated that the concept of ‘native speaker’ needs careful defining, and native speakers of English may not be the ultimate criterion group for an English test for non-native speakers, because native speakers’ ability vary, but on average, native speakers should be reasonably proficient in English and perform well on tests like this both in absolute terms and relative to non-native speakers. The result of this study demonstrated the feasibility and usefulness of utilizing a native speaker sample as a measurement “filter” for quality control in terms of analyzing the level of difficulty and general appropriateness of test questions provided an adequate definition of native speaker is used.
The Impact of the Number of Multiple-Choice, Test-Item Options on a Test’s Construct: Do Five Options Versus Four or Three Really Change Things?

Multiple-choice tests are popular because they are objective, reliable, easy to score, and efficient (Haladyna, 2004). In a high-stakes testing situation, affective distractors should be the norm, but often they are not. This is because good multiple-choice items are hard to write. Thus, researchers have spent more than 80 years trying to determine the best number of options in a multiple-choice item (Rodriguez, 2005). The basic question is, should it be three, four, or five? Educational measurement researchers conclude that three-option items are best (Crehan, Haladyna, & Brewer, 1993; Delgado & Prieto, 1998; Haladyna & Downing, 1993; Landrum, Cashin, & Theis, 1993; Owen & Froman, 1987; Rodriguez, 2005; Trevisan, Sax, & Michael, 1991; Trevisan, Sax, & Michael, 1994) because, as summed by Crehan et al., they are easier to write, have more effective distractors, and take less time to administer. Yet still, high stakes L2 tests often have four or more options. Why?

This study investigates the claim that the number of items matter in the context of an operational, high-stakes, L2 English listening test (the CSAT, which is a college-entrance exam) in Korea that has five-option items. Only one study investigated the optimal number of test-item options in a high-stakes, L2-test situation (Shizuka et al., 2006), and it only compared three versus four options. No study on this topic (that we know of) reported data concerning how tests with differing numbers of distractors correlated (for example, in Shizuka et al., test takers were divided into distinct groups, making correlation analysis impossible), which would provide important information on whether tests that vary in the number of test-item options actually test the same underlying construct. Thus, we created a research design that would investigate how the test construct is affected by the number of test-item options.

We adapted three, practice College Scholastic Aptitude Tests (CSAT) of English listening, each with five-option items, to create four- and three-option versions by asking 73 Koreans to eliminate the least possible options in two rounds, which resulted in nine tests. We placed 264 high school, English-language learners in Korea into three groups. Each took three of the nine tests, one with five-option items, one with four-, and one with three-, with administrations counter-balanced across groups to control for test-content, order, and practice effects. Average test scores differed significantly. Five options were more difficult than three. However, item facility, discrimination, and reliability were similar regardless of the number of options. Low test correlations indicated the number of options affects the test construct: items with more options appear to tap into test-taking strategies and cognitive skills beyond listening ability. CSAT administrators may favor five-option items because five options result in a slightly more normal curve, which may be important in such a high-stakes test. However, learners finished three-option items fastest, which could allow for additional items and concomitantly higher test reliability.

We discuss how test developers may need to consider multiple statistical, cognitive, affective, and contextual factors in determining the optimal number of options.

“Faking” Lower Proficiency or “Just Coping” During Legal Procedures?

This poster presents a framework for analyzing L2 data to document coping strategies that too frequently may be misread by law enforcement as either understanding more than the person actually can or as “faking” a lower than truthful language proficiency. The data come from two proficiency tests commonly used in forensic cases in three actual legal cases. In these cases one side claimed the speakers really understood but were “faking” and the other side, using language tests and some analysis of language data, argued that it was highly likely that they were not “faking”. This poster reports on a follow-up analysis of the data drawing on the work of Canale and Swain (1989) on Communicative Competence (Strategic Competence) and the second language use of strategies used by non-native speakers (NNS) when using Coping Strategies (Compensatory and Cover strategies) Cohen (1996, 2011). However, legal contexts impose a limitation in the application of learner strategy research. The use of retrospection of data, common in strategy research, is not appropriate for testing and analysis of legal evidence. Myths and labels about NNSs language use can lead to errors in judgment which can then affect how a NNS is treated legally. Myths represent a lack of understanding about NNS language use and language learning and may reflect attitudes and negative stereotypes about NNSs. “They lie about their comprehension of English.” “If they’ve been in the US for 10 years and haven’t learned English, they’re lazy.” “Broken English” might suggest something is wrong with the person. Law enforcement officers (LEOs) are trained not to make firm assumptions without adequate evidence. However, if a NNS uses some survival English words or phrases or responds with nods, some LEOs might assume the NNS can also understand and use more complex language. This is not adequate evidence about overall language proficiency. Unfortunately, errors in judgment are sometimes made. Fortunately, forensic linguistics experts can examine the evidence, add assessments, and provide professional opinions. Thus, they can assist the judge/jury in determining whether an error in judgment may have been made about a NNS’s language proficiency. This study represents one modest step in trying to overcome such misunderstandings. Canale and Swain identified, as a component of Communicative Competence, “strategic competence” which includes the need to compensate for communication breakdowns. This might be due to “limiting conditions” or “insufficient competence.” Cohen, in his work on learner strategies, proposed two categories of Coping Strategies: Compensatory and Cover Strategies. Learners use Compensatory Strategies “to allow them to compensate for a lack of some specific language knowledge.” Cover strategies involve “creating an appearance of language ability so as not to look unprepared, foolish, or even stupid.” (Cohen 2011, pp. 13-14).
Assessing Language Proficiency for Air Traffic Controllers: The Brazilian Air Force Exam

In recent years, renewed attention has been devoted to linguistic factors that can contribute to safety in aviation. In 2004 the International Civil Aviation Organization (ICAO), taking note of several accidents and incidents in which problematic communication between pilot and air traffic controller was a determinant or contributory factor, formulated a framework for language proficiency and required a minimum proficiency in English for air traffic controllers and pilots. To comply with that requirement, the Brazilian Air Force, responsible for civilian and military air traffic controllers in the country, developed an exam called EPLIS, which focuses on the speaking and listening skills of those professionals. EPLIS was developed by language teachers and aeronautical subject matter experts based on ICAO’s requirements for radiotelephonic communication which is comprised of two phases. The first one is a computer-based multiple-choice listening test. The second phase consists of an interview conducted by two professionals, a language teacher and an Aeronautical subject matter expert. In this presentation we will describe the EPLIS exam since its creation, in 2006, and comment its developments and adaptations throughout the years. We will also bring into question some criticism of the exam and of the very framework it set out to follow, as seen in a validation study conducted by language testing professionals and in focus groups with the exam developers.

Investigating the Construct Validity of the Writing and Speaking Sections of a Placement Test

Performance assessment has increasingly gained attention in the field of language testing (e.g., Bachman, 2002; McNamara, 1996; Norris, Brown, Hudson & Bonk, 2002; Weigle, 1998). Using performance tasks in order to measure speaking and writing ability is considered a more direct form of testing in that the test takers are required to ‘use’ the actual ability or skill on which they are being measured. However, the discussion and the attempt to identify and explain construct-relevant or irrelevant factors concerning performance assessment have been ongoing in order to establish a better understanding of the construct(s) being measured.

This study attempts to explore the nature writing and speaking ability by investigating the underlying structures of writing and speaking performance tests administered as part of a placement test in an ESL program. We looked at the extent to which scores from the writing and speaking sections of a placement test can be interpreted as indicators of test taker’s writing and speaking ability as defined by the components of the rubrics that were used.

Writing and speaking scores of 144 students were collected from an administration of an ESL program placement test. The writing section included two tasks (writing a postcard and writing an essay), and the speaking section included three tasks (role-playing, decision making, and group discussion). Test takers’ performances on the tasks were rated by two raters on four components, and the average across components and tasks were given as the final writing or speaking score. The four components for the writing test rubric were task completion, content control, organizational control, and language control, and the four for the speaking test rubric were task completion, meaningfulness, grammatical competence, and conversational competence.

The construct validity of the traits specified by the scoring rubrics and their contribution to defining writing and speaking ability in the placement test was investigated by conducting reliability analysis, exploratory factor analysis (EFA), and confirmatory factor analysis (CFA). The high internal consistency reliability of both tests (α = .95 for both tests) and the strong correlations between the rating scales (.70 to .89 for writing and .85 to .94 for speaking) showed that the components for each test collectively contribute to the measurement of the constructs of interest (i.e., writing and speaking).

However, the results from the CFA indicated that the model which best accounted for the two tests involved both trait and method factors. In other words, in addition to the traits specified by the components of the rubric, task characteristic was found to have a non-negligible effect in understanding the constructs of writing and speaking ability. The two writing tasks differed in genre (i.e., general vs. academic), and the speaking tasks differed to a greater extent in terms of both type of interaction (i.e., pair vs. group) and test condition (i.e., administration). The results of this study raise the importance of understanding test method and task characteristics and their potential effect on test performance (Bachman & Palmer, 1996).
Creating a Test of British English Collocation Knowledge for Young Learners

Assessment of vocabulary in reading comprehension has tended to emphasize counting the number of words in a text an individual knows, or vocabulary breadth (Pearson, Heibert & Kamil, 2007). However, in order to process these lexical items effectively the learner must also have knowledge of grammatical functions, register, stylistic characteristics, appropriate contextual usage, idioms and collocations, or vocabulary depth (Milton, 2009). One area of vocabulary depth of increasing interest is the role multi-word phrases, collocations, play in language acquisition and language processing. For the purposes of the current study, a ‘collocation’ will be defined as a composite unit of words that expresses meaning as a whole and whose components co-occur more often than would be expected by chance (e.g. take vs take place). While it has been long established through corpus linguistic evidence that such multi-word items are ubiquitous in naturally-occurring discourse (Erman & Warren, 2000; Sinclair, 1991), and that such items can negatively affect the reading comprehension of adult L2 learners of English (Bishop, 2004; Martinez & Murphy, 2011), at present the nature of collocational knowledge in young learners and its possible correlates with literacy remain relatively under-explored, largely because there are no available measures appropriate for use with children. The current study details the creation and validation of a measure of British verb + noun collocations aimed at children between age 7 and 11. Test items ranged from high to low frequency and from literal phrases to opaque idioms. This paper reports the findings from the validation with 131 English as an Additional Language (EAL) learners and monolingual English speaking children in years 3, 4, and 5. Analysis shows high internal consistency and reliable test scores. The test discriminates well between learners of different ages, as well as between monolingual speakers of English and EAL learners. Scores correlate highly with other measures of vocabulary knowledge, indicating a relationship between vocabulary size and collocation knowledge. Item facility and individual item analysis will also be presented with an analysis of item difficulty in light of item frequency and opacity. This paper also provides a discussion of study limitations, and offers an outlook on future applications and possible further research.

Development of a Spanish-English Language Proficiency Scale (SELPs)

This project aimed to develop a criterion-referenced Spanish-English Language Proficiency Scale (SELPs) that yields valid and reliable score interpretations based on contemporary measurement standards (AERA, APA, & NCME, 1999) and language acquisition theory. The SELPs is designed to measure the level of oral language proficiency (LP) in 4-8 years old Spanish-English bilinguals. This poster presents results of the scale development for assessment of English LP. The study consisted of three development phases. Phase I was dedicated to the development of the framework and operational definitions of the LP domains; development of a rating scale and language tasks that elicit behavioral evidence of the LP construct; and evaluation of parallel task forms using the same scale for possible construct-sampling errors. Phase II evaluated the construct validity of the score inferences based on the SELPS by examining the relationship of the scale scores and domain scores with external measures. Phase III evaluated the effects of language ability (language impairment or typical language development) on the SELP scores. Phase I: The SELPS assessed the following language domains: syntactic complexity, grammatical accuracy, verbal fluency, and lexical diversity based on a story retelling task that is an alternative to standardized language assessments for bilingual children (Gutiérrez-Clellen, 2002). The scale is intended to differentiate between five levels of oral LP with each higher level subsuming lower levels based upon modified stages of L2 acquisition developed by Tabors (2008). Each domain is assessed on a 4-5 point scale, and then a scale score is assigned. The SELPS rating criteria for each language domain and number of domains went through multiple iterations. To elicit behavioral evidence of LP, two story retelling tasks were developed and administered to 40 bilingual children ages 4-5-8.4 years. Children passed a hearing and cognitive screening. Reliability of the two tasks was evaluated: (a) inter-rater reliability assessed by the weighted kappa value yielded a value of 0.81; (b) a Wilcoxon test indicated n.s. differences between the domain and scale scores on the two tasks; (c) the Spearman’s rho correlations between the domain scores ranged from $\rho = 0.59$ to $\rho = 0.82$; (d) the percent agreement between the two scale scores was 87.5% with the linear weighted kappa coefficient of 0.62. Phase II: In evaluating the construct validity of the SELPS scores, language sample measures and teacher’s report of LP were used as external measures. Seventy-six bilingual children participated. Pearson product-moment correlations between the language sample measures and the domain and scale scores ranged from $r = 0.34$ to $r = 0.66$. The Spearman rho correlation between the scale scores and the teachers’ ratings of English LP was $\rho = 0.39$. Current findings provide initial evidence of construct validity. Phase III: Fifty-two bilingual children with language impairment and 66 with typical language development who were matched on chronological age within one year (ages 5-8), participated in the study. Results indicated a significant effect of language impairment ($\eta^2 = .12$) on LP scores suggesting that assessment of LP is confounded by children’s language ability.
Veronika Timpe  
Dortmund University  

**Decoding Socio-pragmatic Utterances in Intercultural Contexts — A Think-Aloud Study**

Byram’s (1997) Model of Intercultural Communicative Competence includes three language competences — linguistic, sociolinguistic, and discourse competence — which he regards as essential in intercultural communicative encounters. Sociolinguistics has proven to be a challenging field in language assessment given that pragmatic expectations and assessments are highly culture- and context-specific (Blum-Kulka, 1991; Spencer-Oatey, 1993; Liu, 2007). Hence, it is challenging to avoid construct-irrelevant variance. This study focused on Byram’s idea of sociolinguistic competence by investigating the cognitive processes of university-level German learners of English when solving receptive intercultural socio-pragmatics assessment tasks. The study was intended to answer the following questions: 1. What strategies do students apply, i.e., what steps/actions do they take, in order to answer the items? 2. Do the strategies differ between the two groups of test-takers? 3. Do the strategies mirror the assumptions underlying the test construct? To answer these questions two groups of university-level EFL students with different intercultural experiential backgrounds were identified. One group (N=7) had never been to the United States, while students from the second group (N=7) had participated in study abroad at an institution of higher education in the United States for at least one semester. Both groups were asked to answer 24 multiple choice discourse completion tasks, taken from an intercultural socio-pragmatics comprehension test that focuses on U.S.-American English as well as the academic context in the United States. Think-aloud methodology, in particular concurrent verbal reports, was used to access respondents’ cognitive processes and to collect verbal protocols from the students while they were working on the tasks. The verbal reports were coded and analyzed for strategic processing and knowledge. Findings revealed that students highly draw upon personal experiences and cultural knowledge, involving consistent, yet varying degrees of cross-cultural comparisons and critical evaluations. As one of the first studies to investigate intercultural socio-pragmatic comprehension processes by means of think-aloud methodology, my study provides insights into the ways in which students make sense of socio-pragmatic utterances. These insights will be relevant for ESL/EFL teachers and test developers alike. Instructors may benefit from a clearer understanding of strategies and knowledge involved here insofar as it could assist them to direct their instructions more efficiently at developing and promoting sociolinguistic competence in their learners. Test developers would profit from the information in terms of construct representation.

Tatyana Vdovina  
Center for Applied Linguistics  

Jennifer Christenson  
Center for Applied Linguistics  

**Alternate ACCESS for ELLs Assessment**

This proposal is for a poster presentation of an assessment that is currently being developed as a collaborative effort by the multistate WIDA Consortium. The poster will present the purpose of the assessment as well as the process of test development. It will outline the unique features of the assessment, discuss challenges of test development, and solicit feedback from the audience. The Alternate ACCESS for ELLs™ is a large-scale assessment of English language proficiency (ELP) for students in grades 1-12 who are classified as English language learners (ELLs) and who have significant cognitive disabilities that prevent their meaningful participation in the ACCESS for ELLs®. It is based on the WIDA Alternate English Language proficiency levels. The Alternate ACCESS for ELLs is designed for students who are classified as ELLs and have significant cognitive disabilities. Students with significant cognitive disabilities are identified as having one or more of the existing categories of disabilities under IDEA (e.g., intellectual disabilities, autism, traumatic brain injury, multiple disabilities, etc.) and their cognitive impairments may prevent them from attaining grade-level achievement standards, even with the best instruction. (adapted from U.S. Department of Education: Alternate Achievement Standards of Students with the Most Significant Cognitive Disabilities Non Regulatory Guidance, August 2005). The purpose of the assessment is to meet the No Child Left Behind Act (NCLB, 2001) and Individuals with Disabilities Education Improvement Act (IDEA, 2004) accountability requirements; facilitate the involvement of ELLs in participating states’ accountability systems; provide a method for monitoring the English language proficiency growth of ELLs with significant cognitive disabilities; and provide guidance to Individualized Education Program (IEP) teams in developing appropriate language proficiency IEP goals and objectives. Students with significant cognitive disabilities may require additional instructional support in their educational environment. Alternate ACCESS for ELLs recognizes and addresses this need by allowing everyday instructional support and practices to be utilized during testing. Additionally, assistive devices are allowable in order for the student to perform their best. Lastly, the assessment has been designed and scripted to include repetition of tasks, cueing, modeling, and auxiliary questions in order to provide student with multiple opportunities to demonstrate their English language proficiency.
Developing a Validity Argument for Score Meaning of a Computer-Based ESL Collocational Ability Test

Among the language capacities that L2 learners of English develop as they increase their proficiency is the ability to use appropriate collocations in the relevant registers of language use (Lewis, 2001; O’Dell & McCarthy, 2008). Assessment of this capacity may therefore provide an efficient means of distinguishing among examinees with various levels of language proficiency within a particular register, although research has not yet attempted to operationalize such a measure. The purpose of this research in progress is to explore the utility of a measure of developmental, register-specific collocational knowledge as a means of making relevant distinctions among examinees in a minimal amount of time. Ultimately, such a test could be used in conjunction with other information for norm-referenced decisions such as placement, with a convincing argument for such uses. The first step, however is to develop the interpretive argument and validity argument for score meaning of such a measure. This study followed a mixed methods embedded and sequential explanatory design consisting of quantitative and qualitative data collection (Creswell & Plano Clark, 2007). Quantitative data were collected from performance on the test of collocational ability, a vocabulary size test and a reading test by more than 200 Chinese-speaking learners of English at three levels of English proficiency. Qualitative data were collected for a sample of the population during and after the quantitative data collection and analyzed in order to identify evidence to explain initial quantitative results from the collocational ability test including test taking processes. Qualitative data included screen-capturing during test administration, post-test interviews and a test reflection survey. The test development, piloting, data collection and analysis provide backing for the assumptions underlying inferences in the interpretive argument to help support in intended interpretations. The poster will present the development of a validity argument for the meaning of a score on a computer-based ESL collocational ability test which could be used to contribute to the decision to allow a test taker to participate in English-medium instruction or place the test taker in an appropriate English language course. The first stage in the validity argument begins with the interpretive argument which defines the grounds, inferences, warrants, and claims which provide the foundation for score interpretation (Bachman & Palmer, 2004; Kane, 1992; Mislevy, Steinberg, & Almond, 2003). The claim that the interpretive argument needs to support is that the test score reflects the ability of the test taker to use and understand lexical collocations as they are written in college and university settings. Backing for the inferences in the interpretive argument includes a theoretical analysis and empirical data. Theoretical evidence provides the backing for the domain definition, evaluation and part of the explanation inferences. These empirical results provide evidentiary support for assumptions backing part of generalization, part of the explanation, and the extrapolation inference in the interpretive argument. The backing for the assumptions that underlie the interpretive argument provide the basis for the validity argument.

Deconstructing the Construct of Teachers’ Diagnostic Competence: Investigating Teachers’ English Language Learning Assessments Practices

Many teachers have not received any formal training in assessment, and are not equipped with a sound understanding of assessment practices and literacies (Black & Wiliam, 1998; Childs & Lawson 2003; Stiggins, Conklin, & Bridgeford, 1986; Taylor, 2008). In classrooms with students learning English as an additional language, teachers need to be able to integrate information gleaned from formative and summative assessments into their practices in order to provide students with the necessary feedback and instruction to advance their language learning development. Edelenbos and Kubanek-German (2004) refer to this ability of teachers to describe and interpret students’ emerging language competencies in order to use the data to inform their instruction as “diagnostic competence”. The goal of this study was to observe and document the assessment practices of teachers across K-12 learning contexts and identify the specific competencies that they need in order to: (1) design effective assessment tasks, and (2) use tasks to elicit diagnostic information about students’ language skill development. This research serves as the foundation for developing an instrument to evaluate teachers’ diagnostic competence to guide teachers’ professional development and assessment training. We investigated 40 teachers’ use of a state-mandated formative language assessment framework. Teachers from different regional districts and different grade clusters implemented the framework into their practice for a two-month period. Our sources of evidence included: (1) pre- and post- surveys which gathered data about teachers’ assessment practices, and reflections on the use of the assessment framework respectively; (2) video documentation of teachers’ use of tasks to elicit formative and diagnostic assessment information about students’ language learning skills; and (3) interviews with teachers to identify how and why teachers use assessments in the classroom and the extent to which it informs their instruction of language. Analyses of the teachers’ surveys revealed that the two primary purposes for teachers’ use of assessments are to inform instruction, and to provide evidence for provision of feedback and reporting. Teachers primarily rely on three types of assessments: informal classroom observations; student work products; and students’ self- and peer-assessments. Analyses of video and interview data following the use of the assessment framework identified five integral components comprising teachers’ diagnostic competence: (1) ability to interpret assessment information and identify language learning skills; (2) ability to respond to and use diagnostic information; (3) ability to observe language behaviours; (4) perception of self-efficacy; and (5) confidence about the reliability of their assessments. These results serve as the foundation for the development of a teachers’ diagnostic competence measure to guide teacher training.
The Cross-National Comparability of the Internal Structure of the TOEFL Junior Test

TOEFL Junior is a relatively new test developed for the purpose of measuring young adolescent EFL students’ English proficiency for English-medium instructional environments at the middle-school level. The paper-based version of TOEFL Junior consists of three sections: Listening Comprehension (LC), Reading Comprehension (RC), and Language Form and Meaning (LFM) sections. Since its launch in October 2010, this test has been administered in various countries in Asia, Europe, the Middle East, and South America. In an effort to provide construct validity evidence for this new test, this study aimed to investigate the internal structure of the test and its comparability across test takers from multiple countries. This study also attempted to replicate previous dimensionality studies to examine whether a multi-componential language model with a higher-order factor of general English ability holds true for the TOEFL Junior test and its test-taker population, i.e., young adolescent students. Considering the intended use of the test scores as an objective indicator of students’ proficiency worldwide, this study is significant in that it provides supportive evidence for this intended test use and also advances our understanding of an EFL/ESL language ability model for younger learners from various educational contexts. The data for this study included 3,990 students from Korea (N = 1,366), France (N = 776), and Vietnam (N = 1,848). The data also included ESL students (N = 1,026) enrolled in U.S. middle schools. Since the test was designed to measure English language proficiency for English-medium instructional settings, the ESL students in the U.S. schools were of particular interest. The study employed two analytic methods: (1) a content analysis of TOEFL Junior items to identify the language skills measured in each test item, and (2) a series of confirmatory factor analyses (CFAs) to identify an underlying test structure. A content analysis protocol was developed to consider two aspects of each test item: the sub-skills tested and the contexts in which language use is involved. The sub-skill categories included comprehending explicit meaning and key details, getting the main idea, and comprehending implicit meanings such as inferences and predictions. The contextual categories included social/interpersonal, instructional, and academic contexts. Two raters coded each item using this coding protocol. A series of CFAs were performed with item parcels constructed based on the content analysis results. Hypothesized models included a single-factor model, a two-correlated-factor model (LC and RC/LFM), a three-correlated-factor model (LC, RC, LFM), and a higher-order factor model. Additionally, a bi-factor model specifying the relationship between the sub-skills and contexts was also hypothesized. Upon identifying the best-fitting model, a multi-group confirmatory factor analysis was conducted to examine the invariance of the test structure across test taker samples from different countries. Preliminary results indicated that a correlated-three-factor model was the best representation of the test’s internal structure across the sample test takers from different countries. Interestingly, there was some variation in factor loadings and factor correlations across the countries. In this poster session, we will present the complete results of the study as well as their implications for the score inferences and an EFL/ESL ability model.

Mikyung Kim Wolf
Educational Testing Service
Jonathan Steinberg
Educational Testing Service
Investigating Effective Strategies for Teaching Assessment Literacy to Instructors of Less Commonly Taught Languages

This work in progress (WIP) will present preliminary findings from a study of assessment literacy of instructors of less commonly taught languages (LCTLs). Assessment literacy is defined as the knowledge and skills instructors need to accurately and effectively plan for and administer assessments and interpret and apply the results (Taylor, 2009). Effective classroom assessment is a necessary component of today’s foreign language programs because it provides insight for instructors on how to improve teaching and learning (Inbar-Louie, 2008). Assessment results also provide key data for accountability and arguing for maintenance of language programs and can be used for program advocacy, improvement, and expansion (Reese, 2010). This WIP describes the changes in instructor knowledge resulting from a blended learning assessment course that combines online and face-to-face formats. The nine-week course was designed for in-service K-16 LCTL instructors. These instructors often have limited teaching resources in their target languages and limited training in pedagogy (Wan, 2009), thus creating a great need for assessment literacy in this population. The study described in this WIP provides insight into how to cultivate assessment literacy by tracking the progress and depth of understanding of course participants over time. Data were collected from pre- and post-course surveys, assessment tasks developed during the course, discussion board posts, and course instructors’ observations of participant learning. These data show what LCTL instructors know about assessment, how their knowledge emerges through assessment practices, and what opportunities exist for providing teacher professional development in order to strengthen their use of assessment. The data suggest that course participants made gains in their awareness of the utility of aligning assessment with instruction, including the importance of developing tasks that elicit the expected proficiency level of the students and embedding these tasks in relevant instructional contexts. Participants showed a need for further development in the actual writing of proficiency level-appropriate tasks and in developing appropriate criteria for accompanying rubrics. During the session, presenters will discuss effective strategies for fostering assessment literacy in LCTL instructors and for improving future assessment literacy training. Because this study explores the evolution of LCTL instructors’ assessment knowledge, beliefs, and practices, it provides an important and often-overlooked foundation for the development of assessment literacy training. Feedback will be elicited from attendees about the data as well as future directions for research about classroom-based assessment practices. Finally, the session will discuss implications for additional assessment and training resources that may be needed in programs that teach LCTLs.

Empirically Verifying Douglas’ ESP Ability Hypotheses: An IRT-SEM Approach

Although English for specific purpose tests are being developed and used, the construct of ESP ability is understood vaguely. Douglas (2000) proposed that ESP ability consists of three major components, i.e., language knowledge, background knowledge, and strategic competence, and that the focus of ESP ability is on the interactions between these components, especially, the interaction between background knowledge and the other two. Empirical verification of this hypothesis has not been confirmed. The purpose of this study was to validate Douglas’ hypothesis in the context of ESP testing in China. To realize this purpose, this study examined the direct and indirect effects of specific-purpose background knowledge (medical knowledge in this case), grammatical knowledge (lexical and syntactical knowledge), and strategic competence (metacognitive competence) on test takers’ performance on the reading section of the Medical English Test System Level Two (METS-2), an ESP test recently developed in China. Subsequently, all possible two-way interactions between each pair of these components were tested. This large-scale study involved about 1600 nurse students from eight medical colleges across China. Participants were asked to respond within two hours to a set of tests and one questionnaire presented in the following order: an medical knowledge test (in Chinese) assessing specific-purpose background knowledge, an English grammar test assessing lexical and syntactical knowledge, an METS-2 reading test assessing ESP reading test performance, and a strategy use questionnaire evaluating metacognitive competence. The response data were first analyzed using 2PL-IRT models and composite scores formed based on these analyses. Subsequently, CFAs (confirmatory factor analyses), single group SEMs (structural equation models), and SEMs with latent variable interactions were performed with these composite scores. Findings from the study revealed that all three components had significant effects on ESP reading test performance. Specifically, the effect of language knowledge was direct and strongest, while the effect of background knowledge and that of strategic competence were less important and indirect and by way of language knowledge. Further, two sources of interaction were detected in predicting ESP reading test performance: the moderation effect of background knowledge and that of strategic competence on language knowledge. These results confirmed Douglas’ hypothesis that ESP ability includes language knowledge, background knowledge, and strategic competence. Nevertheless, the central role of background knowledge that was implied in Douglas’ model was not supported, given the absence of direct effects from background knowledge and its interaction with strategic competence on ESP reading test performance. The results seemed to suggest that, language knowledge plays the most important part in affecting ESP reading test performance, either in terms of its relative importance in predicting ESP reading test performance, or in terms of its functioning as a bridge between background knowledge and strategic competence and ESP reading test performance. In addition, the use of composite scores based on IRT models and IRT for SEM analyses appeared to be successful and implied the possibility of combining IRT and SEM in investigating the relations between different language ability components.
**Teacher Assessment Practices for ELLs in the Content Areas**

Beth Clark-Gareca  
New York University

Classroom tests are an important tool used to evaluate student achievement. A key component in the determination of report card grades, classroom test scores greatly influence placement and tracking decisions for ELLs in K-12 environments. Despite the significant effect that these scores have on ELLs’ academic trajectories, little is known about how teachers administer content tests (e.g., science and math) to ELLs in mainstream classrooms (Cizek, 2007). As part of a larger dissertation study, the work presented here investigates the following question: What are the assessment practices of elementary mainstream teachers for English Language Learners in the areas of math and science?

Ten 4th grade teachers and 50 ELLs in six Pennsylvania schools were the focus of an initial classroom observation and two subsequent observations during routine math and science tests. Teacher practices and student testing behaviors were documented, including teacher-student interactions, testing conditions, and accommodations implementation. Semi-structured interviews were then conducted with each teacher to learn about general assessment practices, knowledge of tests and accommodations, grading practices, and the principles that guide decision-making for ELLs. An online survey of K-6 teachers is in process and will report broadly on classroom teachers’ assessment practices. Teachers from 8-11 school districts across Pennsylvania (including the two districts which contain the focal 4th grade classrooms) were invited to participate based on the following inclusion criteria: 1) ELLs make up at least 5% of the student body of the district (N=100), 2) School districts are located in town, suburban, and small city settings (NCES, 2008a). The survey will capture large-scale teacher report of math and science assessment practices, including accommodations implementation, grading practices, and teacher opinion relating to ELL assessment. Observation and interview protocols and transcripts were transcribed, analyzed, and coded inductively. Data were categorized according to observed and reported implementation of accommodations/modifications and grading practices. Typical high-stakes accommodations were documented (e.g., bilingual tests, use of dictionaries, translator/interpreter, additional time) as well as ad hoc practices implemented by teachers (e.g., use of preferential seating, reading items aloud, paraphrasing, retesting). When completed, survey results will be analyzed quantitatively through descriptive statistical methods. Analyses will include calculations of teacher report of assessment practices by education levels, foreign language proficiency, opinions on accommodation effectiveness, level of training, and student language proficiency levels. This study widens the scope of classroom and content assessment research for ELLs by beginning to document what actually happens during regular content test administrations. Considering that many state departments of education (e.g., Pennsylvania, Florida, Texas, North Carolina) call upon accommodations implementation in the classroom as a model for high-stakes accommodations practices (Pennsylvania Department of Education, 2009b, among others), the results could have implications for state policy-makers if practices are found to be inconsistent. This work will also have many applications for professional development at the elementary level relating to assessment principles. Finally, this study will help to further the discussion of classroom assessment as an accurate and valid measure of ELL academic achievement.

**The Utility of Subscores for Automated Assessment of English Speaking Proficiency**

Keelan Evanini  
Educational Testing Service

Lei Chen  
Educational Testing Service

This study considers the use of subscores in the context of an automated assessment of spoken English proficiency. From a pedagogical and diagnostic standpoint, subscores are highly desired by end users of spoken language assessments, since they can provide more specific information about how the test takers can improve specific aspects of their spoken language proficiency (such as pronunciation, grammar, fluency, etc.). However, recent studies have shown that subscores often provide little to no added value over the holistic scores (see Sinharay (2010) for a review). Furthermore, the Standards for Educational and Psychological Testing (AERA, APA & NCME 1999) state that subscores should only be reported if they are reliable, comparable, valid (Standard 5.12), and distinct (Standard 1.12). This study reports on on-going work investigating to what extent automatically generated subscores for a proficiency assessment of spoken English can meet these criteria. High-stakes international assessments of spoken English typically provide a single speaking proficiency score (IELTS, TOEFL iBT), although some report more detailed subscores for different components of speaking proficiency (for example, the PTE Academic reports subscores on oral fluency and pronunciation). Despite the fact that such detailed feedback about a test taker’s spoken English proficiency would be beneficial to the end user, little research has been published about the reliability and distinctness of such scores (see Xi and Mollaun (2006) and Xi, Zechner & Bejar (2006) for a notable exception). This is especially true for the rapidly growing area of automated spoken language assessment. This study examines a set of responses containing restricted speech (reading a paragraph out loud) that were obtained in the context of an assessment of spoken English proficiency. For each response, human raters provided subscores on a three-point scale for the speaker’s pronunciation and intonation / stress. Then, each response was processed by a system for automatically assessing English speaking proficiency, and metrics relating to the scoring rubrics for pronunciation and intonation / stress were extracted from the speech signal using speech recognition and audio analysis software. In total, 10 pronunciation metrics (for example, the speech recognizer’s average confidence value and the acoustic model score, etc.) and 30 intonation / stress metrics (for example, average distance between stressed syllables, pitch range, etc.) were extracted. Then, the responses were partitioned into training (N=3046) and evaluation (N=1852) sets, and a decision tree classifier was trained to model the human subscores. On the evaluation set, the classifier achieved an exact agreement rate of 57.8% for the pronunciation subscores and 56.8% for the intonation / stress subscores. These results are not too far below the exact agreement rate shown by two human raters on the evaluation set. While this initial result is promising, work is still on-going to determine how reliable these automatically predicted subscores can be (for example, by incorporating additional metrics), and whether they can provide additional information beyond a holistic score.
Marta Fairclough  
*University of Houston*  

**A Working Model for Assessing Spanish Heritage Language Learners’ Language Proficiency through a Placement Exam**

A heritage language learner is someone who has been raised in a home where a minority language is spoken, who speaks or at least understands the HL, and who is to some degree bilingual in the majority and heritage languages (adapted from Valdés, 2000). Despite the many differences that characterize each minority language situation, the teaching of heritage languages shares some common phenomena: heritage learners come from diverse social, economic, and linguistic backgrounds, and they possess a certain affective connection with their heritage language. These and other factors explain the varied proficiency of these speakers in the heritage language, which ranges from barely receptive to completely productive (Carreira 2003, Fairclough 2005, Valdés 2001). Having a clear idea of the knowledge in the heritage language that a student brings to the classroom is essential for a successful language-learning experience; for that reason, research in heritage language education has been focusing increasingly on assessment issues, especially language placement exams. Professionals debate whether assessment procedures should measure what students know and can do with the heritage language (Polinsky and Kagan 2007, Tucker 2005), or what deficiencies they have (Zentella 1997, Ziegler 1981). Ideally, a balance of both will produce a better placement exam (Fairclough et al. 2010). This paper presents a working model of language placement based on current second-language theory and research and on what is known of heritage language learning. It explains the rationale for content selection and the design of effective tasks in a language placement exam for incoming university students; and it briefly voices some key logistical considerations.

Anna Fletcher  
*Charles Darwin University*  

**Using Student-Directed Assessment as a Learning Process**

This study originated in reaction to the 2008 introduction of a Nation-wide Assessment Program in Literacy and Numeracy (NAPLAN) by the Federal Government in Australia. Assessment is well documented to play a key role in establishing and raising standards of learning but with the current political emphasis on using assessment for accountability and to monitor macro educational standards, its formative practice is less noted. Consequently this mixed-methods study into using Student-Directed Assessment (SDA) as a learning process was conducted at a Primary (Elementary) School in The Northern Territory of Australia. The study draws on Formative Assessment principles, Constructivist learning approaches and Self-Regulated Learning. A simultaneous design has been employed in which both qualitative and quantitative data is given equal priority. The study explores how students engage in learning when they are able to identify their own learning goals, determine their assessment criteria and how they will demonstrate their mastery of learning outcomes. Early findings indicate that when students in the participating year levels (Years Two, Four and Six) are involved in their own assessment and when assessment is central to learning, that deep and meaningful learning experiences occur. Through this collaboration with teachers, students display behavioural, emotional and cognitive engagement with their assessment tasks. In the case of the Year Six students, the Student Directed Assessment group who actively planned and took charge of their learning in the assessment process achieved significantly higher scores in their writing sample compared to their peers in the Teacher Directed Assessment group.
Karen Fung  
The University of Alberta  

Samira Elatia  
The University of Alberta  

Identifying Test-Takers Characteristics in Assessing a Language in Minority Contexts  

Using data from the French version of the 2007 Pan-Canadian Assessment Program (PCAP) and the 2004 School Achievement Indicators Program (SAIP), this research identifies the patterns and subgroups within the students’ population who take the French version of the test. Two research questions are raised: first, is there a difference in the performance of the students based on their ethnic and schooling background? We believe that differences in test performance would be found between the French subpopulations depending on their background. Second, is there a ‘distance’ factor that needs to be considered in the performance of the students? We assume that as we move away from the French hub, more group differences in the French used would be noted; since in these minority situations, the linguistic models are scarce and the performance tend to pidginize. Results of this study would shed light on assessing another language in a majority and minority contexts as well as raise awareness among different stakeholders to the complexities in assessing with one model of language.

Raili Hildén  
University of Helsinki  
Outi Hakola  
University of Helsinki  

Assessment Practices on a Special Oral Language Course in the Finnish Upper Secondary School  

The main focus of this presentation is on the validity of assessment practices of oral foreign language proficiency at the upper secondary level of general education in Finland. A course specialized on oral proficiency was introduced into the national core curriculum in August 2010. The course grade is partially based on teacher’s views on students’ overall performance during the course and partially on test tasks drawn on a test bank authorized by the National Board of Education. The theoretical framework encompasses conceptions of multi-level language curriculum, teacher and student perceptions and oral language assessment. The research approach is primarily determined by assessment theory viewed from the perspective of Assessment justification proposed by Bachman (2010) and informed by the works of Messick (1989) and Kane (2006). Teacher perceptions are investigated in the framework of assessment justification which includes Assessment Use Argument (AUA) enriched by collecting evidence to support or challenge a set of claims concerning consequences, decisions, interpretations and reports.

Claim 1. Consequences: The current practices of assessing the oral foreign language proficiency of Finnish upper secondary students enable valid decisions and are beneficial to the major stakeholders, e.g. the students, teaching staff, tertiary level institutions and employers.  

Claim 2. Decisions on equitability and values sensitivity: Test takers whose oral proficiency is at the equivalent levels are assigned to the same group, and relevant legal requirements and community values are taken into account when considering the decisions based on these practices.

Claim 3. Interpretations: The current assessment practices provide stakeholders with sufficient information about oral proficiency and convey this information in comprehensible terms. The format and content of assessment practices and all aspects of administration treat all students equally. There is a reasonable correspondence between the current assessment practices and the target language use in real-life. The interpretation of the assessment outcomes provides relevant information to decision makers.

Claim 4. Assessment reports: Students performance on different assessment practices yields essentially the same records. The aim of the project is to carry out a multi-method survey including two inquiries addressed to teachers (n~600) and students (n~3.000) of foreign languages at the upper secondary level who have taken the oral course. The inquiries include both closed, Likert type questions and open questions. After the quantitative and qualitative analyses of the inquiries (group comparisons, correlations and frequencies) classroom observations will be carried out to ensure the reported classroom and assessment practices. A first phase inquiry (n=173) for teachers show that in the assessment process they tend to stress fluency and pronunciation and pay less attention to the vocabulary and accuracy. The foreseen presentation will look at the general beliefs about the current practices held by teachers and students, but in the future the project also aims to complete the evidence by describing, analyzing and explaining these perceptions with in-depth interviews and classroom observations, as well as statistical studies on student performances on the oral test attached to the course.
**Alexis Lopez**  
*Educational Testing Service*

**Interaction Between Summative and Formative in English Language Classrooms in Colombia: Teachers’ Perceptions**

Current classroom assessment practices require English language teachers to collect, analyze and evaluate evidence about the teaching and learning process, and about learning outcomes. From this perspective we see that language assessment fulfills at least two basic functions. On the one hand, assessment allows teachers and students to reflect on classroom practices and to seek ways to improve the teaching and learning process. On the other hand, assessment is a tool for accountability of student learning and provides information about what and how much students are learning. That is, classroom assessment practices should be used as assessment of learning (formative assessment) and for learning (summative assessment). Thus, it is important to examine how English language teachers integrate both formative and summative assessment practices to enhance and facilitate language learning.

The purpose of this ongoing study is to investigate the interactions between formative and summative assessment practices in English language classrooms in Colombia. This study will be conducted between October 2011 and January 2012. It will employ a sequential mixed method approach in which the researcher first collects quantitative data and then collects qualitative in order to help support, explain, or elaborate on the quantitative results. The data collection will be done in two distinct phases. In the first phase, the researcher will gather data on teachers’ perceptions regarding the use of formative and summative assessment in English language teaching using an online questionnaire. This questionnaire will be sent out to English language teachers around the country. Teachers will be contacted via the Colombian Association of English Teachers and the Colombian English Teachers Network. The teachers who will complete this questionnaire are English language teachers from different parts of the country and teach at different institutions: primary schools, secondary schools, universities and language institutes. In the second phase, the researchers will collect qualitative data through the use of semi-structured individual interviews to try to elaborate on, refine or explain the findings in the first phase. Approximately 30 teachers from around the country will be interviewed. The teachers will be selected using a maximum variation sample (teachers with a lot of experience, teachers with little experience, teachers in public institutions, teachers in private institutions, teachers in rural areas, teachers in urban areas, teachers in primary and secondary schools, teachers in universities and teachers in language institutes). These interviews will address topics such as routine assessment practices in those teachers’ classrooms, the relationship of assessment with teaching and learning, and the interaction of formative and summative assessment. The researcher expects this study to uncover meanings of teachers’ experiences with the use of formative and summative assessments, how these assessments shape classroom instruction and language learning, and how formative and summative assessments interact in the teaching and learning process.

**Jia Ma**  
*Queen’s University*

**English Language Learning through TOEFL iBT Preparation: A Case Study of Chinese Test Takers**

To be granted the academic admissions into overseas postsecondary institutions, Chinese students must submit proof of English proficiency through test scores, such as the TOEFL iBT. Although Chinese test-takers constituted more than 20 percent of the global TOEFL candidates (Qian, 2009), few studies have exclusively examined Chinese TOEFL test takers except for Chau’s (2008) study on the Chinese students’ TOEFL writing preparation. As the preparation courses that Chinese students take may improve their ability to pass the English language test instead of their English proficiency (Fox & Curtis, 2009), it is necessary to understand how Chinese test takers’ are prepared for English language proficiency tests. Therefore, this study employed a case study approach (Yin, 2009) to investigate Chinese TOEFL iBT test takers’ unique preparation experience at one popular commercial test preparation center in northern China.

To delineate the case’s uniqueness, data were collected from multiple sources (document analysis, classroom observations, and interviews), and from multiple stakeholders (teachers, students and administrative staff) at this center. The collected documents included: in-house textbooks designed and used by four TOEFL iBT teachers responsible for the listening, reading, writing and speaking sections of the TOEFL iBT courses, the descriptions (accompanied by videos) of two non-test-oriented English language skill courses, and the advertisements and videos of two series of events/seminars developed to foster motivation, interest and confidence in learning English as a foreign language. The seven classroom observations covered the TOEFL iBT instructions by the four TOEFL iBT teachers, and five interviews were conducted—one interview with administrative staff during the document collections and four interviews (two teachers and two students) following the classroom observations.

The preliminary results of the document analysis suggest that the test preparation center tried to create an encouraging learning environment by presenting the experiences and methods of successful English language learners. The initial analysis of classroom observations and interviews reveals that: 1) the teachers made efforts to draw students’ attention onto the construct that TOEFL iBT measures—English language proficiency, 2) in addition to help students familiarize the TOEFL iBT, the teachers repeatedly presented the methods they have summarized to exemplify how to build up the four English language skills (such as four-step intensive listening method); 3) the students learned some practical ways to direct their own TOEFL iBT preparation and their English language learning, such as writing effective English essays by following their own thoughts rather than by mechanically applying some all-purpose essay templates. Taken together, these preliminary findings suggest that there are promising TOEFL iBT test preparation practices in China, in contrast from a negative washback effect of TOEFL preparation courses reported in Chau’s (2008) study.

Different from the previous studies (Alderson & Hamp-Lyons, 1996; Chau, 2008; Wall & Horák, 2006; 2008) that take TOEFL preparation classroom as unit of analysis, this study investigated the contextual factors and interactions among the multiple stakeholders within a popular test preparation center in China, and the detailed results will help better understand the washback effects of TOEFL iBT in China.
WORKS-IN-PROGRESS ABSTRACTS

David MacGregor
Center For Applied Linguistics

Stephanie Gibson
Center for Applied Linguistics

Dorry Kenyon
Center for Applied Linguistics

Margaret E. Malone
Center for Applied Linguistics

Francesca Di Silvio
Center for Applied Linguistics

Megan Montee
Georgia State University

Developing a Computer-Mediated Spanish Academic Language Assessment for Young Bilinguals

The Illinois State Board of Education, in a partnership with a large multi-state consortium, was awarded an Enhanced Assessment Grant from the U.S. Department of Education to develop Spanish Academic Language Standards for students in grades K-12 and an accompanying standards-based, computer-mediated assessment for students in grades K-2. This presentation describes the current status of the development of Prueba Óptima del Desarrollo del Español Realizado (PODER) assessment and the challenges related to developing a computer-based assessment of academic Spanish for this age group. A growing body of research supports a connection between L1 proficiency and success in second language development (Escamilla & Hopewell, 2010; Tabors, 2008; Schleppegrell & Colombi, 2002). Developing literacy in students’ L1 can foster the development of literacy in English (Bialystok, Luk, & Kwan, 2005; Cárdenas-Hagan, Carlson, & Pollard-Durodola, 2007; Roberts, 1994). In addition, students who are receiving content-area instruction in Spanish need to develop the language skills necessary to access that instruction. Currently, however, there are no standards-referenced Spanish academic language proficiency assessments for L1 Spanish speakers and other students who receive content instruction in Spanish (including students in two-way, full, or partial immersion programs, and students in Puerto Rico for whom Spanish is a second language). As a result, academic programs have no instrument to assess a student’s baseline Spanish academic language proficiency level and to measure a student’s progress in developing Spanish academic language proficiency over time. In order to support the academic success of such students, it is imperative to understand where they fall on the continuum of language development in Spanish and, in bilingual contexts, in English as well. PODER operationalizes Spanish language development (SLD) standards and will provide a tool for assessing a student’s academic language development in Spanish. Additionally, the SLD standards are analogous to English language development standards, so scores on PODER, when considered in conjunction with a student’s performance on an academic English language proficiency test, will provide a fuller profile of the student’s language development as a bilingual learner. This work in progress session seeks input on how we are addressing the unique developmental needs of young language learners (Kindergarten to grade 2) and feedback on item prototypes. Additionally, we seek advice on options for integrating technology into the assessment and in what capacity students in this age group are able to navigate technology.

Assessment Literacy Across Applied Linguistics

Knowledge about language assessment is important for applied linguistics researchers and practitioners across a variety of contexts. However, most research on language assessment literacy, or the knowledge users need to have about language assessment to make informed decisions (Inbar-Lourie, 2008; Taylor, 2009), has focused primarily on instructional contexts and the needs of language teachers to conduct fair and valid assessments of their students. As language testers continue to develop research about and resources for language assessment literacy, additional contexts and groups of assessment users across disciplines must be considered. This work in progress session will discuss the results of a multi-year project that examines the assessment literacy needs of practitioners in two fields, Second Language Acquisition (SLA) researchers and professionals who work in Language Teacher Education (LTE). The purpose of the project is to develop assessment literacy resources for each group. The project investigates the current assessment knowledge base of both groups to determine how resources should be developed to meet the needs of these audiences. For the SLA context, the project focuses on how collaboration can be facilitated between the fields of SLA and language testing in order to support SLA researchers’ use of language tests in research contexts and enhance research methodologies. For the LTE context, the project focuses on the essential knowledge of assessment needed by language teacher educators to provide appropriate pre- and in-service professional development for language instructors. This session discusses results from four focus groups conducted with SLA and LTE researchers and practitioners. Focus groups were recorded, transcribed, and analyzed for major themes related to participants’ current practices, challenges, and resources needed in language assessment. The results of the focus groups will inform the development of future phases of the project, including large-scale surveys of both groups’ assessment needs. During the work in progress session, researchers will discuss focus group results and planned project developments and facilitate a discussion about collaboration between language testers and practitioners in diverse disciplines. The session will raise questions about the responsibility of the language testing community to make testing research and materials accessible to colleagues in related fields. Feedback from this session will be used to improve language assessment materials designed for the SLA and LTE groups. Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. Language Testing, 25, 328-402. Taylor, L. (2009). Developing assessment literacy. Annual Review of Applied Linguistics, 29, 21-36.
Michelle Raquel
Hong Kong University of Science and Technology
University of Tasmania

Towards a Framework for Assessing English through Drama: A Dynamic Assessment Approach

Task-based performance assessment is the common approach to the assessment of English language proficiency when students are engaged in dramatic activities for English learning. This approach is popular because alignment between assessment tasks and learning objectives is transparent, and, because it works well with formative assessment classroom practices which promote development of creativity and critical thinking skills. In Hong Kong, commonly used classroom assessment methods in assessing English development through drama involve pre- and post-testing, and/or portfolio assessment. In theatrical productions, continuous and summative assessments are popular; students’ English is assessed either on their final performance, and/or continuously assessed through the completion of mini-tasks such as script-writing to assess writing skills, or reading aloud to assess pronunciation. Although these assessment methods are effective in taking a “snapshot” of students’ current proficiency level, they fail to provide information on students’ individual development throughout the process. In addition, these grades only reflect what students demonstrate in performance and do not take into account other English skills that develop as they engage in other activities in the learning environment. An additional issue is the pretense of using drama to promote creativity and language awareness caused by pragmatic pressure to prepare students for public examinations makes teachers teach and assess only linguistic competence. This study aims to examine the feasibility of adopting Dynamic Assessment (DA) as a framework for the teaching and assessment of English through participation in full-scale theatrical productions. DA is a development-oriented assessment approach that aims to promote learner development by directing teaching and assessment to students’ potential ability. As students attempt to complete tasks, teachers provide mediated assistance by engaging students in collaborative dialogue. This mediation process triggers students’ zone of proximal development (ZPD), which is the optimal activity that facilitates internalisation of the target language. Using case study methodology, the study investigates the English proficiency development of four Hong Kong Chinese tertiary students, of varying levels of English proficiency and theatre experience, engaged in the process of preparing to perform a scripted text for an audience. Video recordings of rehearsals, director and student journals, and in-depth interviews have been collected throughout the production process and analysed for potential DA interactions. To identify how and what specific areas of English were developed, these interactions have been further analysed using Poehner’s (2008) model of profiling learner development in DA. After parallels between the production process and DA were drawn, the data reveals that DA and consequently, English proficiency development, can naturally occur in the process of preparing for and performing in a theatrical production provided that the director promotes learner development of theatrical skills through mediated assistance. Elements of the production process that operate as DA interactions thus serve as a basis for the development of a DA framework for the teaching and assessment of English through drama.

Xiaomei Song
Queen’s University

Fairness of A Large-Scale High-Stakes Language Test: Investigations With the Test Itself, Test Users, and Test-Takers

Concerns about fairness among test stakeholders are paramount in high-stakes decision-making tests such as admission. Given the importance of fairness in admission testing, my research examined fairness from multiple perspectives with one of the high-stakes tests in China—the Graduate School Entrance English Exam (GSEE). Based on a synthesis of the 11 standards and codes of practices (e.g., AERA, APA, & NCMER, 1999) and scholarly work (e.g., Willingham & Cole, 1997) in testing and assessment, I build a conceptual framework and an operating model of fairness to guide fairness investigations. The framework conceptualizes fairness as comparability investigations across the four stages of the testing process (test design, development, administration and scoring, and use) from multiple perspectives within the context in which a test operates. This conceptualization enhances the scope of fairness investigations, emphasizes the importance of stakeholder involvement, highlights the role of value judgments informed by the testing context, and promotes more methodologically diverse conception of fairness. Predominant research on test fairness has been conducted with groups of test-takers examining psychometric properties of tests developed in North America (e.g., Angoff, 1989). There is insufficient research to examine fairness from multiple perspectives, especially stakeholder perspectives, in language testing and assessment. Studies of this nature are scarce in the Chinese testing context, which influences a huge number of test-takers and other test stakeholders. The major purposes of the GSEE are to measure English proficiency of test-takers and to provide information for educational institutions to select candidates for their Master's programs. The research used two phases to investigate fairness of the GSEE. Phase 1 examined one specific aspect in test development—potentially biased items using SIBTEST and content analysis. Specifically, Phase 1 examined whether items/bundles of the 2009 GSEE administration functioned differentially and advantaged/disadvantaged towards test-taker groups of gender (male vs. female) and academic background (Humanities/Social Sciences vs. Sciences). SIBTEST found two grouping variables had a mixed effect on the GSEE test scores, with some favouring the focal group and others favouring the reference group. A review of the flagged items/bundles by three test reviewers concluded that gender and academic background had minimal influence on the GSEE test scores, thereby playing limited role in affecting test-taker’s admission status for Master’s programs in China. Using focus group interviewing and one-on-one interviewing, Phase 2 investigated fairness perceptions as expressed by the GSEE test stakeholders—test-takers and two groups of test users (graduate program administrators and teachers). The preliminary results found that discrepancies existed regarding fairness perceptions within and across groups of test-takers and test users. Results also described how the overall testing-oriented educational system in China significantly influenced stakeholder fairness perceptions on the GSEE as well as English teaching and learning. By drawing on the accounts of those test stakeholders, it was possible to identify fairness concerns that may otherwise remain undetected by statistical procedures such as DIF. The two-phase research provides empirical information about fairness of the GSEE and elaborates on the complex nature of test fairness.
Testing Arabic Speaking Proficiency: An Examination of Test Language and Breakdown in Dialect, MSA, and Combined Tests

Testing the speaking proficiency of candidates who offer Arabic language speaking skills to U. S. federal government agencies is a challenge for a number of reasons. The diversity of colloquial Arabic variants (dialects) and the existence of Modern Standard Arabic (MSA) as a lingua franca among literate speakers of Arabic from all areas of the Arabic-speaking world lead to the question of the appropriate speaking proficiency test (SPT) to administer to determine a candidate’s suitability to perform required work. Native and heritage speakers have access to a colloquial variant, often the language of the home. Educated native speakers have exposure to MSA and are likely to have some level of MSA speaking proficiency. Heritage speakers often have had no formal exposure to MSA. Finally, learners of Arabic have often been exposed only to MSA. Operational needs for translators and interpreters often involve the ability to use and understand a dialect of Arabic. This might lead to the assumption that dialect SPTs are sufficient for hiring purposes. However, the characterization of Arabic as strictly diglossic is somewhat misleading. Well-educated native speakers of Arabic—those who score in the range of 4 and 5 when rated using the Interagency Language Roundtable (ILR) Skill Level Descriptions (SLDs) for Speaking—use their colloquial variant in certain circumstances and MSA in other circumstances; there also exists a range of circumstances where an ‘educated colloquial’ is employed. In July of 2011, one federal agency, working with its trained and experienced Arabic SPT testers, combined dialect and MSA into one SPT. The agency now awards one proficiency rating for candidates who take these combined tests. This approach permits both more efficient and more authentic testing of Arabic speaking proficiency. This exploratory study investigates the use of language variants by testers and examinees during archived SPTs conducted in dialect and in MSA, and on current SPTs using the combined approach. It also investigates where linguistic breakdown by examinees occurs, considering the task and the language of the prompt, and the language of the response. For the study, 15 SPTs (five of each type), each rated and certified at the level of 2+ to 3, will be transcribed and translated. The translations will be coded to indicate the use of dialect, educated colloquial, and MSA by testers and examinees. Test transcripts will also be coded to indicate the language use tasks elicited by test prompts. The percentage of each test that is conducted in each variant of Arabic will be reported. A task-by-variant analysis will be conducted to determine relationships that may exist; an analysis of the language of the prompt and the corresponding response will also be conducted. Segments of the transcripts that were identified in rater reports as examples of linguistic breakdown will be coded and a variant-by-breakdown analysis will be conducted. Implications for continued refinement of the combined Arabic SPT method will be reviewed.

Interacting with Video in L2 Listening Tests: An Eye-Tracking Study

Visuals such as video have been used in second language (L2) testing for a number of years; however, their role in assessing the listening comprehension of L2 learners is not fully understood (Buck, 2001; Ockey, 2007). Some researchers suggest that the use of video makes listening tests more authentic because in most real-life communicative situations people engage both auditory and visual senses for receiving information (e.g., Coniam, 2001). However, existing empirical studies on the effect of video on L2 learners’ listening comprehension have yielded inconclusive results. While some studies evinced that video can facilitate test-takers’ performance on L2 listening tests (e.g., Ginther, 2002; Wagner, 2010), others found detrimental or no effect of video on test-takers’ L2 listening comprehension (e.g., Coniam, 2011; Gruba, 1993; Ockey, 2007). In addition, since research in this area has been limited almost exclusively to comparative studies of test-takers’ performance, more qualitative investigations of test-takers’ viewing behavior as well as the extent of their interactions with and use of visuals during L2 listening tests are clearly necessary. To address this gap in research, the present study utilized an eye-tracking system to investigate L2 learners’ viewing behavior during an L2 listening test with audio-only and video academic lectures that used both content and context visuals (Bejar et al., 2000). The study participants were five native speakers of English who were graduate students and 70 ESL students enrolled in five sections of an intermediate-level English listening course at a large university in the Midwest of the US. The instruments used in this mixed-methods study included (a) an online pre-test questionnaire that elicited background information about participants, (b) a computer-based English listening test consisting of eight academic lectures (four in audio-only format and four video lectures) and 40 questions, (c) an online post-test questionnaire that inquired into students’ perceptions about the role of video in their listening comprehension during the test, and (d) stimulated recall aimed at eliciting students’ perceptions of their viewing behavior and the effect of video on their listening comprehension during the test. The ESL students were divided into two groups: (a) a test-only group (60 students) who completed the questionnaires and the listening test in a computer lab; and (b) an eye-tracking group (10 students) who, in addition to the questionnaires and the test, participated in stimulated recall and whose eye movements during the test were recorded in an eye-tracking lab. The five native speakers of English went through the same procedure as an eye-tracking group of ESL students. The qualitative analysis will involve the recorded eye-movement data from an eye-tracking group of ESL students and native speakers, the latter being used as the source of the baseline data (i.e., a record of normal eye movements when English listening comprehension is not an issue). The quantitative data (i.e., scores from the listening test) obtained from all study participants will be analyzed using descriptive and inferential statistics. The findings and implications of this work in progress will be discussed.
Test Design Considerations in Assessing Young Adolescent English Learners on Computerized Language Tests

In the 21st century, everyday language use is so commonly tied to technology that technology-based language learning has become a part of life starting from a very early age. In the current era of assessment design, computer-delivered language assessments take precedence as well. Particularly, young adolescent learners of English as a foreign and second language (EFL/ESL) are increasingly taking computer-delivered assessments. However, little is known about young adolescent learners’ (at the age range of 11-15) test-taking processes on computerized language assessments. Thus, there is a need to understand test-taker computer interface considerations, serving the broader goal of making valid inferences about the language abilities of young adolescent test-takers taking computerized language assessments. In light of this need, this study aimed to identify key considerations for assessing English language ability of young adolescent English learners by examining the relationship between the degree of computer familiarity and performance on computer-delivered test items designed for this age range. Study methods included conducting cognitive interviews with international and domestic middle school students on a sample of these items and in a larger sample investigating the relationship of computer familiarity with test performance.

A computer familiarity survey of fourteen items was administered to 2,000 international test takers who were taking a computerized English language proficiency pilot test. Test takers’ performance data on reading, speaking, listening and writing sections of the test and survey data were analyzed shortly after the pilot test. A cross tab of high achieving and high computer-familiar test takers, and low achieving and computer-unfamiliar test takers was computed. The significance of relationships between the computer familiar and computer unfamiliar examinees was subjected to MANOVA to account for relationship among the dependent measures. Degree of computer familiarity was the independent variable while performance on computer-based items or tasks was the dependent variable.

Further, cognitive interviews were conducted to explore how domestic and international English learners with varying degrees of computer familiarity perceived the computerized items. Cognitive interviews with ten international test takers and ten domestic English language learners were conducted using a sample of items. The items were selected based on the challenges that the design features of the computer-test taker interface might cause for English learners. Each interview lasting no more than forty-five minutes included two main steps: 1) having the student read the item aloud; 2) having the student think-aloud using concurrent and retrospective techniques. Interviews were audio-taped and coded for possible construct irrelevant variance on the following usability considerations: a) navigation; b) terms used in instructions; c) page layout; d) text; e) icons/graphics; f) space provided on constructed response writing items. Interviews with both the international and domestic English learners suggested ways in which terms used in test and item level instructions could be made less confusing. Also, strategies for helping young learners cope with difficulties in processing both textual and auditory input simultaneously were identified.

Tense and Aspect Avoidance in Testing Procedures: The Case of Elicited and Guided Measures

The use of tense and aspect markers (as verb inflections) in two different testing methods (the Oral Proficiency Interview (OPI) and elicited imitation paradigms) is compared to determine whether avoidance in one measure is more likely than avoidance in the other. The two measures elicit responses in different ways: the Oral Elicited Imitation (EI) method requires recall from Working Memory (WM) of stimulus sentences which include the target structures; while the OPI involves open-ended guiding questions of a pre-selected topic. EI selects the topic and manipulates the target structures (tense and aspect), such that their position within the sentence and the roots with which they are associated are systematically varied. Achieving these systematic variations typically does not result in contrived and artificial constructions in Spanish (the language of testing) as a consequence of the language’s rather free word order, and its reliance on inflectional morphemes to represent the grammatical categories of tense and aspect, i.e., the word order of declarative sentences in Spanish can be: SVO, VSO, or OVS in the case of object pronoun inclusion. Conversely, OPI procedures provide a theme or topic for the interview; and in the course of the response, the interviewer poses questions regarding the theme to elicit reference to time frames other than the present. This method is implemented to elicit spontaneously the interviewee’s comfort level with, and consistent and accurate use of, tense and aspect markers. Establishing the frequency and accuracy of use and form of tense and aspect markers within EI and OPI can determine which type of test is a more reliable measure of these powerful devices in language — more spontaneously elicited responses or highly structured ones. Twenty university L2 Spanish (third-year undergraduate) students will be given both an oral EI test and a Simulated OPI (SOPI). All subjects will be tested individually by the same evaluator, with the same EI stimuli and OPI topic; and their responses will be recorded. The oral EI consists of two paragraphs, each of which contains twelve stimulus sentences (half preterite, half imperfect verbs). EI stimulus sentences are audio recordings of the same native Spanish speaker. Subjects are instructed to repeat each sentence after hearing it, without altering it in form or content. After hearing each stimulus sentence, but before the response, subjects are instructed to provide a word association for each stimulus sentence. The latter procedure ensures that subjects are focusing on meaning and provides a short interval to guard against verbatim recall. The same evaluator (trained in OPI testing) will administer the SOPI to each student within three weeks of the EI test. Responses will be coded by two independent raters to determine tense and aspect accuracy, and meaningfulness. ANOVAs will determine levels of significance between proportions of form and use accuracy within and across subjects. Expectations are that subjects will avoid use of tense and aspect markers more often on the SOPI than on the oral EI test; but at the expense of greater proportions of form-based errors.
Kimberly Woo  
New York University

**Defining Oral Social Language in the Assessment of English Proficiency in Kindergarten-Age ELLs**

In recent years, there has been growing emphasis in the United States on the assessment of English language proficiency in school-age ELLs, which frequently takes the form of high-stakes large-scale standardized tests (Abedi, 2007). Often included among the tasks on such tests are those purporting to measure “social” language, especially in the oral language domain. However, despite the presence of these tasks on ELP tests, social language is a construct typically viewed as secondary to academic language and is weakly defined, usually discussed only in counterpoint to academic language. Lack of clarity surrounding the definition of “social language” questions the validity of these tasks, which is of great importance when they are used with ELLs in the earliest grades who are developing social skills concurrently with language development in both their first and second language (Coltrane, 2003). This work-in-progress presents the first part of a three-phase dissertation study exploring the validity of tasks used to measure oral social language on large-scale standardized ELP tests used with kindergarten ELLs. Specifically, this phase examines how the construct “oral social language” is defined and measured by large-scale ELP tests used with early childhood ELLs, with the definitions and measures utilized by teachers of young ELLs identified as a point of comparison. Content review of test materials (e.g. technical manuals, training guides, test samplers) from large-scale standardized ELP tests used under NCLB Title III (e.g. WIDA, NYSESLAT), and analysis using Bachman & Palmer’s (2010) test task characteristic framework identify the definition(s) of oral social language used by these tests and how these definitions are operationalized. A web-based survey is conducted to determine teachers’ definition(s) of the construct and identify the tasks and measures used to assess social language in the classroom. Teacher data are collected in New York City’s Chinatown, an immigrant enclave community in which development of children’s English oral social language is expected to be emphasized in early-grade ESL instruction. The survey targets classroom teachers and ESL service providers working with K-1 students in all six public elementary schools in the neighborhood (N=30) and explores teachers’ perceptions of social language and its role in kindergarten/ESL curricula. Teachers are also asked to identify the kinds of assessments they use with their students, to characterize common school-based oral interactions as more or less “social,” and to describe “social language” according a set of adjectives and linguistic features. This session presents preliminary descriptive findings from this first phase of work and seeks to contribute to the literature by addressing the need for clarification in the definition of “social language” in the assessment of young language learners (YLLs). In addition, this work provides an intersection between assessment of YLLs, high-stakes standardized tests, and second language classroom assessment, three major ongoing lines of research identified in the broader field of language assessment for school-age language learners (McKay, 2005).

Amy Yamashiro  
Alexandria City Public Schools

**Investigating Student Growth Percentiles as a Measure of ELL Instructional Effectiveness for Teacher Performance Evaluations**

Recent educational policy initiatives focused on school reform at the federal and state levels seek to require approximately 40% to 50% of teacher performance evaluations to include measures of teacher effectiveness to be calculated through a value-added growth model based on a standardized or other empirically-validated assessment of student achievement. Much of the research has been focused at the state level for the two most commonly administered assessments, English Language Arts and Mathematics. However, policy makers seek to expand teacher evaluations to include all of the other academic subjects (e.g., Science, Social Studies, Physical Education, Art, etc.) and specialists, such as teachers of Special Education students and of English language learners (ELLs). For example, in Virginia about 40% of an ELL teacher’s annual performance evaluation could be based on student growth on an ELL assessment, such as the one developed by the WIDA (World-Class Instructional Design and Assessment) Consortium. With over 22 states in the consortium, WIDA (World-Class Instructional Design and Assessment) has already begun to conduct research on value-added growth models using ELL assessment data at the national, state, and school district level. This work-in-progress session will investigate the application of student growth percentiles for ELL teacher evaluations within a small, highly diverse school district having a large proportion of ELLs. For this study, student growth percentiles will be examined using WIDA test results from the Spring 2009, 2010, and 2011 administrations. A brief overview of the teacher evaluation model to be used in Virginia along with the method for reporting student growth percentiles to teachers and administrators and the various ELL delivery models used within the district will precede the discussion of the preliminary results. Since WIDA provides average district growth patterns on its ELL assessment, growth patterns for the district used for this study will be compared to this external benchmark for all available cohorts. District growth patterns use the WIDA Composite Scale Scores and the WIDA Proficiency Levels by the following five grade clusters: Kindergarten, Grades 1 and 2, Grades 3 to 5, Grades 6 to 8, and Grades 9 to 12.
Investigating the Construct Validity of a Diagnostic Test of Range and Strength of Grammatical Knowledge

Assessment of grammatical knowledge based on influential models of communicative language ability (e.g., Bachman & Palmer, 1996) is a rather neglected area of research in the field with many open questions (Purpura, 2004). This work-in-progress argues that the construct of grammatical knowledge can be enriched by incorporating more recent proposals about the nature of grammatical development in language learners along dimensions of range i.e., type and frequency of the forms used and their complexity, and accuracy (Rimmer, 2006). The current study proposes an additional viable dimension namely strength of grammatical knowledge. This can be studied by measuring learners’ ability to perform on a series of items with increasing difficulty all of which test the same grammatical form. Since it is generally accepted that as language learners progress, their grammar becomes more sophisticated (Ortega, 2003), such sophistication can be studied by examining amount and type of change along all three dimensions of type and frequency, complexity, and strength. Such a conceptualization of the construct of grammatical knowledge can be particularly valuable for diagnostic tests, by providing a more detailed feedback to teachers, test-takers and the other stakeholders about the development of grammatical knowledge. This work-in-progress report details the rationale for the suggested framework of grammatical knowledge for diagnostic purposes and how it is operationalized with reference to a model for test specifications proposed by Davidson and Lynch (2002). Furthermore, a plan for test piloting, analysis, and interpretation is proposed addressing the effect driven theory of test development (Fulcher & Davidson, 2007). Given not only grammar assessment but also diagnostic assessment is a neglected concept in applied linguistics and very little research is available on developing diagnostic tests and their validation (Alderson, 2007; Huhta 2008; Kunnan & Jang, 2009) the study attempts to contribute to the field by investigating the validity of a more comprehensive construct of grammatical knowledge for the purpose of developing a diagnostic test to be used by teachers to plan remedial tasks, or for students for self-assessment.
The TOEIC® tests — Research-based assessments for English in the workplace

Based on 60+ years of world-renowned research by ETS, the TOEIC® tests were designed to provide:

• An accurate and relevant measure of English-language skills used in the workplace
• Test content based on real-life workplace situations relevant to global organizations
• Scoring held to the highest quality-control standards resulting in fair, valid, and reliable scores
• Security of the highest level to ensure unmatched test integrity
• A standardized testing administration for equal opportunity for all test takers

See the research behind the assessment used by more than 10,000 organizations in 120 countries.

Visit www.ets.org/toeic/research
<table>
<thead>
<tr>
<th>Name</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ackermann, Kirsten</td>
<td>42</td>
</tr>
<tr>
<td>Alderson, Charles</td>
<td>32</td>
</tr>
<tr>
<td>Alt, Meredith</td>
<td>43</td>
</tr>
<tr>
<td>Amoroso, Luke</td>
<td>46</td>
</tr>
<tr>
<td>Artzi, Lauren</td>
<td>60</td>
</tr>
<tr>
<td>August, Diane</td>
<td>60</td>
</tr>
<tr>
<td>Bach, Aileen</td>
<td>73</td>
</tr>
<tr>
<td>Bachman, Lyle F.</td>
<td>55, 58, 60</td>
</tr>
<tr>
<td>Bailey, Alison L.</td>
<td>53</td>
</tr>
<tr>
<td>Baker, Beverly</td>
<td>61</td>
</tr>
<tr>
<td>Barni, Monica</td>
<td>56</td>
</tr>
<tr>
<td>Bridgeman, Brent</td>
<td>59, 81</td>
</tr>
<tr>
<td>Brooks, Rachel</td>
<td>80</td>
</tr>
<tr>
<td>Brophy, Stephanie</td>
<td>29</td>
</tr>
<tr>
<td>Butler, Yuko Goto</td>
<td>53</td>
</tr>
<tr>
<td>Byrnes, Heidi</td>
<td>56</td>
</tr>
<tr>
<td>Cai, Yuyang</td>
<td>73</td>
</tr>
<tr>
<td>Carless, David</td>
<td>52</td>
</tr>
<tr>
<td>Chambers, Lucy</td>
<td>36</td>
</tr>
<tr>
<td>Chapelle, Carol A.</td>
<td>59</td>
</tr>
<tr>
<td>Chen, Lei</td>
<td>74</td>
</tr>
<tr>
<td>Cheng, Liying</td>
<td>77</td>
</tr>
<tr>
<td>Chi, Youngshin</td>
<td>43, 62</td>
</tr>
<tr>
<td>Cho, Bokyoung</td>
<td>61</td>
</tr>
<tr>
<td>Choi, Ikyu</td>
<td>31</td>
</tr>
<tr>
<td>Christenson, Jennifer</td>
<td>70</td>
</tr>
<tr>
<td>Chung, Sun Joo</td>
<td>62</td>
</tr>
<tr>
<td>Chung, Yoo-Ree</td>
<td>63</td>
</tr>
<tr>
<td>Clark-Gareca, Beth</td>
<td>74</td>
</tr>
<tr>
<td>Colby-Kelly, Christian</td>
<td>37</td>
</tr>
<tr>
<td>Cook, H. Gary</td>
<td>45</td>
</tr>
<tr>
<td>Cox, Troy</td>
<td>62</td>
</tr>
<tr>
<td>Crusan, Deborah</td>
<td>63</td>
</tr>
<tr>
<td>Dambock, Barbara</td>
<td>60</td>
</tr>
<tr>
<td>Davis, Larry</td>
<td>64</td>
</tr>
<tr>
<td>Donovan, Anne</td>
<td>73</td>
</tr>
<tr>
<td>Downey, Ryan</td>
<td>35</td>
</tr>
<tr>
<td>Duijm, Klaartje</td>
<td>64</td>
</tr>
<tr>
<td>Eckes, Thomas</td>
<td>65</td>
</tr>
<tr>
<td>Elatia, Samira</td>
<td>44, 76</td>
</tr>
<tr>
<td>Evanini, Keelan</td>
<td>74</td>
</tr>
<tr>
<td>Fairclough, Marta</td>
<td>75</td>
</tr>
<tr>
<td>Fan, Jinsong</td>
<td>39</td>
</tr>
<tr>
<td>Fang, Felicia</td>
<td>65</td>
</tr>
<tr>
<td>Farnsworth, Timothy</td>
<td>49</td>
</tr>
<tr>
<td>Fletcher, Anna</td>
<td>75</td>
</tr>
<tr>
<td>Fung, Karen</td>
<td>44, 76</td>
</tr>
<tr>
<td>Gebril, Atta</td>
<td>63</td>
</tr>
<tr>
<td>Gennaro, Kristen di</td>
<td>40</td>
</tr>
<tr>
<td>Gibson Stephanie</td>
<td>78</td>
</tr>
<tr>
<td>Gonzales, Maria Lourdes Lira</td>
<td>61</td>
</tr>
<tr>
<td>Gorin, Joanna</td>
<td>69</td>
</tr>
<tr>
<td>Grant, Rosalie</td>
<td>45</td>
</tr>
<tr>
<td>Gray, Shelley</td>
<td>69</td>
</tr>
<tr>
<td>Haapakangas, Eeva-Leena</td>
<td>32</td>
</tr>
<tr>
<td>Hakola, Outi</td>
<td>76</td>
</tr>
<tr>
<td>Hamp-Lyons, Liz</td>
<td>41</td>
</tr>
<tr>
<td>Han, Sumi</td>
<td>38</td>
</tr>
<tr>
<td>Härmlä, Marita</td>
<td>33</td>
</tr>
<tr>
<td>Harman, Reanna</td>
<td>37</td>
</tr>
<tr>
<td>He, Lianzhen</td>
<td>47</td>
</tr>
<tr>
<td>Hegelheimer, Volker</td>
<td>34</td>
</tr>
<tr>
<td>Hildén, Riiil</td>
<td>76</td>
</tr>
<tr>
<td>Houston, Lorena</td>
<td>53</td>
</tr>
<tr>
<td>Hout, Roeland van</td>
<td>43</td>
</tr>
<tr>
<td>Hsu, Huei-Lien</td>
<td>35</td>
</tr>
<tr>
<td>Huang, Becky</td>
<td>38</td>
</tr>
<tr>
<td>Huhta, Ari</td>
<td>32</td>
</tr>
<tr>
<td>Jang, Eunice Eunhee</td>
<td>71</td>
</tr>
<tr>
<td>Jeon, Eun Hee</td>
<td>30</td>
</tr>
<tr>
<td>Jeon, Hun</td>
<td>61</td>
</tr>
<tr>
<td>Jin, Tan</td>
<td>44</td>
</tr>
<tr>
<td>Jin, Yan</td>
<td>39, 41</td>
</tr>
<tr>
<td>Jones, Glyn</td>
<td>66</td>
</tr>
<tr>
<td>Jong, John de</td>
<td>66</td>
</tr>
<tr>
<td>Judith, Ainsworth</td>
<td>61</td>
</tr>
<tr>
<td>Jung, Woojin</td>
<td>61</td>
</tr>
<tr>
<td>Kane, Michael</td>
<td>58</td>
</tr>
<tr>
<td>Kenyon, Dorry</td>
<td>78</td>
</tr>
<tr>
<td>Kim, Ahyoung Alicia</td>
<td>32</td>
</tr>
<tr>
<td>Lam, Jo Fond</td>
<td>35</td>
</tr>
<tr>
<td>Laurier, Michel</td>
<td>61</td>
</tr>
<tr>
<td>Lee, Hyesun</td>
<td>67</td>
</tr>
<tr>
<td>Lee, Yong-Won</td>
<td>34, 38</td>
</tr>
<tr>
<td>Li, Zhi</td>
<td>34</td>
</tr>
<tr>
<td>Lim, Gad</td>
<td>36, 40</td>
</tr>
<tr>
<td>Link, Stephanie</td>
<td>34</td>
</tr>
<tr>
<td>Liu, Li</td>
<td>44</td>
</tr>
<tr>
<td>Lopez, Alexis</td>
<td>77</td>
</tr>
<tr>
<td>Lu, Xiaofei</td>
<td>39</td>
</tr>
<tr>
<td>Lucks, Patricia Tosqui</td>
<td>68</td>
</tr>
<tr>
<td>Ma, Hong</td>
<td>34</td>
</tr>
<tr>
<td>Ma, Jia</td>
<td>77</td>
</tr>
<tr>
<td>MacGregor, David</td>
<td>78</td>
</tr>
<tr>
<td>Malone, Margaret E.</td>
<td>78</td>
</tr>
<tr>
<td>Mak, Barley</td>
<td>44</td>
</tr>
<tr>
<td>Massoud, Lindsey</td>
<td>60</td>
</tr>
<tr>
<td>McGinnis, Jenn</td>
<td>37</td>
</tr>
<tr>
<td>McNamara, Tim</td>
<td>55</td>
</tr>
<tr>
<td>Min, Shangchao</td>
<td>47</td>
</tr>
<tr>
<td>Moere, Alistair Van</td>
<td>35, 47</td>
</tr>
<tr>
<td>Montee, Megan</td>
<td>78</td>
</tr>
<tr>
<td>Murphy, Victoria</td>
<td>69</td>
</tr>
<tr>
<td>Naerssen, Margaret van</td>
<td>67</td>
</tr>
<tr>
<td>Neittaanmäki, Reeta</td>
<td>33</td>
</tr>
<tr>
<td>Nieminen, Lea</td>
<td>32</td>
</tr>
<tr>
<td>Nier, Victoria</td>
<td>80</td>
</tr>
<tr>
<td>Nikolov, Marianne</td>
<td>52</td>
</tr>
<tr>
<td>Oh, Saerhim</td>
<td>68</td>
</tr>
<tr>
<td>Phakiti, Aek</td>
<td>33, 45</td>
</tr>
<tr>
<td>Pill, John</td>
<td>42</td>
</tr>
<tr>
<td>Plakans, Lia</td>
<td>63</td>
</tr>
<tr>
<td>Raquel, Michelle</td>
<td>79</td>
</tr>
<tr>
<td>Raymundo, Natalia de Andrade</td>
<td>68</td>
</tr>
<tr>
<td>Restrepo, Maria Adelaia</td>
<td>69</td>
</tr>
<tr>
<td>Roever, Carsten</td>
<td>29</td>
</tr>
</tbody>
</table>
PTE ACADEMIC™

Standing out from the crowd!

PTE Academic is a leading computer-based English language test recognised worldwide.

- Assesses performance using innovative integrated item types
- Uses real-life tasks and authentic academic content
- Feedback on grammar, oral fluency, pronunciation, spelling, vocabulary and written discourse
- Comprehensive test preparation materials including: The Official Guide to PTE Academic
- The latest in biometric security to ensure test taker identity

To find out more visit www.pearsonpte.com or see us at our stand to find out more.

Presentations at LTRC

PAPERS
Investigating item variables influencing test taker performance on multiple choice items
Dr. Ying Zheng and Kerstan Ackermann

POSTERS
The EAQUALS CEFR-Based Assessment Special Interest Project
Glyn Jones
Native Speaker responses in language test development
Prof. Dr. John de Jong and Dr. Ying Zheng
Errata

Correction of two abstracts

Page 32

Charles Alderson, *Lancaster University*
Ari Huhta, *Jyväskylä University*
Lea Nieminen, *Jyväskylä University*
Riikka Ullakonoja, *Jyväskylä University*
Eeva-Leena Haapakangas, *Jyväskylä University*

Tuesday, April 3 4:30–5:00 Columbus–Savannah

**The Diagnosis and Treatment of Weaknesses in Reading and Writing in a Foreign Language**

The ability to read and write in a second or foreign language (SFL) is increasingly important in the modern world and therefore diagnostic language testing is potentially an important area of language test development and research. However, relatively little attention has been paid in language testing to the need to diagnose strengths and weaknesses in these important skills. Proficiency tests dominate the testing world, and although attempts have been made to derive useful diagnostic information from proficiency tests such tests are not designed to provide detailed, usable information. Furthermore, most testing studies concern English as the language to be learned and mastered; research on non-Indoeuropean languages is rare.

This paper reports on the design and results of a three-part, four-year (2010-2013) research project into the diagnosis of SFL reading and writing. The first part was a cross-sectional study comprising over 600 Finnish-speaking learners of English and about 250 Russian-speaking learners of Finnish. The Finnish-speakers comprised three age groups: 4th graders in primary school (age 10), 8th graders in lower secondary (age 14) and upper secondary students (age 17). The Russian-speakers consisted of primary and lower secondary students aged 10 to 15. The participants completed a range of L1 and SFL reading, writing and other linguistic tests, background and motivation questionnaires, as well as individually administered cognitive / psycholinguistic tasks.

The second study is longitudinal in nature, beginning in October 2011 and continuing until December 2013. In this study we are following primary and lower secondary students who participated in the first study. In addition to administering tests of reading and writing in English, we are also administering other instruments used in the first study as well as new instruments. We are observing the teaching of reading and writing in English and Finnish on a regular basis, and interviewing class teachers and selected learners.

In addition, within the longitudinal studies, we are carrying out a number of intervention studies, which are aimed at improving learners' abilities in a number of relevant areas: oral and silent reading fluency, vocabulary development, morphological and phonological awareness, reading and writing strategies. We are also investigating reading and writing processes with eye-tracking and keystroke logging approaches.
We will present key findings of the cross-sectional study on the predictors of SFL reading and writing based mainly on regression and SEM analyses of the data and focus on the similarities and differences of the two language groups. We will also report on the preliminary results of the ongoing longitudinal and intervention studies.

The results of these studies will give new insights into the process and components of perceiving and producing written form of language. They will also enable the construction of component profiles and a comparison of the development paths of reading and writing. The studies will thus be of considerable importance for a better understanding of how to both diagnose and treat learners' weaknesses, as well as contributing to a theory of SFL reading and writing development and diagnosis.
Elicited imitation, or sentence repetition, is a task that features in several automatically scored speaking tests. Test-takers listen to sentences one by one and must repeat each sentence word for word immediately after hearing it. Maximum credit is earned by repeating the sentence verbatim, but partial credit may be awarded depending on the number of mistakes or missing words in the test-taker’s utterance.

The task is suited to automated scoring because it elicits controlled language, and efficiently gathers rich information about a speaker’s proficiency in a short amount of time (Bernstein, Van Moere & Cheng, 2010). However, it is sometimes questioned whether the technology has driven the use of elicited imitation as a test task, or whether the task would also be valid in human administered tests. Moreover, since sentence repetition is not the same as spontaneous speech, it should be investigated whether the construct of the task generalizes to real-life communication.

Decades of research on elicited imitation (Vinther, 2002; Erlam, 2006) in the fields of second language acquisition (SLA) and bilingualism have revealed that the task has much to offer language assessment: it yields stable test-retest correlation; is a good predictor of performance in natural language situations; reflects stages of language development; mirrors familiar classroom activities such as rehearsal and drilling, which reinforce language learning; consists of a process of decoding, interpretation and subsequent (re)production of the stimulus sentence; consists of both comprehension and production skills; and is a test of knowledge of structures and semantics, since sentences are decontextualized and learners cannot rely on contextual clues or strategies for assistance. Further, parroting can only occur when the stimulus sentence is short enough to be retained in immediate memory as an acoustic image, which is approximately seven syllables (Miller, 1956).

This presentation critically reviews the literature on elicited imitation with regard to its application in language assessment. It then reports on several experiments (n>100) involving (1) comparison of scores on elicited imitation with scores on tasks involving spontaneous speech, including storytelling, using correlation, multi-method multi-trait analysis, and G-theory, and (2) comparison of elicited imitation under two different conditions: immediate repetition and delayed repetition. The rationale for investigating (2) is because some research has shown that delayed repetition does more to prevent parroting and alters the construct to require the test-taker to focus on meaning rather than form; thus, meaning is held in short-term memory and sentence syntax is reconstructed from meaning. Together, the experiments address the research questions: “To what extent does performance on elicited imitation generalize to performance on test tasks with spontaneous speech?” and “Which variation of the elicited imitation task is most reliable for separating test-takers according to ability?”

The presentation concludes by discussing sentence repetition in the context of language assessment and technology. Further research required to support the validity of the task includes memory variables, coaching, and perceived narrowness of the construct with respect to the social aspects communication.
IELTS Research Reports
Volume 13

Now available

Volumes 1 – 13 are available at: www.ielts.org

For IELTS, quality and fairness have always been paramount. Our continuing programme of research and validation activities is a key element in delivering quality, fairness and a fit for purpose test.

www.ielts.org
Assessment and Learning: Bridging Diverse Disciplines and Domains