From language testing to language assessment: Connecting teaching, learning and assessment.
More than four decades of research supports the validity of the TOEFL® Family of Assessments

Read the latest journal articles and reports on the TOEFL® Family of Assessments.


<table>
<thead>
<tr>
<th>Page</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ILTA President’s Welcome Message</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>LTRC 2015 Chair’s Welcome Message</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>LTRC 2015 Co-Host CALA/ACEL Welcome Message</td>
<td>42</td>
</tr>
<tr>
<td>8</td>
<td>Samuel J. Messick Memorial Lecture</td>
<td>46</td>
</tr>
<tr>
<td>11</td>
<td>The Davies Lecture</td>
<td>66</td>
</tr>
<tr>
<td>13</td>
<td>Cambridge/ILTA Distinguished Achievement Award Lecture</td>
<td>82</td>
</tr>
<tr>
<td>14</td>
<td>Conference Sponsors</td>
<td>101</td>
</tr>
<tr>
<td>17</td>
<td>Conference Exhibitors</td>
<td>123</td>
</tr>
<tr>
<td>19</td>
<td>Conference Support</td>
<td>145</td>
</tr>
<tr>
<td>21</td>
<td>ILTA 2015</td>
<td>161</td>
</tr>
<tr>
<td>22</td>
<td>Award Winners</td>
<td></td>
</tr>
</tbody>
</table>

Conference Overview
Conference Schedule
Pre-Conference Workshop Abstracts
Paper Abstracts: Wed, Mar. 18
Paper Abstracts: Thurs, Mar. 19
Paper Abstracts: Fri, Mar. 20
Symposium Abstracts
Poster Abstracts
Work-in-Progress Abstracts
Index
Studies in Language Testing

An indispensable resource for anyone interested in new developments and research in language testing

Examining English – the complete set

To find out more about our full list of publications:

www.cambridge.org/elt/silt
www.cambridgeenglish.org/silt
Welcome to LTRC 2015

As the incoming President of the International Language Testing Association, it is my great pleasure to welcome you to the 37th Language Testing Research Colloquium. After ten years it's great to be back in Canada, in many ways our spiritual home since ILTA was established here in 1992. This time we are in the dynamic city of Toronto, which, just like ILTA, is home to a diverse and lively (but generally harmonious) population. Reflecting ILTA's links with regional associations, this year's conference is held in conjunction with the Canadian Association of Language Assessment/Association canadienne pour l'évaluation des langues (CALA/ACEL), one of ILTA's growing list of affiliates.

This year's conference theme – From Language Testing to Language Assessment: Connecting Teaching, Learning, and Assessment – reflects the broadening scope of our field and increasing recognition of the pivotal role of testing and assessment in educational systems. It is therefore particularly appropriate that this year LTRC is timed to coincide with two other conferences relating to language education: AAAL (March 21-24) and TESOL (March 25-28). Indeed, AAAL will open on the morning of 21 March with a joint colloquium, open to LTRC delegates. The theme will be 'Revisiting the Interfaces between SLA and Language Assessment Research.' Many of the authors contributing to the landmark 1999 publication edited by Lyle Bachman and Andrew Cohen will take a fresh look at the issues it raised.

As usual LTRC, with generous support from our sponsors, gives us the chance to recognize the achievements of many of our members. The ILTA/Cambridge Distinguished Achievement Award recognizes outstanding contributions to our field. The ILTA Best Article Award and Sage/ILTA Book Award acknowledge excellence in scholarship and research. We encour-

age the professional development of MA and PhD students in language assessment through the Student Travel Award and the Lado Award for the best student presentation. We promote education in language assessment with our annual Workshops and Meetings Awards. These and other awards will be presented and celebrated in the magnificent setting of Hart House at a ceremony held during the gala banquet on Friday night.

Before relaxing and enjoying the banquet, we also have ILTA business to conduct and I hope as many members as possible will attend the Annual Business Meeting, to be held on Thursday at 12.30 pm. There is plenty to discuss and important questions to decide: venues for future LTRCs, plans for future initiatives and updates on ILTA activities. Not yet a member of ILTA? There is still plenty of time to go online and join before the meeting.

ILTA only exists because of the hard work of our many volunteers. This year I'd particularly like to thank Erik Voss for the excellent and extensive work he has done to improve the ILTA website at iltaonline.com. If you have not yet visited the new site launched in January, please do take a look. Of course, organising a conference is also a monumental task and I am most grateful to the conference chair, Liying Cheng, the organizing committee of Khaled Barkaoui, Heike Neumann, Beverly Baker and Christine Doe and everyone involved for arranging and managing such a packed and varied program: 63 research papers, 28 posters, 20 works in progress and seven symposia. Each day starts off with a plenary session: the Samuel J. Messick Memorial Lecture, the ILTA/Cambridge Distinguished Achievement Award Lecture, and the Davies Lecture. In addition to the established session types, this year there are two Professional Events (one on publishing and the other on career opportunities in Language Assessment) and the conference closes with another exciting innovation: a debate that is set to keep us all engaged in lively (but generally harmonious) discussion until we gather again in Palermo in 2016.

Best wishes,

Tony Green,

ILTA President
As Paragon develops and produces its assessments, it strives for the following:

INNOVATION

VALIDITY

FAIRNESS

ACCURACY

Paragon encourages and financially supports collaborative research relationships with professional researchers and graduate students.

For more information, visit www.paragontesting.ca.

Paragon Testing Enterprises is a subsidiary of the University of British Columbia
Welcome to LTRC 2015

As the conference chair of LTRC 2015 Toronto, I warmly welcome you to the 37th Language Testing Research Colloquium. I’d like to take this opportunity to first thank the organization committee: Khaled Barkaoui (local chair), Beverly Baker, Christine Doe, and Heike Neumann for their countless hours, enormous efforts, and constant commitment to bringing this conference to its fruition.

This year, we received 246 proposals. Each proposal was reviewed by 2-3 reviewers. After a thorough and rigorous review process, we proudly present 63 papers, 7 symposia, 28 posters and 20 Works in Progress. My special thanks go to those colleagues who took the time and reviewed 12-14 proposals within 10-12 days! Many thanks also go to the generous financial support of all our LTRC 2015 sponsors. I am grateful for the unfailing support of the presidents and the executive board of ILTA, and Robert Ranier, the ILTA Account Manager. I also would like to thank our workshop and conference presenters, and all the conference participants. There is no LTRC 2015 without your support and your contributions!

It has been three years since I first proposed to host LTRC2015 jointly with our newly established Canadian Association of Language Assessment/Association canadienne pour l’évaluation des langues. I’d like to take this opportunity to share with you three personal connections that I have with LTRC 2015 Toronto: the theme, the interdisciplinary nature, and the location of this colloquium. The theme From Language Testing to Language Assessment: Connecting Teaching, Learning, and Assessment reflects the growth of our field, its movement away from a narrow focus on tests and testing to a broader focus on assessment, and connects to my own research on washback over the past 20 years.

LTRC 2015 takes place in the same city and month as the annual conference of the American Association of Applied Linguistics (AAAL) and the TESOL International Association convention. This joint opportunity offers the time and space for us as a distinct, yet interrelated, sub-field within applied linguistics, to revisit the interdisciplinary nature of our field. Who are we? Do we remember the days and nights when we were the language teachers? This is also the first time that we have a representative from the TESOL International Association speaking at LTRC. We continue to strengthen our relationships with AAAL by presenting a joint session at AAAL 2015 on Revisiting the Interfaces between SLA and Language Assessment Research. LTRC 2015 participants can attend this session for free. Lastly, I’d like to connect this LTRC 2015 to Canada. Ten years ago (2005), the 27th LTRC was held in Ottawa, Canada. And fifteen years ago (2000), the 22nd LTRC was held in Vancouver, Canada. It was during LTRC 2000 that the ILTA code of ethics was established.

I hope everyone in our community will embrace the opportunities presented during LTRC 2015, and I look forward to meeting all of you.

Liying Cheng,
LTRC 2015 Conference Chair
The Canadian Association of Language Assessment (CALA) was officially established in 2009. We have over 100 members from academia as well as other educational, institutional and government settings. The general objective of the Association is to promote best practices in language assessment design, practice, research, and scholarship in the Canadian context, in both academic and occupational/professional settings.

L’Association canadienne pour l’évaluation des langues (ACÉL) est un regroupement de plus de 100 chercheurs et d’autres professionnels dans le domaine d’évaluation de la langue au Canada. L’ACÉL existe depuis 2009. L’Association a pour but général de promouvoir des pratiques de qualité dans le domaine d’évaluation des langues au Canada, en tenant compte des domaines académiques ainsi que des domaines professionnels.
The International Center for Language Research and Development (ICLRD) is cofounded by Higher Education Press (HEP) (http://eng.hep.com.cn/EnglishHep/index.shtml) and Guangdong University of Foreign Studies. The Center’s mission is to promote the development of language education through research, testing and evaluation, teacher training and academic exchange. It provides consulting services to universities and government on the reform and development of language education.

Considering the status of foreign language education in China, the language assessment policy and development as well as the current needs of foreign language talents in society, ICLRD is committed to developing the assessment programs in accordance with the tendency of talents cultivation of foreign language education. The national wide assessment programs and business is under construction in areas of Business English, Vocational English, Foreign Language Translation and Interpretation, etc.

To facilitate ELT professional development, ICLRD has been organizing series of international, national and local teacher training programs in cooperation with the Higher Education Institutions Teacher Online Training Center under the Ministry of Education. The training programs mainly focus on teachers teaching ability and research ability development, which has been greatly welcomed and appreciated by EFL teachers.

In addition, ICLRD has also organized a number of major events and activities like Micro-course Contest of Foreign Languages in China and Asia TEFL Conferences (upcoming 13th Asia TEFL will be held in Nanjing, Nov. 6-8 2015) to promote the integration of language teaching, research and technology.

Contact Information: Email: irdef@126.com Tel: 86 010-58556584

iSmart (Digital intelligent solution of language teaching and learning)

iSmart test
iSmart is a digital intelligent solution developed by HEP. It’s open, stable, rich in comprehensive skills training and practice forms. It provides social function as well as gaming motivation. It provides intelligent, digital resources and services for colleges and universities to achieve “iSmart intelligent language learning campus”. The scheme includes special training system such as listening, speaking, reading, writing and translating, and multi-evaluation system, to realize unified identity authentication, unified learning process management, unified score report. iSmart platform fully supports ios and Android mobile devices to provide excellent service of ubiquitous learning and lifelong learning for both college teachers, students and social users. Meanwhile, iSmart platform is seamlessly connected to iCourse, which is the campus end of fine-designed open course of the Ministry of Education, it also provides teaching management, curriculum management and communication platform between teachers and students.

System features

- Perfect solution to the daily homework, grading test, mid-term exams, final exams and other assessment requirements
- Fully support mock tests of written and computer-based test such as College English Test (CET) 4, CET6, the National Business English Level 4 and Level 8 Examinations, Practical English Test for Colleges Level A and Level B, the National Vocational Skills Contest, PETS, BETS, IELTS, TOEFL IBT etc.
- Support time-limited exams, computer-based tests, completely open independent evaluation and daily operations
- Support oral English debating contest for multi-users at the same time
- Support automatic speech recognition of oral testing
- With powerful data analysis function, support deep mining of performance information to analyses students’ knowledge and skills and perfectly support teaching, research as well as evaluation of teaching reform
- Discipline: English for non-English majors English for English majors English for postgraduate level
- Knowledge framework: Subdivide knowledge according to skills of listening, speaking, reading, writing, translating, or topics of grammar, vocabulary, etc.

Contact Information: hepismart@139.com
Validity theory and empirical methods have largely been developed in the context of large-scale testing (Bonner, 2013; Messick, 1989) and do not always fit classroom assessment purposes (Brookhart, 2003). Appropriate methods and good examples of validation of classroom assessment information are scarce (Bonner, 2013). Gaining perspective on how validation should proceed in classroom contexts is important for the theme of this conference: Connecting Teaching, Learning, and Assessment. It is also important for language testing more generally because the intended use of much language testing information is referenced to students’ future learning in the classroom.

The purpose of this lecture is to suggest a framework for validation research that is appropriate in the classroom context. It builds on previous work in this area (Bonner, 2013; Brookhart, 2003; Llosa, 2008; Moss, 2003). It is informed by recent developments in the theory of self-regulation of learning (Zimmerman & Schunk, 2011) and by the understanding that learning is a social and emotional, as well as cognitive, process (Leighton, Chu, & Seitz, 2013). It is also informed by multiple perspectives on validity: Kane (2012) identifies the measurement perspective, the contestant perspective (see also Dorans, 2012), and the pragmatic perspective on validity.

The measurement perspective on validity treats students as examinees. As such they are (ideally) randomly sampled, representative of some population, and, importantly, non-reactive. The contestant perspective treats students as contestants who aim to do as well as possible on an assessment. The pragmatic perspective, which is most often taken by those who run accountability programs, treats students as participants, consumers of educational programs where the purpose of the program is to maximize achievement for groups of students. Drawing some concepts from each of these approaches, a classroom-centered (or “classroometric,” Brookhart, 2003) perspective on validity considers students as learners, both individually and in groups.

As self-regulated learners, students sit on both sides of the testing event, as both “those who are measured” and as one of the users of the information (“those who must do something about it”). In other words, the inferences made and actions taken are part of the learning process, not external to it as they are for large-scale assessment (Brookhart, 2003). The interpretive argument for validity (Kane, 2006) shifts from meaningful inferences and effective uses of test score information (Messick, 1989, 1996) to an understanding of how students’ work compares to learning goals (or, in some language testing contexts, readiness for learning goals). When the users are learners, their information needs and their use of assessment information for learning become relevant to validation. As a corollary, the measurement context (how and when students are able to use their knowledge) is construct-relevant.
The proposed framework for validation research will meet two challenges: to exploit these rich connections among teaching, learning, and assessment and at the same time to seek parsimony so that the classroom perspective remains a useful tool for examining validity and is not overly broad.

**Susan M. Brookhart, Ph.D.,** is an independent educational consultant and author based in Helena, Montana. She also currently serves as a Senior Research Associate in the Center for Advancing the Study of Teaching and Learning in the School of Education at Duquesne University, where she was a full-time faculty member for fourteen years. Dr. Brookhart’s research interests include the role of both formative and summative classroom assessment in student motivation and achievement, the connection between classroom assessment and large-scale assessment, and grading. Dr. Brookhart was the 2007-2009 editor of Educational Measurement: Issues and Practice, a journal of the National Council on Measurement in Education. She is author or co-author of sixteen books and over 70 articles and book chapters on classroom assessment, teacher professional development, and evaluation. She serves on several editorial boards and national advisory panels. She has been named the 2014 Jason Millman Scholar by the Consortium for Research on Educational Assessment and Teaching Effectiveness (CREATE).
English courses and testing aligned to a single scale: the Global Scale of English

The Global Scale of English is a standardised granular scale from 10-90 for measuring English language proficiency. It forms the backbone of Pearson English course materials and assessment, including Progress the new fully automated online test. Progress accurately measures student progression in English, highlighting strengths and weaknesses to inform teaching.

Visit English.com/GSE or PearsonELT.com/Progress to learn more.
Four conceptually-distinct options exist for relating assessments of writing to teaching and learning in programs of language education, each fulfilling different but interrelated purposes. The assessment purposes may either be normative, diagnostic, formative, or summative. The options are (a) proficiency tests and curriculum standards, each based on different kinds of normative principles and data, (b) diagnostic and dynamic assessments focused on individual learners and their learning potential within a specific educational context, (c) responding to students’ written drafts for formative purposes of informing their improvement, and (d) grades or local tests of summative achievements in a particular course.

In educational practices these assessment options often converge, and their combinations may even be necessary to achieve effective instruction and program organization, but their purposes and conceptualizations are easily confused. For these reasons, assessors and educators need to distinguish the respective values, functions, and limitations of each assessment option so as to avoid confusing them when making particular policy and pedagogical decisions.

Alister Cumming, Ph.D., is a professor in the Centre for Educational Research on Languages and Literacies (CERLL, formerly the Modern Language Centre) at the Ontario Institute for Studies in Education, University of Toronto, where he has been employed since 1991 following briefer periods at the University of British Columbia, McGill University, Carleton University, and Concordia University. His research and teaching focus on writing in second languages, language assessment, language program evaluation and policies, and research methods. Alister’s recent books are Agendas for Language Learning Research (with Lourdes Ortega and Nick Ellis, Wiley-Blackwell, 2013), Adolescent Literacies in a Multicultural Context (Routledge, 2012), A Synthesis of Research on Second Language Writing in English (with Ilona Leki and Tony Silva, Routledge, 2008), and Goals for Academic Writing (John Benjamins, 2006).
Shanghai Foreign Language Education Press (SFLEP) was founded in 1979 as an affiliate to Shanghai International Studies University. It dedicates to foreign language education, promotion of academic achievements and dissemination of culture.

It has established itself as:
- An award-winning publisher for over 35 years
- More than 1000 institutions of higher education in China are using College English course books it publishes
- 6000 titles of books in 23 languages, among them, 500 have won awards at provincial and national level
- A state-of-the-art expertise in publishing and a reputation for high quality course books, academic works, reference books, dictionaries, and more

Website: www.sflep.com

Foreign Language Testing and Teaching (FLTT) is an academic journal published by Shanghai Foreign Language Education Press. Making its debut in 2011, the quarterly publishes original researches on foreign language testing and assessment in China. It reports theoretical issues, empirical studies, and reviews on the major language testing and assessment researches, such as CET (college English tests), TEM (tests for English majors), NMET (national matriculation English test) and international languages tests. Currently, the journal is interested in publishing more studies on testing policies, technical matters, testing literacy, washback and impact of language tests. With an eye on future, it encourages pioneering interdisciplinary researches, submissions on theoretical and methodological researches in different fields of applied linguistics and relevant disciplines.

Eminent testing professionals from prestigious colleges across China sit on this journal’s editorial board. Recently, FLTT joined efforts with National College English Testing Committee and Advisory Board for Teaching and Testing of Higher Education Foreign Language Majors, together with local testing research institutes, to host a series of workshops and symposiums, and effectively promoted testing researches in China.

FLTT is also actively seeking more interaction and cooperation possibilities with testing research bodies and individuals outside China, and welcomes submissions that methodically study various aspects of tests and assessments in both international and Chinese contexts.

ISSN 2095–1167
Contact us: fltt@sflep.com
The trajectory of research in any field is a mixture of opportunities created by broader developments in theory and practice and the chance elements of individual histories. I try to illustrate this by looking back at turning points in my own career as a language tester as it unfolded in the context of broader trends in communicative language testing, performance assessment, the development of new measurement techniques, particularly Rasch measurement, the discursive and social turn in applied linguistics and language testing, and the challenge presented by the relatively new field of English as a lingua franca. What enables a language tester to grow as a researcher in an unpredictable and constantly evolving intellectual and research environment? Good luck, certainly, particularly the luck of encountering and learning from the work of colleagues who act as role models and mentors and who are leading the exploration of points of significant change. But also openness to the unexpected, underpinned by intellectual interests that go beyond the specific familiar topics of a single research field. I argue that breadth as well as depth of intellectual preparation, and the encouragement to think laterally, should be emphasized in the education of the next generation of language testing researchers.

Tim McNamara, has taught Applied Linguistics at Melbourne since 1987. He helped establish the graduate program in Applied Linguistics from 1987, and with Professor Alan Davies founded the Language Testing Research Centre. His language testing research has focused on performance assessment, theories of validity, the use of Rasch models, and the social and political meaning of language tests. Tim has acted as a consultant with Educational Testing Service, Princeton where he worked on the development of the speaking sub-test of TOEFL iBT; he was also part of the team involved in the original development of IELTS. His work on language and identity has focused on poststructuralist approaches to identity and subjectivity. Tim is the author of Measuring Second Language Performance (Longman, 1996), Language Testing (OUP, 2000) and co-author (with Carsten Roever) of Language Testing: The Social Dimension (Blackwell, 2006). He is currently working on a book entitled Language and Subjectivity, to be published by De Gruyter. Tim has been elected 2nd Vice-President of the American Association for Applied Linguistics (AAAL) and will be Conference Chair for the 2017 conference.
Conference Sponsors

British Council
http://www.britishcouncil.org/aptis
Banquet Gala
Newcomer Session

ETS TOEFL and TOEIC
www.ets.org
Conference Bags
Conference T-shirts
Subsidy for Student Registration
Subsidy for Student Volunteers

University of Cambridge ESOL Examinations
www.cambridgeesol.org
Opening Reception

Paragon
www.paragontesting.ca.
Preconference Workshop Refreshments
Conference Badges
Local Organization Sponsor

Pearson Tests of English
www.pearsonpte.com
Refreshment Breaks (March 20)

Shanghai Foreign Language Education Press
http://www.sflep.com/english/about_us.html
Refreshment Breaks (March 18)
Conference Sponsors

Foreign Language Teaching and Research Press
http://en.fltrp.com/
Refreshment Breaks (March 19)

Canadian Academic English Language (CAEL) Assessment
www.cael.ca
Refreshment Breaks (March 19)

Faculty of Education, York University
www.edu.yorku.ca
Laptops

OISE, University of Toronto
www.oise.utoronto.ca
Banquet location

Routledge
www.routledge.com/
Refreshment Breaks (March 18)

Faculty of Education, Queen’s University
http://educ.queensu.ca
Graduate Students and Photographers

Touchstone Institute
www.touchstoneinstitute.ca
Toronto Information Flyer
Find out why so many organizations trust CaMLA for their language assessments.

CaMLA, a joint venture between the University of Cambridge and the University of Michigan, combines the expertise of two world-class universities, each in its own right a leader in language assessment. Together, we offer more research to support quality test development. More exams in more places. More availability when you need it.

With CaMLA, you can count on an experienced, international, responsive partner.

CambridgeMichigan.org
CHINA LANGUAGE ASSESSMENT

中国外语测评中心

Provide diversified language assessment services
Develop standards for evaluating language teachers
Conduct language education and assessment research

IN THE SERVICE OF ALL LEARNERS

Co-established by

www.fltrp.com
www.sinotefl.org.cn
LTRC Organizing Committee 2015

Chair: Liying Cheng, Faculty of Education, Queen’s University
Local Chair: Khaled Barkaoui, Faculty of Education, York University
Beverly Baker, University of Ottawa
Christine Doe, Mount Saint Vincent University
Heike Neumann, Concordia University

LTRC 2015 Website
Tom Fullerton
Beverly Baker

LTRC 2015 Program Book and Logo Designer
Kris Pierre Johnston, kpj@yorku.ca

LTRC 2015 Program Preparation Support
Jessica Chan
Jia Ma
Yi Mei
Cheng Zhou

LTRC 2015 Session Chairs
Poster chair: Yong-Won Lee
WIP chair: Lorena Llosa

Professional Events Chairs
Publishing: Guoxing Yu
Career: Eunice Jang

Abstract Reviewers
Beverly Baker
Jayanti Banerjee
Liying Cheng
Ikkyu Choi
Christian Colby
Alister Cumming
Fred Davidson
Larry Davies
Barbara Dobson
Samira ElAtia
Janna Fox
April Ginther
Anthony Green
Luke Harding
Ching-Ni Hsieh
Ari Huhta
Talia Isaacs
Eunice Jang
Yan Jin
Dorry Kenyon
Yong-won Lee
Lorena Llosa
Sari Luoma
Lynnette May
Heike Neumann
Spiros Papageorgiou
Lia Plakans
David Qian
John Read
Daniel Reed
Steven Ross
Shahrzad Saif
Yasujo Sawaki
Toshihiko Shiotse
Youngsoon So
Charles Stansfield
Lynda Taylor
Carolyn Turner
Elvis Wagner
Yoshinori Watanabe
Sara Weigle
Guoxing Yu
Xiaoming Yu
Ying Zheng

Volunteers
Farhana Ahmed
Shaheda Akter
Poonam Anand
Bo Bao
Esther Bettney
Tina Beynen
Qinghui Chen
Andrew Coombs
Shaima Dashti
Jessica Freitag
Wenqian Fu
Valerie Kolesova
Jeannie Larson
Zhi Li
Jia Ma
Yi Mei
Jason Yoongoo Nam
Gina Park
Shujiao Wang
Yiqi Wang
Yan Wei
Yongfei Wu
Luna Yang
Cheng Zhou
IELTS Joint Funded Research Program 2015

Educational institutions and suitably qualified individuals are invited to apply for funding to undertake applied research projects in relation to IELTS for a period of one or two years. Projects will be supported up to a maximum of £45,000/AU$70,000.

Application deadline: 30 June 2015

Download an application form from: www.ielts.org/research_grants
ILTA Executive Board 2015

President: Anthony Green, University of Bedfordshire, UK
Vice-President: Catherine Elder, University of Melbourne, AUS
Secretary: Margaret Malone, Center for Applied Linguistics, US
Treasurer: Jayanti Banerjee, CAMLA, US
Immediate Past President: Dan Douglas, Iowa State University, US

Members at Large
Kirby Grabowski, Teachers College, Columbia University, US
Eunice Eunhee Jung, OISE, University of Toronto, CAN
Ute Knoch, University of Melbourne, AUS
Elvis Wagner, Temple University, US

LTRC 2015 Chair
Liying Cheng, Queen’s University, CAN

ILTA Nominating Committee 2015
Chair: Barbara Dobson, CAMLA, US
Rachel Brooks, Federal Bureau of Investigation, US
Luke Harding, Lancaster

University, UK
Mikyung Kim Wolf, Educational Testing Service, US

2015 Award Committees

Cambridge/ILTA Distinguished Achievement Award
Chair: Cyril Weir, University of Bedfordshire, UK
Hanan Khalifa, Cambridge English Language Assessment, UK
Antony Green, University of Bedfordshire, UK
Dan Douglas, Iowa State University, US
Meg Malone, Center for Applied Linguistics, US

ILTA Student Travel Awards
Chair: Dan Douglas, Iowa State University, US
Tony Green, University of Bedfordshire, UK
Ari Huhta, University of Jyväskylä, FI
Jessica Wu, The Language Training & Testing Center (LTTC), TWN

Lado Award
Chair: Jonathan Schmidgall, Educational Testing Service US
Eunice Jang, University of Toronto, CAN
Maryam Wagner, University of Toronto, CAN
Kirby Grabowski, Columbia University, US

ILTA Best Article Award 2012
Chair: Ute Knoch, University of Melbourne, AUS
Dory Kenyon, Centre for Applied Linguistics, US
Nathan Car, California State University, Fullerton, US
Jessica Wu, Language Training and Testing Centre, TWN
Rie Koizumi, Juntendo University, JPN

ILTA Best Article Award 2013
Chair: Elvis Wagner, Temple University, US
Rie Koizumi, Juntendo University, JPN
Ute Knoch, University of Melbourne, AUS
Mehdi Riazi, Macquarie University, AUS
Claudia Harsch
University of Warwick, UK
Cambridge/ILTA Distinguished Achievement Award 2015

Tim McNamara, University of Melbourne, AUS

ILTA Best Article Award 2012


ILTA Best Article Award 2013


ILTA Best Student Paper 2015

TBA March 20

ILTA Student Travel Awards 2015

Laura Ballard & Shinhye Lee, Michigan State University

Yueting Xu, Hong Kong University, HK

Naoki Ikeda, University of Melbourne, AUS

Jacqueline Ross TOEFL Dissertation Award 2015

Hans Rutger Bosker, Utrecht University, NLD. The Processing and Evaluation of Fluency in Native and Non-Native Speech. Supervisors: Dr. Nivja H. de Jong & Dr. Hugo Quené

TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment 2014

<table>
<thead>
<tr>
<th>RECIPIENT</th>
<th>INSTITUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ashley Chrzanowski</td>
<td>University of Colorado at Boulder</td>
</tr>
<tr>
<td>Justin Cubilo</td>
<td>University of Hawaii at Manoa</td>
</tr>
<tr>
<td>Iftikhar Haider</td>
<td>University of Illinois at Urbana-Champaign</td>
</tr>
<tr>
<td>Wei-Li Hsu</td>
<td>University of Hawaii at Manoa</td>
</tr>
<tr>
<td>Naoki Ikeda</td>
<td>University of Melbourne</td>
</tr>
<tr>
<td>Kunlaphak Kongsuwannakul</td>
<td>University of Leicester</td>
</tr>
<tr>
<td>George LaFlair</td>
<td>Northern Arizona University</td>
</tr>
<tr>
<td>Jia Ma</td>
<td>Queen’s University</td>
</tr>
<tr>
<td>Khoi Ngoc Mai</td>
<td>The University of Queensland</td>
</tr>
<tr>
<td>Nathaniel Owen</td>
<td>University of Leicester</td>
</tr>
<tr>
<td>Woranon Sitajalabhorn</td>
<td>University of Melbourne</td>
</tr>
<tr>
<td>Jayoung Song</td>
<td>University of Texas at Austin</td>
</tr>
<tr>
<td>Jing Xu</td>
<td>Iowa State University</td>
</tr>
</tbody>
</table>

Caroline Clapham IELTS Masters Award, 2014

Lorraine Briegg-Jones, Newcastle University, UK. An investigation into nonverbal behaviour in the oral proficiency interview. Supervisor: Dr Dawn Knight
TIRF Doctoral Dissertation Award

Justin Cubilo, University of Hawaii at Manoa, US. Video-Mediated Listening Passages and Typed Notetaking: Investigating Their Impact on Comprehension, Test Structure, and Item Performance.

Supervisor: James Dean Brown

Iftikhar Haider, University of Illinois at Urbana-Champaign. Re-Envisioning Assessment of Interlanguage Pragmatics (ILP) through Computer Mediated Communicative Tasks.

Supervisors: Fred Davidson & Melissa Bowles

Jia Ma, Queen’s University, CAN. Chinese Students’ Journey to Success on High stakes English Language Tests: Nature, Effects and Values of Their Test Preparation.

Supervisor: Liying Cheng

Jing Wei, New York University, US. Assessing Speakers of World Englishes: The Roles of Rater Language Background, Language Attitude and Training.

Supervisor: Lorena Llosa

Canadian Academic English Language (CAEL) Assessment

- Recognized as a reliable assessment of English language proficiency for over 20 years.
- Tests a student’s ability to use English in an academic context.
- Closely resembles the experience of being in a tertiary level classroom where the readings, lecture and the writing components are all on the same topic.
- Presents a constructed-response question format with tasks common to the Canadian university context.
- Item types include: filling in the blanks; labeling diagrams; completing tables; note-taking; short answer and longer essay.
- Promotes the interests of higher education in Canada.
- Accepted across Canada and in a growing number of European and American institutions.

www.cael.ca
Introducing Queen’s New Online Graduate Programs
Part-time professional programs for teachers and other education professionals

Queen’s Faculty of Education is pleased to announce the launch of its new online graduate programs, which are designed for educators who wish to improve their professional practice and/or take their career to the next level. Courses are approximately 36 hours online with your professor & 70-100 hours independently. To accommodate work schedules, the courses are 7 weeks in summer and 10 weeks in fall and winter. You can enrol in the graduate diploma program & then decide whether to apply to the professional master’s or you can enrol directly in the professional master’s degree program.

The Graduate Diploma in Professional Inquiry
Expand your knowledge of professional inquiry by completing 5 online courses.

The Professional Master of Education
Choose a concentration, such as Assessment and Evaluation, Literacy Education or Teaching Abroad, and complete 5 graduate diploma courses and 5 professional master’s courses (3 courses in a concentration, a course in organizational leadership and an elective).

Contact us
For further information about our new programs, or our on-campus Master’s and PhD programs, please contact us:

Online: educ.queensu.ca   Email: edonline@queensu.ca   Phone: 613-533-6000 ext. 78530
Official Languages and Bilingualism Institute (OLBI)
613-562-5743 | olbi@uOttawa.ca
olbi.uOttawa.ca

Institut des langues officielles et du bilinguisme (ILOB)
613-562-5743 | ilob@uOttawa.ca
ilob.uOttawa.ca

Proud leader in the assessment of Canada’s official languages
Contact Information: 613-562-5743 | l2test@uOttawa.ca

Offering a Master of Arts in Bilingualism Studies with the opportunity
to specialise in language assessment
Contact information: 613-562-5800 ext. 3384 | paribakh@uOttawa.ca
Mon, Mar 16
8:30am–4:00pm
Conference Registration (Churchill Court)
9:00am–4:00pm
Workshops 1, 2, & 3 (Churchill A & B and Rossetti)

Tues, Mar 17
8:30am–6:30pm
Conference Registration (Churchill Court)
9:00am–4:00pm
Workshops 1, 3 & 4 (Churchill A & B and Rossetti)
12:00–5:00pm
ILTA Executive Board Session (Windsor)
4:30–6:00pm
LTRC Newcomers’ Session (Rossetti)
6:00–8:00pm
Opening Reception (Churchill Court)

Wed, Mar 18
8:00am–5:00pm
Conference Registration (Churchill Court)
8:30am–5:00pm
Exhibition (Churchill Court)
8:45–9:00am
Opening Welcome & Announcements (Churchill Ballroom)

9:00–10:00am
Samuel J. Messick Memorial Lecture (Churchill Ballroom)
10:00–10:20am
Coffee Break (Churchill Court)
10:20–11:45am
Paper Session 1 (Churchill A & B, Mountbatten A & B)
11:45–1:30pm
Group Photo and Lunch Break
11:45–1:30pm
LAQ Editorial Board Meeting (Rossetti)
1:30–3:00pm
Poster Session (Mountbatten Lane)
3:00–4:25pm
Paper Session 2 (Churchill A & B, Mountbatten A & B)
4:25–4:45pm
Break (Churchill Court)
4:45–6:15pm
Symposium Session 1 (Churchill Ballroom)
6:30–7:30pm
Publishing in Language Assessment (Mountbatten Salon)

Thurs, Mar 19
8:30am–5:00pm
Conference Registration (Churchill Court)
8:30am–5:00pm
Exhibition (Churchill Court)
8:45–9:00am
Announcement (Churchill Ballroom)

9:00–10:00am
The Davies Lecture (Churchill Ballroom)

10:00–10:20am
Coffee Break (Churchill Court)

10:20–11:45am
Paper Session 3 & Symposium Session 2 (Churchill A & B, Mountbatten A & B)

11:45–1:30pm
ILTA Business Meeting (Churchill Ballroom)

Lunch Break

1:30–3:00pm
Work in Progress (Mountbatten Salon)

3:00–4:25pm
Paper Session 4 (Churchill A & B, Rossetti)

4:25–4:45pm
Coffee Break

4:45–6:15pm
Symposium Session 3 (Churchill A & B)

Fri, Mar 20

6:30–7:30 pm
Career in Language Assessment (Mountbatten Salon)

Sat, Mar 21

8:00–11:00am
Fairmont Royal York, Main Mezzanine, Territories (Room 19)

AAAL/ILTA Joint Session - Revisiting the Interfaces between SLA and Language

Free for all LTRC 2015 registrants!
York University’s Graduate Program in Education
Language, Culture and Teaching

Our unique program will allow you to study education within the broad fields of Language, Culture and Teaching. As a student in our program, you will share our commitment to the interdisciplinary study of education through rigorous intellectual inquiry.

Master of Education
Doctor of Philosophy
Specialized Diplomas

Flexibility is key – students are free to develop individual programs of study and research as they work with faculty supervisors.

Small classes allow for individual attention and collaboration with peers.

Guaranteed funding is available for all full-time graduate students.

www.edu.yorku.ca
Collaborating across Languages and Borders

La collaboration au-delà des langues et des frontières

26-28 • MARCH / MARS • 2015
NIAGARA FALLS | ON | CANADA

www.caslt.org/WCML-CMLV-2015
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30am-4:00pm</td>
<td>Conference Registration</td>
<td>Churchill Court</td>
</tr>
<tr>
<td>9:00am-12:00pm</td>
<td><strong>Workshop 1 Part 1</strong>: Working with (and teaching) language test specifications for large-scale testing and classroom assessment</td>
<td>Churchill Ballroom A</td>
</tr>
<tr>
<td></td>
<td>Fred Davidson, Youngshin Chi, Sun Joo Chung, and Stephanie Gaillard</td>
<td></td>
</tr>
<tr>
<td>9:00am-12:00pm</td>
<td><strong>Workshop 2 Part 1</strong>: Using Rasch analysis to gain insights into test development</td>
<td>Churchill Ballroom B</td>
</tr>
<tr>
<td></td>
<td>Rita Green</td>
<td></td>
</tr>
<tr>
<td>12:00-1:00pm</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>1:00-4:00pm</td>
<td><strong>Workshop 1 Part 2</strong>: Working with (and teaching) language test specifications for large-scale testing and classroom assessment</td>
<td>Churchill Ballroom A</td>
</tr>
<tr>
<td></td>
<td>Fred Davidson, Youngshin Chi, Sun Joo Chung, and Stephanie Gaillard</td>
<td></td>
</tr>
<tr>
<td>1:00-4:00pm</td>
<td><strong>Workshop 2 Part 2</strong>: Using Rasch analysis to gain insights into test development</td>
<td>Churchill Ballroom B</td>
</tr>
<tr>
<td></td>
<td>Rita Green</td>
<td></td>
</tr>
<tr>
<td>1:00-4:00pm</td>
<td><strong>Workshop 3</strong>: Learning oriented language assessment for the classroom: A primer</td>
<td>Rossetti</td>
</tr>
<tr>
<td></td>
<td>Anthony Green &amp; Liz Hamp-Lyons</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td>Location</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>8:30am-6:30pm</td>
<td>Conference Registration</td>
<td>Churchill Court</td>
</tr>
<tr>
<td>9:00am-12:00pm</td>
<td><strong>Workshop 1 Part 3:</strong> Working with (and teaching) language test specifications for large-scale testing and classroom assessment</td>
<td>Churchill Ballroom A</td>
</tr>
<tr>
<td></td>
<td>Fred Davidson, Youngshin Chi, Sun Joo Chung, and Stephanie Gaillard</td>
<td></td>
</tr>
<tr>
<td>9:00am-12:00pm</td>
<td><strong>Workshop 3:</strong> Learning oriented language assessment for the classroom: A primer</td>
<td>Rossetti</td>
</tr>
<tr>
<td></td>
<td>Anthony Green &amp; Liz Hamp-Lyons</td>
<td></td>
</tr>
<tr>
<td>9:00am-12:00pm</td>
<td><strong>Workshop 4 Part 1:</strong> Mixed methods research in language assessment: Evolution and practice</td>
<td>Churchill Ballroom B</td>
</tr>
<tr>
<td></td>
<td>Carolyn E. Turner &amp; Janna Fox</td>
<td></td>
</tr>
<tr>
<td>12:00-1:00pm</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>12:00-5:00pm</td>
<td>ILTA Executive Board Session</td>
<td>Windsor</td>
</tr>
<tr>
<td>1:00-4:00pm</td>
<td><strong>Workshop 1 Part 4:</strong> Working with (and teaching) language test specifications for large-scale testing and classroom assessment</td>
<td>Churchill Ballroom A</td>
</tr>
<tr>
<td></td>
<td>Fred Davidson, Youngshin Chi, Sun Joo Chung, and Stephanie Gaillard</td>
<td></td>
</tr>
<tr>
<td>1:00-4:00pm</td>
<td><strong>Workshop 4 Part 2:</strong> Mixed methods research in language assessment: Evolution and practice</td>
<td>Churchill Ballroom B</td>
</tr>
<tr>
<td></td>
<td>Carolyn E. Turner &amp; Janna Fox</td>
<td></td>
</tr>
<tr>
<td>4:30-6:00pm</td>
<td>LTRC Newcomers’ Session</td>
<td>Rossetti</td>
</tr>
<tr>
<td>6:00–8:00pm</td>
<td>Opening Reception</td>
<td>Churchill Court</td>
</tr>
</tbody>
</table>
**Wed, Mar 18**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30am-5:00pm</td>
<td>Conference Registration</td>
<td>Churchill Court</td>
</tr>
<tr>
<td>8:30am-5:00pm</td>
<td>Publishers’ Exhibits</td>
<td>Churchill Court</td>
</tr>
<tr>
<td>8:45-9:00am</td>
<td>Welcome to LTRC2015</td>
<td>Churchill Ballroom</td>
</tr>
<tr>
<td>9:00-10:00am</td>
<td><strong>Messick Memorial Lecture:</strong> Susan Brookhart</td>
<td>Churchill Ballroom</td>
</tr>
<tr>
<td>10:00-10:20am</td>
<td>Break</td>
<td>Churchill Court</td>
</tr>
<tr>
<td>10:20-11:45am</td>
<td><strong>Paper Session 1</strong> (4 Parallel)</td>
<td></td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Gu, Yan, &amp; Tian: Impact of an international English proficiency test on young learners: Voices from learners and parents</td>
<td></td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Gu &amp; Hsieh: Distinguishing features in scoring young language students’ oral performances</td>
<td>Churchill Ballroom A</td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Sawaki &amp; Koizumi: Japanese students’ and teachers’ perception and use of score reports for two large-scale EFL tests</td>
<td></td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Tsagari &amp; Vogt: Assessment literacy of foreign language teachers: Research, challenges and future prospects</td>
<td></td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Chapman: Task equivalence in writing assessment: Which task features can and should be controlled?</td>
<td>Churchill Ballroom B</td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Salamoura, Guetterman, Khalifa &amp; Docherty: Enhancing test validation through rigorous mixed methods design features</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Llosa, Malone, Wei, Donovan, &amp; Stevens: Comparability of writing tasks in TOEFL iBT and university writing courses: Insights from students and instructors</td>
<td></td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Nakastuhara, Inoue, Berry &amp; Galaczi: A multifaceted study using web-based video conferencing technology in the assessment of spoken language</td>
<td></td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Wagner, Chen, Park, Stone, &amp; Jang: Evaluating and revising a reading test blueprint through multiple methods: The case of the CELPIP-G</td>
<td></td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Elwood &amp; Petchko: Development and validation of a placement test for a multi-polar graduate institute</td>
<td></td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Brau &amp; Brooks: Testing the right skill: The misapplication of reading scores as a predictor of translation ability</td>
<td></td>
</tr>
<tr>
<td>11:45-1:30pm</td>
<td>(Lunch) <strong>Group Photo</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>LAQ Editorial Board Meeting</strong></td>
<td></td>
</tr>
<tr>
<td>1:30-3:00pm</td>
<td><strong>Poster Session</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Session Chair: Yong-Won Lee</td>
<td></td>
</tr>
<tr>
<td>3:00-4:25pm</td>
<td><strong>Paper Session 2</strong> (4 Parallel)</td>
<td></td>
</tr>
<tr>
<td>3:00-3:25pm</td>
<td>Iwashita &amp; Vasquez: Features of discourse competence in IELTS speaking task performances at different band levels</td>
<td></td>
</tr>
<tr>
<td>3:30-3:55pm</td>
<td>Youn: The interlocutor effect in paired speaking tasks to assess interactional competence</td>
<td></td>
</tr>
<tr>
<td>4:00-4:25 pm</td>
<td>Acedo Bravo, Diaz Moreno &amp; Baten: Challenges of blended learning assessments in network-limited areas</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>3:00-3:25pm</td>
<td>Phakiti: Structural equation models of calibration, performance appraisals, and strategy use in an IELTS listening test</td>
<td></td>
</tr>
<tr>
<td>3:30-3:55pm</td>
<td>Cai: Producing actionable feedback in EFL diagnostic assessment</td>
<td></td>
</tr>
<tr>
<td>4:00-4:25pm</td>
<td>Suvorov: Test-takers’ use of visual information from context and content videos in the video-based academic listening test</td>
<td></td>
</tr>
<tr>
<td>3:00-3:25pm</td>
<td>Qian: Applying academic vocabulary lists to validating test contents</td>
<td></td>
</tr>
<tr>
<td>3:30-3:55pm</td>
<td>Harsch &amp; Hartig: C-tests outperform Yes/No vocabulary size tests as predictors of receptive language skills</td>
<td></td>
</tr>
<tr>
<td>4:00-4:25pm</td>
<td>Kramer &amp; McLean: Comparing aural and written receptive vocabulary knowledge of the first 5k and the AWL</td>
<td></td>
</tr>
<tr>
<td>3:00-3:25pm</td>
<td>Dunlop: Eye tracking of learner interactions with diagnostic feedback on French as a Second Language skill reports</td>
<td></td>
</tr>
<tr>
<td>3:30-3:55pm</td>
<td>Ballard &amp; Lee: How young children respond to computerized reading and speaking test tasks</td>
<td></td>
</tr>
<tr>
<td>4:00-4:25pm</td>
<td>Fleckenstein, Leucht &amp; Koller: B2 or C1? Investigating the equivalence of CEFR-based proficiency level classifications</td>
<td></td>
</tr>
<tr>
<td>4:25-4:45pm</td>
<td>New Perspectives in Language Assessment Series Launch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>4:45-6:15pm</td>
<td>Symposium Session 1 (Plenary)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Symposium 1: Testing as social practice: Canada and context in construct representation, test interpretation, and arguments for consequential validity.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fox, Laurier, Baker, Mellow, Begg, Slomp, Zumbo, Maddox, &amp; Turner</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td>Location</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------------------------------------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>6:30-7:30pm</td>
<td><strong>Professional Event 1</strong>: Publishing in Language Assessment</td>
<td>Mountbatten Salon</td>
</tr>
<tr>
<td></td>
<td>Special Sponsor: Higher Education Press</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chair: Guoxing Yu</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Guoxing Yu, Executive Editor, Assessment in Education (Taylor &amp; Francis)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Liz Hamp-Lyons, Editor, Assessing Writing (Elsevier)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>James Purpura, Editor, Language Assessment Quarterly (Taylor &amp; Francis)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>April Ginther, Co-Editor, Language Testing (Sage)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fred Davidson, Distinguished Editorial Adviser, Language Testing in Asia (Springer)</td>
<td></td>
</tr>
<tr>
<td>8:30am-5pm</td>
<td>Conference Registration</td>
<td>Churchill Court</td>
</tr>
<tr>
<td>8:30am-5pm</td>
<td>Publishers’ Exhibits</td>
<td>Churchill Court</td>
</tr>
<tr>
<td>8:45-9:00am</td>
<td>Announcements</td>
<td>Churchill Ballroom</td>
</tr>
<tr>
<td>9:00-10:00am</td>
<td><strong>The Davies Lecture: Alister Cumming</strong></td>
<td>Churchill Ballroom</td>
</tr>
<tr>
<td>10:00-10:20am</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>10:20-11:45am</td>
<td><strong>Symposium Session 2 &amp; Paper Session 3</strong> (4 Parallel)</td>
<td></td>
</tr>
<tr>
<td>10:20-11:45am</td>
<td><strong>Symposium 2</strong>: Assessment for learning (AFL): Teaching, learning and assessment in the writing classroom.</td>
<td>Mountbatten A</td>
</tr>
<tr>
<td></td>
<td>Zhang, Sun, Shen, Huang, Gao &amp; Wu</td>
<td></td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Colby-Kelly: Assessing the unobservable: Can AFL inform on nascent learning in an L2 classroom setting?</td>
<td></td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Saito: Junior and senior high school EFL teachers’ practice of formative assessment: A mixed method study</td>
<td>Churchill Ballroom A</td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Winke, Lee, Walter, Weyant, Thirakunkovit, &amp; Yan: What does the cloze test really test? A replication with eye-tracking data</td>
<td></td>
</tr>
</tbody>
</table>

**Thurs, Mar 19**
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:20-10:45am</td>
<td>Volkov, Stone &amp; Gesicki: Empirically derived rating rubric in a large-scale testing context</td>
<td></td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Wu &amp; Wu: Constructing a common scale for a multi-level test to enhance interpretation of learning outcomes</td>
<td>Churchill Ballroom B</td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Margolis: Factors contributing to fluency ratings in classroom-based assessment</td>
<td></td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Ma: Chinese students’ test preparation for success on high-stakes English language tests</td>
<td></td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Bai: A students’ perspective in a high-stakes test: Attitudes, contributing factors and test performance</td>
<td>Mountbatten B</td>
</tr>
<tr>
<td>11:45-1:30pm</td>
<td>(Lunch) <strong>ILTA Membership Meeting</strong></td>
<td>Churchill Ballroom</td>
</tr>
<tr>
<td>1:30-3:00pm</td>
<td><strong>Work in Progress</strong></td>
<td>Mountbatten Salon</td>
</tr>
<tr>
<td></td>
<td>Session Chair: Lorena Llosa</td>
<td></td>
</tr>
<tr>
<td>3:00-4:25pm</td>
<td><strong>Papers Session 4</strong> (3 Parallel)</td>
<td></td>
</tr>
<tr>
<td>3:00-3:25pm</td>
<td>Knoch, Macqueen, May, Pill, &amp; Storch: Test use in the transition from university to the workplace: Stakeholder perceptions of academic and professional writing demands</td>
<td>Churchill Ballroom A</td>
</tr>
<tr>
<td>3:30-3:55pm</td>
<td>Kim &amp; Billington: Tackling the issue of L1 influenced pronunciation in English as a lingua franca communication contexts: The case of aviation</td>
<td></td>
</tr>
<tr>
<td>4:00-4:25pm</td>
<td>van Naerssen: Promoting credible language assessment in courts: Two forensic cases</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td>Location</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>3:00-3:25pm</td>
<td>Papageorgiou &amp; Ockey: Does accent strength affect performance on a listening comprehension test for interactive lectures?</td>
<td></td>
</tr>
<tr>
<td>3:30-3:55pm</td>
<td>Min &amp; He: Examining the effect of DIF anchor items on equating invariance in computer-based EFL listening assessment</td>
<td>Churchill Ballroom B</td>
</tr>
<tr>
<td>4:00-4:25pm</td>
<td>Saif, Cheng, &amp; Rahimi: High-stakes test preparation programs and learning outcomes: A context-specific study of learners' performance on IELTS</td>
<td></td>
</tr>
<tr>
<td>3:00-3:25pm</td>
<td>Yang &amp; Wu: An investigation into cognitive evidence for validating a visual literacy task in listening assessment</td>
<td></td>
</tr>
<tr>
<td>3:30-3:55pm</td>
<td>Lidster &amp; Shin: Developing and validating achievement-based assessments of student learning outcomes in an intensive English program</td>
<td>Rossetti</td>
</tr>
<tr>
<td>4:00-4:25pm</td>
<td>Park, Dunlop, van der Boom, &amp; Jang: A mixed-methods investigation into the young learners' cognitive and metacognitive strategy use in reading test</td>
<td></td>
</tr>
<tr>
<td>4:25-4:45pm</td>
<td>Break</td>
<td>Churchill Court</td>
</tr>
<tr>
<td>4:45-6:15pm</td>
<td><strong>Symposium Session 3</strong> (2 Parallel)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Symposium 3</strong>: Roles and needs of learning-oriented language assessment</td>
<td>Churchill Ballroom A</td>
</tr>
<tr>
<td></td>
<td>Purpura, Waring, Hamp-Lyons, Green &amp; Hill</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Symposium 4</strong>: New models and technologies for blurring the distinction between language testing and language learning</td>
<td>Churchill Ballroom B</td>
</tr>
<tr>
<td></td>
<td>Stenner, O'Sullivan, Dunlea, Sandvik, &amp; Geva</td>
<td></td>
</tr>
</tbody>
</table>
### Fri, Mar 20

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30am-5:00pm</td>
<td>Conference Registration</td>
</tr>
<tr>
<td>8:30am-5:00pm</td>
<td>Publishers’ Exhibit</td>
</tr>
<tr>
<td>8:45-9:00am</td>
<td>Announcements</td>
</tr>
<tr>
<td>9:00-10:00am</td>
<td>Cambridge/ILTA Distinguished Achievement Award: Tim McNamara</td>
</tr>
<tr>
<td>10:00-10:20am</td>
<td>Break</td>
</tr>
<tr>
<td>10:20-11:45am</td>
<td>Papers Session 5 (4 Parallel)</td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Isaacs &amp; Turner: Aligning teaching, learning, and assessment in EAP instruction? Stakeholders’ views, external influences, and researchers’ perspectives</td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Van Gorp: Integrating task-based language teaching and assessment: Towards a unified task specification framework</td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Erickson: Large-scale assessment for learning – a collaborative approach</td>
</tr>
</tbody>
</table>

**Professional Event 2:** Careers in language assessment
Chair: Eunice Jang
Ardeshir Geranpayeh, Cambridge English Language Assessment
Jonathan Schmidgall, Educational Testing Service
May Tan, Canadian Armed Forces
Eunice Jang, OISE/University of Toronto
Jake Stone, Paragon Testing Enterprises (UBC)
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:20-10:45am</td>
<td>Huhta, Alderson, Nieminen, &amp; Ullakonoja: Diagnostic profiling of foreign language readers and writers</td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Jin, Zou, &amp; Zhang: What CEFR level descriptors mean to college English teachers and students in China</td>
</tr>
<tr>
<td></td>
<td><em>Churchill Ballroom B</em></td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Zhao, Jin, &amp; Wang: Detecting evidence behind the college English curriculum requirements in China: A mixed-methods study</td>
</tr>
<tr>
<td></td>
<td><em>Windsor</em></td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Roever, Elder, &amp; Fraser: Validation of a web-based test of ESL sociopragmatics</td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Ryu: Investigating a construct of pragmatic and communicative language ability through email writing tasks</td>
</tr>
<tr>
<td></td>
<td><em>Mountbatten A</em></td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Liu: Automated writing instructional tool for English language learners: A case study of WriteToLearn</td>
</tr>
<tr>
<td></td>
<td><em>Windsor</em></td>
</tr>
<tr>
<td>10:20-10:45am</td>
<td>Deygers: The concurrent and predictive validity of university entrance tests</td>
</tr>
<tr>
<td>10:50-11:15am</td>
<td>Bi: Modeling the relationships between construct-relevant and -irrelevant strategic behaviors and lexi-co-grammar test performance</td>
</tr>
<tr>
<td></td>
<td><em>Mountbatten B</em></td>
</tr>
<tr>
<td>11:20-11:45am</td>
<td>Ikeda: Investigating constructs of L2 pragmatics through L2 learners' oral discourse and interview data</td>
</tr>
<tr>
<td>11:45-1:15pm</td>
<td>(Lunch) <em>LT Editorial Board Meeting</em></td>
</tr>
<tr>
<td></td>
<td><em>Windsor</em></td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>1:15-1:40pm</td>
<td>Papers Session 6 (3 Parallel)</td>
</tr>
<tr>
<td>1:15-1:40pm</td>
<td></td>
</tr>
<tr>
<td>2:15-2:40pm</td>
<td></td>
</tr>
<tr>
<td>1:15-1:40pm</td>
<td></td>
</tr>
<tr>
<td>1:45-2:10pm</td>
<td></td>
</tr>
<tr>
<td>2:15-2:40pm</td>
<td></td>
</tr>
<tr>
<td>1:15-1:40pm</td>
<td></td>
</tr>
<tr>
<td>1:45-2:10pm</td>
<td></td>
</tr>
<tr>
<td>2:15-2:40pm</td>
<td></td>
</tr>
<tr>
<td>2:40-3:00pm</td>
<td></td>
</tr>
</tbody>
</table>
### Conference Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Session/Activity</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:00-4:30pm</td>
<td><strong>Symposium Session 4</strong> (3 Parallel)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Symposium 5</strong>: Vocabulary in assessment: What do we mean and what do we assess?</td>
<td>Mountbatten A</td>
</tr>
<tr>
<td></td>
<td>Schoonen, Schmitt, Kremmel, Jarvis, &amp; Segalowitz</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Symposium 6</strong>: The evaluation of school-based EFL teaching and learning in China</td>
<td>Mountbatten B</td>
</tr>
<tr>
<td></td>
<td>Wu, Qi, Han, Liu &amp; Wei</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Symposium 7</strong>: Applications of automated scoring tools for student feedback and learning</td>
<td>Rossetti</td>
</tr>
<tr>
<td></td>
<td>Weigle, Deane, Geranpayyeh, Suzuki, Davis &amp; Ginther</td>
<td></td>
</tr>
<tr>
<td>4:35-5:35pm</td>
<td><strong>Inaugural ILTA Debate</strong></td>
<td>Churchill Ballroom</td>
</tr>
<tr>
<td></td>
<td>Liz Hamp-Lyons, Tim Mcnamara &amp; Barry O’Sullivan</td>
<td></td>
</tr>
<tr>
<td>5:35-5:45pm</td>
<td>Closing</td>
<td>Churchill Ballroom</td>
</tr>
<tr>
<td>6:30-10:00pm</td>
<td><strong>Banquet and Awards</strong></td>
<td>Hart House</td>
</tr>
</tbody>
</table>

### Ph.D. and M.A. in LINGUISTICS and APPLIED LINGUISTICS

http://www.yorku.ca/gradling/

Located in Toronto, Canada’s largest city, the Graduate Program in Linguistics and Applied Linguistics at York University is well known for the excellence of its faculty, students and teaching. Faculty research and supervision interests cover a broad spectrum of areas in the two offered fields of Linguistics and Applied Linguistics.

We offer a one-year M.A. and a four-year Ph.D. program with courses and specialization in either in Linguistics or Applied Linguistics.

In **Linguistics**, students can focus on sociolinguistics (language variation and change, discourse analysis), formal linguistics (phonology, syntax), or psycholinguistics.

In **Applied Linguistics**, students can focus on second language education and assessment, language policy and planning, narrative studies, as well as multi-modal and e-learning.

**For More Information Contact:**
Rose Frezza-Edgecombe, Program Administrator
Telephone: 416-650-8046 or E-mail: gradling@yorku.ca
Mon, Mar 16 & Tues, Mar 17

Working with (and teaching) language test specifications for large-scale testing and classroom assessment

Fred Davidson, Youngshin Chi, Sun Joo Chung & Stephanie Gaillard

A test specification is a generative document from which many equivalent test tasks or items can be produced. It is a well-established classical tool in test development. “Specs” are relevant at all language testing contexts: from small-scale classroom-based assessment to large high-stakes testing programs. Recent years have seen the emergence of empirical research and statements of theory about the use of specs in language testing, inter alia:

- Chi (2011) featured test specs in her exploration of factors in the assessment of listening
- Chung (2014) reported on the evolution and changes in a major high-stakes writing placement test, for which test specs (and their changes) provided key documentary evidence
- Davidson (2012a) and Davidson (2012b) are recent statements of theoretical considerations in specs in language testing
- Gaillard (2014) utilized specs in a central way during her development of an elicited imitation task for the assessment of French (for both practical and research purposes)
- Li (2006) investigated how specs change over time and, in so doing, explored the metaphor of an “audit” of evolution of test specs
- Kim, J.T. (2006) employed test-takers in the development of his test, and specs were the common vehicle by which all parties communicated
- Kim, J. et al. (2010) reported on a spec management project with a group of trained item writers who were producing operational test items and tasks
- Furthermore, personal experience (Davidson) and anecdotal reports (from other language testers) suggest that test specs are a useful feature in the teaching of language testing and growth of assessment literacy.

The workshop is designed for maximum applicability to many language test settings from classroom testing, to medium-stakes or institutional testing, to national and major high-stakes assessment projects. The goals of this workshop are to:

1. Teach about language test specs
2. Critically analyze and obtain feedback on specs brought by workshop participants (a “spec clinic”)
3. Practice with spec editing and revision
4. Explore and expand spec theory (Davidson 2012b)
5. Explore how specs apply to different settings
6. Provide information and tips on how to teach specs

The two days will be organized along themes as shown here: (a) Overview and basic theory of specs, (b) Exploration of various models of test specs (this includes the bring-your-own Spec Clinic), (c) The role of test specs in evolutionary test change and validation arguments, (d) Use of test specs in development of research instruments, (e) Management of operational test-spec driven projects, (f) Theory, teaching (of specs), and Wrap-up

Workshop activities will follow a general model as follows: presentation of specs (e.g. samples brought by participants and published
samples), then critical discussion of the samples (in small groups or as a whole), then revision of the specs (with justification as to why), capped by reflection about the revision and its potential contribution to test validity (Davidson 2012a; Li, 2006). Depending on the goal and theme of the moment, this general activity model may be tailored to a specific need -- e.g. in the theme of management of operational spec-driven testing projects, we will also explore matters of budgeting and organizational management of item/task production and how revision of a given spec should take those matters into account.

Fred Davidson is a Professor Emeritus of Linguistics at the University of Illinois at Urbana-Champaign. He currently resides in Phoenix, Arizona. His interests include language test specifications and the history and philosophy of testing.

Youngshin Chi is an independent consultant working in the United States. She received her PhD in Educational Psychology from the University of Illinois at Urbana-Champaign. She has research interests in test development, test validation, and second language listening.

Sun Joo Chung is a Lecturer at Hankuk University of Foreign Studies in Seoul, Korea. She received her PhD in Educational Psychology from the University of Illinois at Urbana-Champaign. She has research interests in placement testing, quality management in language assessment programs, and second language writing.

Stephanie Gaillard is an Assistant Professor of French and the language program coordinator at Louisiana State University. She received her PhD in French with a concentration in Second Language Acquisition and Teacher Education at the University of Illinois at Urbana-Champaign. Her research interest includes test validation, placement testing, test specifications and especially their impacts in second language classroom curricula.

WORKSHOP 2 (1 day)
Mon, Mar 16
Using Rasch analysis to gain insights into test development

Rita Green

This one-day workshop has two main aims: firstly, to provide a basic introduction into the insights that Rasch analysis can offer test developers; and secondly, to encourage participants to use such procedures to help them better understand the tests they have developed. Participants will learn how to construct and run a Winsteps control file based on a set of raw data. Item difficulty and person ability estimates, person and item reliabilities, and the amount of error associated with each item and person will be explored. In addition, distracter performance will be investigated. Furthermore, participants will be expected to bring their own laptop to the workshop. Information about the availability of the Winsteps software (Linacre) will be provided in due course. This workshop is for Rasch beginners.

Dr. Rita Green has been involved in the field of language testing for more than 30 years, as a trainer and adviser to many projects around the world. She has run numerous seminars and workshops in Europe, Africa, Asia and South America. She is a Visiting Teaching Fellow in the Department of Linguistics & English Language at Lancaster University where she has been the Director of a four-week course in language testing and statistics for language testers since 2001. She is also an EALTA Expert Member. She is the author of Statistical Analyses for Language Testers (2013, Palgrave Macmillan, ISBN 9781137018281) and has presented her work at LTRC and EALTA conferences.
WORKSHOP 3 (2 half days)
Mon, Mar 16 &
Tues, Mar 17
Learning oriented language assessment for the classroom: A primer

**Anthony Green & Liz Hamp-Lyons**

This workshop will introduce participants to the principles which are emerging for learning-oriented assessment (LOA), focusing in particular on learning-oriented language assessment (LOLA) by teachers in the classroom.

Outline:

1. Principles of LOLA
2. Role of teachers in diagnosing learners’ needs for LOLA-focused instruction
3. LOLA tasks
4. LOLA-oriented feedback
5. Helping students to self and peer assess

The purpose is to introduce the concept of learning oriented language assessment as a means of improving the coherence of educational assessment systems that involve both assessment by teachers in the classroom and large scale external testing. The workshop will a) explore how teachers can use self-designed assessments to provide feedback to students and diagnostic evidence to guide instructional practices and b) consider how external tests – often regarded as a threat to preferred classroom practices – might be more effectively integrated into local curricula.

This workshop begins with an overview of the formative functions of learning oriented language assessment, presenting sample performance assessment activities that illustrate key concepts such as task authenticity, cognitive engagement and uptake of feedback. Tasks from large-scale language tests will then be evaluated as tools for LOLA and participants will explore the apparent tensions between test preparation, test administration and learning processes. Participants will explore the differences between formal testing and classroom assessment, consider how formal tests might be reformed to become valuable learning experiences and how tests might be better integrated into the teaching and learning cycle.

**Goal:** To encourage debate about the roles of assessment and testing in educational programs.

**By the end of the workshop, participants will:**

- Understand the key features of learning oriented language assessment
- Understand the role of cognitive engagement in assessment by learners and teachers
- Be equipped to evaluate language test tasks as tools for effective learning
- Build their awareness of how external tests might be more effectively integrated with teaching and learning

**Anthony Green** is Deputy Director of the Centre for English Language Learning and Assessment at the University of Bedfordshire (UK). His publications include the books Exploring Language Assessment and Testing, (Routledge), Language Functions Revisited and IELTS Washback in Context (both Cambridge University Press). He has extensive experience in training teachers in language assessment, including acting as the academic coordinator of the EU funded ProSET project carried out in partnership with the Ministry of Education of the Russian Federation and providing training in assessment for secondary school teachers nationwide. He has carried out assessment training workshops in over 30 countries.

**Liz Hamp-Lyons** is Visiting Professor at CRELLA (English Language Assessment) and at Shanghai Jiao Tong University, China (International Studies). Her research encompasses development and validation of English language performance (i.e. writing and speaking) assessments, classroom assessment and assessment for learning as well as EAP materials and methods. She is the Editor of Assessing Writing and joint Editor of the Journal of English for Academic Purposes.
Situating mixed methods research in language assessment: Evolution and practice

Carolyn E. Turner & Janna Fox

Mixed methods research (MMR) design, which draws on both quantitative and qualitative approaches in addressing a research problem, is increasingly evident in studies within language testing and assessment (hereafter, LT). This is in keeping with the general trend toward MMR, which is evident in: 1) the increasing number of prominent publications describing MMR procedures, pitfalls, and potentials (e.g., Creswell & Plano Clark, 2011; Tashakkori & Teddlie, 2010); 2) the successful launch of the Journal of Mixed Methods Research in 2007, which now ranks 9 out of 92 Social Science/Interdisciplinary journals; 3) the number of doctoral dissertations which apply MMR approaches in addressing research problems and questions (Cheng & Fox, 2013); and 4) revisions taking place in textbooks used in social science research (e.g., McMillan & Schumacher, 2010), which in more recent editions place MMR design alongside quantitative and qualitative as a separate, but equally important approach.

Arguably, the original challenges to MMR (see the paradigm wars, e.g., Caracelli& Greene, 1997) have ceased to be the focus of debate, as MMR gained legitimacy as a pragmatic (Tashakkori& Teddlie, 2010) alternative to purely quantitative or qualitative research approaches for all but the most rigidly conservative researchers. Indeed, some have suggested (e.g., Creswell, 2012) that MMR is a natural and inevitable extension of a priori triangulating strategies in research, which seek alternative perspectives or source of evidence to cross-validate findings. Within LT, the overriding concerns for validity and validation in assessment practices over the past decades

(Cronbach, 1988; Kunnan, 2000; Messick, 1989), and the concomitant need for convergent sources of evidence in arguments for test interpretation, mesh well with a MMR perspective. Increasing recognition of the role of context (see, Fox, 2003; McNamara & Roever, 2007), the greater presence of qualitative approaches in LT research, and the need to address ever more complex research questions, have led to greater use of and interest in MMR approaches in LT. As Turner (2013) points out, researchers in LT are latecomers to considerations of MMR issues, but their “rationale for the use of MMR follows closely the philosophical orientation most often associated with it, that is, pragmatism” (p. 1403).

Carolyn E. Turner, PhD, is Associate Professor of Second Language Education at McGill University where she teaches assessment and research methods. Her research interests include language testing/assessment in educational settings and in healthcare contexts concerning access for linguistic minorities. A former President of the International Language Testing Association (ILTA) and Associate Editor of Language Assessment Quarterly, her work appears in journals such as Language Testing, TESOL Quarterly, and Health Communication. She is currently co-authoring a book, “Learning-oriented L2 assessment,” with James Purpura. Recently she published chapters on mixed methods research (MMR) in the Companion to Language Assessment and the Encyclopedia of Applied Linguistics.

Janna Fox, PhD, is Associate Professor and Director of the Language Assessment and Testing Research Unit in the School of Linguistics and Language Studies at Carleton University where she teaches courses on research methods to graduate students in Applied Linguistics. Her research in language policy, curriculum, and assessment has increasingly drawn on mixed methods approaches to research, when neither a solely quantitative nor qualitative approach was adequate in addressing a research problem. She is currently co-editing a book (with Vahid Aryadoust) entitled Current Trends in Language Testing in the Pacific Rim and Middle East for Cambridge Scholars Press.
Impact of an international English proficiency test on young learners—voices from learners and parents

Xiangdong Gu, Chongqing University, China/ Cambridge English, UK, xiangdonggu@263.net
Qiaozhen Yan, Chongqing University, qiaozhencqu@163.com
Jie Tian, Chongqing University, China, tianjiemolly@qq.com

As learners are the most direct and ultimate stakeholders of any assessment, an increasing number of washback/impact studies have been focused on learners (e.g. Xie 2008; Zhan & Andrews 2014). Noticeably, young learners are often overlooked. However, with the process of globalization, English has become a world language and the starting age to learn English has been getting increasingly younger worldwide (Graddol 2014). Correspondingly, young learners are encouraged to take various English proficiency tests. Thus it is urgent to study the impact of tests on young learners. This study is one of the few studies investigating the impact of Cambridge English: Key for Schools (CEFR Level A2) on young learners in China. To be specific, the study investigates the impact of the exam on young learners’ motivations, self-awareness of their strengths and weaknesses in their English learning, their English proficiency, test pressure and workload. A questionnaire survey designed for the above-mentioned purposes was conducted among 340 young learners in Beijing China in 2012. As the learners’ views are among the most difficult to make sense of and to use, Green (1997) points out that it may be necessary to call on surrogate stakeholders, for example, parents, to speak on behalf of their children. Indeed, there is emerging evidence that parents play a key role in their children’s school attainment (Feignstein et al 2010). Previous studies (e.g. European Commission 2012) also highlight the necessity to investigate parents’ perspectives of their children’s English language learning. Therefore, in our study, semi-structured interviews were also conducted with twenty parents individually to get deeper information about KET for Schools impact on young learners. The findings of student survey revealed that KET for schools have exerted a strongly positive impact on young learners’ learning motivation, self-awareness of learning strengths and weakness and the improvement of their English proficiency. Remarkably, the findings of parent interview were highly consistent with those of the student survey, particularly in terms of promoting young learners’ English proficiency. Parent interview further revealed the main factors contributing to this positive impact, including the comprehensive measurement of KET for schools, the step-by-step learning approach it offered, the relatively higher test difficulty of the exam compared with the tests at school, young learners’ self-esteem, and the trained teachers’ teaching methods. Nevertheless, both student survey and parent interview demonstrated noteworthy negative impact of the exam on young learners in terms of high test pressure and heavy workload due to the test preparation. Further reasons were explored through parent interview. The study provides valuable insights into the attitudes and perceptions of two influential groups of stakeholders—young learners and their parents who are of particular relevance to the exam developers, researchers, and test users. It is hoped that this study merits further investigation into test impact on young learners nationwide in the Chinese context and beyond. More importantly, it is intended to provoke reflective thoughts about how to positively and effectively...
promote young learners’ English language learning with the help of proficiency test.

Assessment literacy of foreign language teachers: Research, challenges and future prospects

Dina Tsagari, University of Cyprus, dinatsa@ucy.ac.cy

Karin Vogt, University of Heidelberg, vogt@phil-heidelberg.de

Numerous references to ‘new’ literacies have been added to the discourse of various academic and public domains in recent years. Among them the term “language assessment literacy” (LAL) used in the field of language testing and assessment to refer to the skills, knowledge, methods and techniques needed by various stakeholders to carry out effective language assessments (e.g. Hasselgreen, et al. 2004; Inbar-Lourie, 2008; 2013; Fulcher, 2012; Taylor, 2009; 2013). In this paper we firstly define and discuss the notion of LAL and its related issues with a particular focus on language teachers and their professional development in this area. In the second part of the paper we consider current LAL levels of pre- and in-service foreign language teachers (FLTs) based on research findings conducted in three European countries. Both quantitative and qualitative data was collected in a mixed method approach via questionnaires (n=589) distributed to FLTs in Cyprus, Greece and Germany and interviews (n=63) with teachers from the same local contexts. The data was analysed using descriptive and inferential statistics (frequency distributions, correlation analyses, regression analysis) and qualitative content analysis (following Mayring 2010). Despite the small differences across countries, the results show that only certain elements of teachers’ LAL expertise are developed such as testing microlinguistic aspects (grammar/vocabulary) and language skills (e.g. reading and writing). LAL aspects such as compiling and designing non-traditional assessment methods (e.g. self-/peer-assessment or portfolio assessment), grading and placing students onto courses as well as establishing quality criteria of assessments (e.g. validity, reliability or using statistics) are not. To compensate for insufficient training, teachers reported that they learn about LAL on the job or use teaching materials for their assessment purposes. Teachers overall expressed a need to receive training across the range of LAL features identified in the study with varying priorities depending on their local educational contexts. The results of the study offer insights into teacher cognition of their LAL needs and outline pressing challenges which exist for fostering LAL among FLTs. Finally the presentation proposes various approaches to the professional development of FLTs and directions for future research in LAL.

Comparability of writing tasks in TOEFL iBT and university writing courses: Insights from students and instructors

Lorena Llosa, New York University, lorena.llosa@gmail.com

Margaret Malone, Center for Applied Linguistics, mmalone@cal.org

Jing Wei, New York University, jing.wei@nyu.edu

Anne Donovan, Center for Applied Linguistics, adonovan@cal.org

Lilly Stevens, New York University, lma350@nyu.edu

Although many studies have been conducted to validate the writing tasks on various versions of the TOEFL, none have explored the comparability of students’ performance on TOEFL writing tasks and actual academic writing tasks. Investigating this comparability is essential for providing backing for the extrapolation inference in the TOEFL validity argument. We conducted a study that examines the comparability of TOEFL iBT writing tasks and tasks assigned in required university writing courses.
Because such courses constitute a graduation requirement for students in many North American universities, they represent an important part of the TLU domain. Specifically, we investigated whether students viewed their performance on the TOEFL iBT writing tasks as comparable to their writing performance in their required writing class. We also examined instructors’ and students’ perceptions of the comparability of the characteristics of the TOEFL writing tasks and rubrics and those assigned in required writing classes. One hundred international, nonnative-English-speaking, undergraduate students enrolled in required university writing courses from eight universities in the U.S. and their writing instructors (N=15) participated in the study. During a two-hour data collection session, students completed a background questionnaire, two TOEFL iBT tasks, one Integrated and one Independent, and a post questionnaire about their perceptions of the TOEFL writing tasks, the level of correspondence between the TOEFL writing tasks and the tasks they encounter in required university writing courses, and their views about the extent to which the TOEFL writing tasks measure their writing abilities accurately. Instructors also completed a questionnaire that included screenshots of the TOEFL writing tasks used in the study and the TOEFL writing rubrics; instructors were asked to comment on the degree of comparability between the characteristics of the TOEFL tasks and rubrics and those of the writing tasks they assign in their courses. Preliminary analyses of the student questionnaire data indicate that the majority of students believed that their performance on the TOEFL tasks was similar to or worse than their performance on tasks in their writing class. Also, 60% of students reported that the Independent task most resembled the types of tasks they encountered in their writing class. By contrast, only 17% of students thought the Integrated task was most comparable. In general, writing instructors reported that the TOEFL tasks only partially resembled the types of writing they assign, with the Independent task perceived as more similar by many of the instructors—an interesting finding considering that integrated tasks are typically believed to be more similar to the types of tasks that students will likely encounter in academic settings. Even though instructors believed that the TOEFL tasks only partially resembled the tasks they assigned, most of them reported that the criteria in the TOEFL rubrics are similar to the criteria they use to assess student writing. Detailed findings of quantitative and qualitative data from the student and teacher questionnaires will be presented and implications of the findings for the validity of score interpretations and the nature of the TLU domain (writing courses vs. content courses) will be discussed.

Evaluating and revising a reading tesblueprint through multiple methods: The case of the CELPIP-G

Maryam Wagner, OISE/University of Toronto, maryam.wagner@utoronto.ca

Michelle Yue Chen, The University of British Columbia / Paragon Testing Enterprises, mchen@paragontesting.ca

Gina Park, OISE/University of Toronto, gina.park@mail.utoronto.ca

Jake Stone, Paragon Testing Enterprises, stone@paragontesting.ca

Eunice Eunhee Jang, OISE/University of Toronto, eun.jang@utoronto.ca

In Canada, prospective immigrants need to demonstrate their language proficiency through one of two approved tests, one of which is the Canadian English Language Proficiency Index Program General (CELPIP-G) (Canadian and Immigration Canada). This high-stakes test has significant consequences for its test-takers; accordingly, the quality of the CELPIP-G is of critical importance (Shohamy, 2007). As part of an on-going development and validation process to evaluate the quality of inferences made from the test, we engaged in a systematic review and analysis of the CELPIP-G. This paper focuses on
distinguishing ideas (e.g., relevant and irrelevant explicit information; inferencing; knowledge of knowledge; textually explicit information; non-iterative process revealed ten reading processing less interpretable construct representation. The investigation involved three main tasks: 1) Developing an overarching reading assessment framework based on a review of relevant theoretical and empirical literature (e.g., Alderson, 2000; Grabe, 1991); 2) Expert analysis of the CELPIP-G reading test items using the framework to identify the knowledge and cognitive skills underlying each item; and 3) Developing a cognitive diagnostic model (CDM) using test performance data to empirically substantiate the proposed knowledge/skill relationships with the items. Based on the literature review, 17 fine-grained sub-skills were identified as underlying assessment of functional reading proficiency which served as the foundation for the expert panel's work of reviewing test items on two operational CELPIP-G reading test forms. The purpose of the panel's work was not to confirm the framework's set of linguistic knowledge and cognitive skills, but to use it as a basis to identify a wide range of knowledge and skills enriching their theoretical meanings through collaboration and negotiations. The skills identified by the expert panel were specified into a Q-matrix (Tatsuoka, 1990). A Q-matrix maps the relationships among knowledge and skills with items. These relationships were empirically evaluated through cognitive diagnosis modelling using the Fusion Model skill diagnosis procedure (Roussos et al, 2007). All discrepancies emerging from the modelling were carefully considered to ensure that changes made to the Q-matrix based on psychometrics did not lead to theoretically less interpretable construct representation. The iterative process revealed ten reading processing skills: vocabulary knowledge; grammatical knowledge; textually explicit information; non-explicit information; inferencing; knowledge of cohesive markers and organizational features; distinguishing ideas (e.g., relevant and irrelevant information); knowledge of pragmatics (e.g., tone, mood, formality); interpreting information across different text forms; and synthesising. The psychometric analysis also provided integral information about the quality of test items in terms of difficulty and discrimination, and further revealed the relationship amongst skills, as well as the distribution of test-takers across skill mastery classes. This paper discusses key lessons learned from our efforts to strike a balance between the two different methods and sources of evidence crossing paradigmatic borders: human judgement and psychometrics.

10:50-11:15am

Distinguishing features in scoring young language students’ oral performances

Lin Gu, Educational Testing Service, lgu001@ets.org
Ching-Ni Hsieh, Educational Testing Service, chsieh@ets.org

Research as examined the components of speaking proficiency and their associations with human ratings among adult learners of English (e.g., Iwashita, Brown, MaNamara, & O’Hagan, 2008; Plough, Briggs, & Van Bonn, 2010), and has contributed significantly to our understanding of the construct of speaking ability. However, little attempt has been made to explore the nature of speaking proficiency among young learners whose communication needs are distinctively different from adult learners. This study examined the spoken features of young English language learners using performance data extracted from operational TOEFL Junior Comprehensive (TJC) speaking test database. Our goal was to gain a detailed understanding of how different linguistic features distinguish levels of oral proficiency among young learners. Performance data samples were produced by 57 TJC test takers, between 9 and 12 years of age with varying degrees of oral proficiency.
Four distinct TJC task types were included in the samples. Oral responses were transcribed by human transcribers. Audio recordings and transcripts were then analyzed by using automated scoring technologies in terms of pronunciation, fluency, and vocabulary use. ANOVA analyses were performed for each feature separately to determine the degree to which each linguistic feature differed across proficiency levels. Significant differences across levels were found for all three aspects of the speaking ability. The results also indicated that features related to fluency have the strongest association with ratings of young learners' oral proficiency. This study provides an important piece of empirical evidence in support of the interpretation about young learners' language abilities made on the basis of the scores awarded by TJC human raters. Implications for the development of young learner language assessments and suggestions for future research will be discussed.

Task equivalence in writing assessment: Which task features can and should be controlled?

Mark Chapman, CRELLA, chmark@umich.edu

One difficulty facing those who attempt to assess spoken or written language on repeated occasions is how to ensure that the tasks learners are presented with are equivalent. If certain test tasks elicit responses that are significantly different from other tasks, then test takers' scores may vary depending on which prompt they respond to. These are serious concerns for high-stakes, standardized assessments where parallel test forms are essential. Although there is a developing understanding of how varying complexity of input in integrated writing tests affects the final written product (Cumming, Kantor, Baba, Erdosy, Keanre, & James, 2005; O'Loughlin & Wigglesworth, 2007), the specific features of the task that may affect the final written product are not well understood. This paper describes a study of task equivalence within a high-stakes test of English as a second language. The study focuses on the writing component of the exam and investigates which features of the writing task may be controlled so as to better establish task equivalence. This paper will report on a study of six retired writing prompts from the Michigan English Language Assessment Battery (MELAB), an advanced level test of English for nonnative speakers. The prompts were selected and categorized by task features after applying previously developed prompt taxonomies (Hirokawa and Swales, 1986; Horowitz, 1991; Swales, 1982; Lim, 2010), and interviewing test takers and experienced MELAB raters. The resulting task features that best categorize the writing prompts are domain, response mode, number of rhetorical cues and focus (the extent to which the writer is constrained when selecting a topic to write on). 20 essays at three distinct proficiency levels were collected in response to each of the six prompts (n=360) and each essay was analyzed with respect to a range of discourse variables, selected from the existing literature on task effect and to represent key features from the MELAB rating scale. The discourse variables were selected to operationalize the traits of fluency (essay length), syntactic complexity (average sentence length and average number of words before the first verb), lexical sophistication (average word length, lexical frequency profile, and type-token ratio), cohesion (incidence of all connectives, causative connectives, and temporal connectives), and accuracy (total number of errors and average number of errors per 100 words). The discourse variable data were analyzed using MANOVA for significant differences across prompts. The results revealed significant differences in lexical sophistication, syntactic complexity, and cohesion. The task features that were associated with these significant differences in written product were response mode, prompt focus, and domain. A series of interviews with test takers who had completed the MELAB Writing test confirmed that domain and prompt focus were characteristics of the task that should be controlled. The interview data also indicated that the number of rhetorical cues within the task was an additional task variable that influenced how test takers responded to the prompts. The talk
will conclude with an account of the revisions that have been made to the test specifications to promote task equivalence for all writing prompts.

Impact and consequences of university-based assessment in China: A case study

Jinsong Fan, Fudan University, jinsongfan@fudan.edu.cn

Xiaomei Song, Georgia Southern University, xsong@georgiasouthern.edu

In many parts of the world today, there is a growing trend of combining assessment of learning with assessment for and as learning to promote students’ learning (Cheng, Andrews, & Yu, 2011). As part of this trend, some universities in China have developed local English language tests in recent years, and used them in combination with the College English Test (CET), the national English test, to assess students’ English achievement and proficiency levels (e.g., TOPE Project Team, 2013; Tsinghua University Testing Team, 2012). This study was conducted within this context in which localized assessments are used in parallel with centralized large-scale testing in China. Adopting the mixed-methods sequential explanatory design (Creswell, 2003), this study investigated the impact and consequences of university-based assessment from the perspective of students, a group of stakeholders who are most significantly influenced by such an educational change. The target test of this study is a typical university-based English test developed by a major university in 2011 in response to the call for using localized assessments to enhance students’ English learning. Specifically, this study investigated students’ perceptions of this test, and its impact on students’ learning practices. On top of that, this study examined the role of students’ gender and English ability levels in shaping the test impact. Research data were collected through structured questionnaires (n=335) and semi-structured individual interviews (n=20). A number of statistical analyses were performed of the questionnaire data, including exploratory factor analysis, descriptive statistics, two-way ANOVA, and two-way MANOVA. The qualitative interview data were analyzed through an inductive approach with the aid of NVivo. Results revealed that: 1) despite students’ overall positive perceptions of the test, most of them merely considered it as another external English test imposed on them; 2) the test had more impact on students’ test preparation strategies than on learning content, attitudes, and motivation; 3) students commented negatively on the transparency of test information and the lack of test preparation materials and learning support; 4) the effect of gender was statistically insignificant in shaping students’ perceptions (F(2, 239)=0.19, p>0.05), but the effect of English ability was significant (F(2, 239)=11.18, p<0.05); and 5) gender and English ability had insignificant effect on students’ reported washback on the different dimensions of their learning practices (gender: Pillai’s V=0.00, F(3, 236)=0.15, p>0.05; ability: Pillai’s V=0.05, F(4, 474)=1.88, p>0.05). In conclusion, this study suggests that the test plays a limited role in enhancing students’ learning, and it does not achieve the intended purpose of assessment for and as learning. The failure of the intended washback, as this study reveals, is not attributable to test design but largely because learning and assessment were treated as two independent and unrelated activities in the development and implementation of this university-based test. The paper goes on to consider the broader implications for the conceptualization and implementation of university-based assessment in China, and argues for a change in test developers’ mindset and the closer link between language assessment and learning in the university setting.
Development and validation of a placement test for a multi-polar graduate institute

James Elwood, Meiji University, elwood@meiji.ac.jp

Katerina Petchko, Graduate Institute of Policy Studies, kpetchko@grips.ac.jp

In this session, we will present results of a project to develop a test of academic writing ability for a multi-national graduate school of policy studies. At the institute in question, the student body comprises an eclectic group, with each cohort including some 200 students from, on average, 60 countries in which English is spoken either as a second language or as a foreign language. Applicants must provide evidence of English proficiency, but that evidence is often spotty or of uncertain provenance and seldom includes evidence of writing ability. Unfortunately, in the course of their study at this institute, many students prove to be, at best, inadequately familiar with the conventions of academic writing, a situation which seriously compromises their prospects of completing their studies and producing an acceptable graduation paper. Given this background, we have constructed an academic writing test to measure the skill set deemed necessary to succeed in this community. The initial section of this presentation focuses on the development and validation of the test to date. In the conceptualization phase, we considered standardized tests of academic writing to assess students’ ability, which often employ “general-purpose” writing tasks that elicit a description of personal opinions and/or experiences. However, we felt that the focus on how students write rather than what they write provided little insight into understanding students’ “true” academic writing ability, which should be assessed using tasks more germane to academic writing. Such tasks would assess students’ ability to understand academic assignments, evaluate the logic and validity of presented arguments, and respond by producing “text-responsible prose” (Leki & Carson, 1997, p. 58). Our dissatisfaction with existing tests resulted in the development of a pre-matriculation test in which examinees respond to two carefully selected, culturally sensitive prompts. The test was administered in three consecutive years to 408 students. Data were examined using multi-faceted Rasch measurement (MFRM) to elucidate the performance of the examinees, raters, and the rating rubric along the dimensions of (a) the extent to which the students were able to evaluate the presented arguments and produce text-responsible prose and (b) language and organization. In the validation phase, we assessed the predictive validity of the test by examining whether it predicted the students’ academic performance as well as, or better than, the evidence of English proficiency the students had submitted to the institute at the time of application. With the goal of optimizing the diagnostic potential and the predictive validity of this test, we will outline lessons from earlier administrations of this test and examine salient points related to the test construction, scoring rubric, data analysis, and test validation. We hope that this discussion will engender further refinements of the test while providing guidance for test developers in similar situations with extremely heterogeneous student populations.

11:20-11:45am

Japanese students’ and teachers’ perception and use of score reports for two large-scale EFL tests

Yasuyo Sawaki, Waseda University, ysawaki@waseda.jp
Rie Koizumi, Juntendo University, rie-koizumi@mwa.biglobe.ne.jp

In response to the recent call for offering detailed feedback about language test results (e.g., Kunnan, 2008; Kunnan & Jang, 2009), some large-scale English language tests have started to report fine-grained information about learners’
enhancing test validation through rigorous mixed methods design features

Angeliki Salamoura, Cambridge English Language Assessment, salamoura.a@cambridgeenglish.org

Tim Guetterman, University of Nebraska, tcguetterman@gmail.com

Hanan Khalifa, Cambridge English Language Assessment, khalifa.h@cambridgeenglish.org

As test validation calls for the integration of multiple strands of evidence, mixed methods research carries many potential ways to add value to it. However, conducting a mixed methods study in language testing can be a challenging task. The purpose of this paper is to discuss essential design features of mixed methods studies in language testing. In order to develop recommendations for designing mixed methods studies in language testing, we reviewed a small body (n = 10) of such studies conducted over the last ten years. For inclusion in the review, the studies had to meet three criteria: (i) addressed a language assessment topic, (ii) described a true mixed methods approach
A multifaceted study using web-based video conferencing technology in the assessment of spoken language

**Fumiyo Nakatsuhara** CRELLA, University of Bedfordshire, fumiyo.nakatsuhara@beds.ac.uk

**Chihiro Inoue** CRELLA, University of Bedfordshire, chihiro.inoue@beds.ac.uk

**Vivien Berry**, British Council, vivien.berry@britishcouncil.org

**Evelnia Galaczi**, Cambridge English Language Assessment, galaczi.e@cambridgeenglish.org

Rapid advances in computer technology over the past few decades have facilitated changes in the practice of language teaching, learning and assessment. As one of the ways to offer spoken language lessons, it is now possible for teachers and students in different locations to ‘meet’ using a web-based video conferencing system, offering opportunities to those who would not otherwise be able to attend classes. While the new technology has made face-to-face teaching and learning less logistically complex and thus accessible to more learners, it seems that the use of the technology has not been fully explored in the field of spoken assessment. Most computer-delivered spoken tests to date are semi-direct tests, in which the test-taker speaks in response to recorded input delivered via a computer. The delivery mode does not permit reciprocal interaction between the test-takers and interlocutor(s) in the same way as a face-to-face format, inevitably constraining the speaking ability construct to be measured. This research therefore explored how web-based video conferencing technology could be used to deliver and conduct the face-to-face version of an existing speaking test, and what similarities and differences between the two modes can be discerned. The study sought to examine (a) test-takers’ linguistic output and scores on the two modes and their perceptions of the two modes, and (b) examiners’ test management and rating.
behaviours across the two modes, including their perceptions of the two conditions for delivering the speaking test. 32 test-takers and four trained examiners participated in the study. The test-takers took two speaking tests under face-to-face and computer-delivery conditions using Zoom technology. The order of the two test conditions was counter-balanced for both the test-takers and examiners. A convergent parallel mixed method design was used to allow for a more in-depth and comprehensive set of findings from multiple types of information. The research included an analysis of feedback interviews with test-takers as well as their linguistic output during the tests (especially types of language functions) and rating scores awarded under the two conditions. Examiners provided written comments justifying the scores they awarded, completed a feedback questionnaire and participated in retrospective verbal report sessions to elaborate on their interlocuting and rating behaviour. Three researchers also observed all test sessions and took field notes. The findings suggested that while the two modes generated similar test score outcomes, there were some differences in functional output and examiner interviewing and rating behaviours. In particular, some interactional language functions were elicited differently from the test-takers between the two modes, and the examiners seemed to use different turn-taking techniques between the two conditions. Although the face-to-face mode tended to be preferred, some examiners and test-takers with more exposure to computer-based communication and teaching felt more comfortable with the computer mode than the face-to-face mode. This presentation will conclude with recommendations for further research, including examiner and test-taker training and resolution of technical issues, which should be addressed before any examination boards make any decisions about introducing (or not) a speaking test using video conferencing technology.

Testing the right skill: the misapplication of reading scores as a predictor of translation ability

Maria Brau, Federal Bureau of Investigation, maria.brau@ic.fbi.gov

Rachel Brooks, Federal Bureau of Investigation, rachel.brooks@ic.fbi.gov

US Federal Government agencies often use a measure of reading comprehension in the foreign language to qualify a person for translation work. This practice is based on two assumptions: that as a citizen of the United States, the examinee has a sufficiently strong proficiency in writing English; and that this assumed English proficiency in combination with a measure of reading comprehension in a foreign language constitute sufficient qualification to perform translation tasks. However, the Interagency Language Roundtable (ILR) Skill Level Descriptions (SLDs) for Translation Performance state that reading comprehension of the source language and ability to write the target language are prerequisite but not sufficient for translation. The SLDs posit that translation also involves “congruity judgment,” defined as the ability to choose equivalent expressions in the target language that best match the meaning intended in the source language (ILR, 2007). Still, there has been little research into the relationship between reading ability and translation ability that prove the need for an additional ability (“congruity judgment”) and therefore the need to measure it. Past research has shown that for Arabic, there is a moderate to weak correlation between reading proficiency and translation performance (Brau & Lunde, 2005), but that research was limited to one language and did not consider demographic variables of the examinees. Therefore, the answer to the research question “is reading comprehension a good predictor of translation ability” is unclear. At the Federal Bureau of Investigation (FBI), applicants for translator positions take both reading and ILR-based translation tests as a part
Features of discourse competence in IELTS speaking task performances at different band levels

Noriko Iwashita, The University of Queensland, n.iwashita@uq.edu.au

Claudia Vasquez, The University of Queensland, claudia.vasquez@uqconnect.edu.au

The present study aims to identify distinctive features of oral performance that characterise test-taker discourse at three proficiency levels. More specifically, the study examines the construct of discourse competence observed in the IELTS speaking test performance and how the distinctive features of performance correlate to the IELTS band descriptors. The investigation of discourse competence has been undertaken within a search for further understanding of communicative competence (e.g., Bachman & Palmer, 1996; Chalhaub-Deville, 2003; Purpura, 2008). The search for deeper understanding has prompted the development of various frameworks attempting to provide dependable theoretical foundations of the nature of communicative language ability: these various theoretical models conceptualize communicative language ability as a composite of different sub-competencies which come to explain the degrees of mastery of a L2. Among the sub-competencies suggested to constitute communicative language ability, discourse competence has been assumed to be at the core of the knowledge needed for L2 communication (Bachman, 1990; Bachman & Palmer, 1996). While there seems to be consensus of the importance of greater understanding of discourse competence as a means for further clarity of communicative language ability and L2 proficiency in general, a detailed study of discourse competence appears to somewhat neglected (Kormos 2011; Purpura, 2008), particularly in speaking performance. Discourse competence is one of the four categories identified in the IELTS Speaking Band Descriptor. In order to fill this gap, the current study has undertaken detailed examination of the test-taker oral discourse at three proficiency levels addressing the research question: What are the distinctive features of performance that characterise test-taker discourse at the three IELTS speaking band levels? The transcribed 60 speech samples (20 at each level) of the IELTS Academic Speaking task 2 was analysed both quantitatively and qualitatively by adapting the method employed by Banerjee, et al. (2004) in the analysis of the IELTS Academic Writing tasks. The features of discourse competence analysed in the current study include both coherence (i.e., text generic structure, theme-rheme development) and cohesive devices (i.e., use of reference, ellipsis and substitution, lexical cohesion, conjunctions). The in-depth analysis has revealed some features of discourse (e.g., use of a wider range of conjunctions; more accurate use of referential expressions) were more distinctively observed in the higher level test-
taker performance than lower level test-takers, but other features (e.g., ellipsis, use of reference) was not clearly distinctive according to the levels. Especially other factors such as test-takers’ L1 and task type have an impact on some aspects of coherence regardless the proficiency levels. These findings contribute to further understanding of the nature of oral proficiency and communicative language ability as well as supplement IELTS speaking band descriptors with oral features of test-taker discourse empirically identified in the test-taker performances. Furthermore, the results provide test-takers useful diagnostic information about test-takers’ oral proficiency and will inform language teachers of characteristics of oral proficiency to be targeted in L2 instruction. The study also contributes to the research in the interface between language assessment and SLA.

Structural equation models of calibration, performance appraisals, and strategy use in an IELTS listening test

Aek Phakiti, The University of Sydney, aek.phakiti@sydney.edu.au

The present study investigates test takers’ calibration and its relationship to cognitive and metacognitive strategy use (CSU and MSU, respectively) in an IELTS (International English Language Testing System) listening test. Calibration is a psychological construct which denotes a perfect relationship between a performance appraisal and actual performance. Calibration, in other words, denotes the concurrent validity of test takers’ performance appraisals (evaluation of one’s performance, Bachman & Palmer, 2010). To date, little is known about test takers’ calibration and how it is related to CSU and MSU. To measure test takers’ calibration and strategy use, the present study employed single-case confidence rating scales embedded in each test question, so that test takers can report on their performance appraisal as soon as they answer a test question or complete a test task. Their CSU and MSU were measured prior to the test taking, during the test taking and after the test taking. 300 English as a second language (ESL) international students took part in the study. The data collection stages were as follows. First, prior to taking the listening test, test takers were asked to report on their general CSU and MSU strategy use in listening comprehension and IELTS listening tests. Second, they took IELTS listening test. While completing each of the IELTS listening test part, they were asked to report on the level of their confidence in their test performance (e.g., 0% to 100%) for each question. There were four parts to this listening test. At the end of each listening IELTS section, they were asked to report on 0.25 in the CSU and MSU during that part. At the end of the entire listening test, they were also asked to report their overall CSU and MSU. Test takers’ calibration, performance appraisals (i.e., confidence of test correctness), and reported strategy use scores were analyzed together using a structural equation modeling (SEM) approach. It was found that on average the test takers were not well calibrated and had a tendency to be overconfident across the test sections. Their calibration scores and confidence ratings in performance were positively, yet marginally related to their reported metacognitive strategy use. The present study sheds light into how various aspects of strategic processing interact with one another during listening test taking and subsequently influence language test performance. Implications for the role of strategic competence in test taking and future research will be articulated.
Applying academic vocabulary lists to validating test contents

David Qian, The Hong Kong Polytechnic University, david.qian@polyu.edu.hk

Academic vocabulary lists play an important role in facilitating English as a second/foreign language learners in their academic studies. Since the introduction of University Word List (Xue & Nation, 1984), more academic vocabulary lists have appeared, such as Academic Word List (AWL, Coxhead, 2000), Academic Formulaic List (AFL, Simpson-Vlach & Ellis, 2010), PHRASE List (Martinez & Schmitt, 2012), and Academic Vocabulary List (AVL, Gardner & Davies, 2013). However, as these lists were created following different theoretical frameworks and selection criteria, they may differ in the scope of coverage of academic lexical elements. The present study aimed to analyse the frameworks and criteria researchers resorted to in creating their respective vocabulary lists, and apply these resultant lists to profiling the academic lexical coverage of the Spoken Sub-Corpus of TOFEL 2000 Spoken and Written Academic Language (T2K-SWAL) corpus, which is used as an important database for developing new test items in the speaking and listening subtests of the TOEFL iBT. The results of our analysis indicate that, in terms of overall coverage, AVL seems to capture a much higher percentage of individual academic words than does AWL, and can therefore be considered more robust in detecting the level of intensity of academic discourse. On the phraseology dimension, however, both AFL and PHRASE List appear to be similarly represented in the T2K Spoken Corpus; nevertheless, when the top 30 high-frequency phrases are examined, nuances appear. The paper will explore those fine-grained differences and identify possible reasons which may have resulted in different scopes of coverage between AWL and AVL, and between AFL and PHRASE List. In addition to determining the usability of the above academic vocabulary lists for supporting validity studies of English language for academic purposes, the results of this study also points to the need that, to make a better judgment of item/test quality, it is advisable to employ multiple vocabulary lists for measuring the academicity of a given text or discourse so that not only word and phraseology coverage but also the ranking of the most frequently occurring words and phraseologies can be more reliably determined. Furthermore, since these lists may be purpose-specific when created, in applying them to test validation, it is advisable to ensure with preliminary screening that the vocabulary lists chosen for such validation are relevant and suitable for the intended purpose of the test being evaluated.

Eye tracking of learner interactions with diagnostic feedback on French as a Second Language skill reports

Maggie Dunlop, OISE/University of Toronto, maggie.dunlop@utoronto.ca

This paper reports a study that investigated the cognitive processes taking place when French as a Second Language (FSL) learners interpret diagnostic feedback presented in different presentation modes (e.g., textual vs. visual). In particular, the study examined how these processes differ among students with different learner characteristics (i.e., FSL proficiency level, goal-orientations (Dweck, 1986)). The present study contributes to current theories of learner-feedback interaction in second language learning, a central component of the assessment-for-learning practices (Hattie & Timperley, 2007) that are a key feature of effective foreign language instruction. Specifically, as learner interaction with feedback greatly influences how feedback is used (VandeWalle, 2003), we need to understand ways in which students process feedback to ensure effective integration of assessment with learning (Jang, 2008). The objectives of this study were to investigate how learners engage with and process feedback differently, and what features of descriptive feedback facilitate deeper cognitive processing among language learners. Non-
Francophone students at a bilingual Canadian undergraduate institution received diagnostic feedback reports on their French language skills after writing a placement test designed according to cognitive diagnostic assessment principles (Leighton & Gierl, 2007). The test and reports were delivered by computer. Six students with various background characteristics participated in this study. The students were randomly assigned to one of two groups, to receive either ‘holistic textual’ or ‘specific visual’ descriptive feedback. The students received their reports while an eye tracking device was operating, then participated in a stimulated recall interview. These activities were designed to elicit data on cognitive aspects of students’ attention, processing and interaction with their report. The study analysed the eye tracking data and interviews and found that learners of different psychosocial backgrounds engage effectively with different forms of feedback. Lower proficiency students, who were also highly motivated to master French, reflected on feedback with few emotions and fairly uncritical acceptance of advice. These students read the whole report carefully but paid little attention to achievement charts, and struggled to define learning plans. The higher proficiency students, who were also highly concerned about performing well, engaged in various reflection patterns. The more emotionally reactive students viewed the feedback as an evaluation of themselves rather than their French, and were less likely to engage in careful reading, deeper reflection, and planning. All the students were prompted to deeper reflection by comparisons of self-assessments with test results, sought to understand skill descriptions, and read study advice carefully. This study suggests that students with highly emotive learning approaches would benefit from feedback that allows them to focus on learning. Secondly, learners may not benefit substantially from achievement charts, which are interpreted similarly to numerical marks even when no numbers are displayed. Thirdly, skill descriptions provide a foundational framework for learners to interpret feedback, and must be clearly written and well-developed. Finally, including self-assessment comparisons is key to prompting learners to meaningfully engage with feedback. These study results enhance our understanding of learner-feedback interaction, with future applications available to delivering second language feedback meaningfully to facilitate continued acquisition.

3:30-3:55pm

The interlocutor effect in paired speaking tasks to assess interactional competence

Soo Jung Youn, Northern Arizona University, SooJung.Youn@nau.edu

Paired speaking tasks can be useful classroom assessment tools to measure interactional competence, as students can engage in meaningful conversations and a teacher can efficiently administer a speaking test. Despite its potential, the co-constructed nature of interactional competence (Chalhoub-Deville, 2003; McNamara, 1997; Taylor & Wigglesworth, 2009) presents validity challenges, for example, in applying rating criteria to assign an objective and fair score to each examinee and in interpreting score meanings. Thus, further research on the co-constructed nature of interactional competence is crucial for valid uses of paired speaking tasks for classroom assessment. One of the research directions in the field of assessing interactional competence is to examine the interlocutor effect in paired speaking tasks (e.g., Davis, 2009; Nakatsuahara, 2004). More specifically, peer examinee’s proficiency level might be a potential source that influences the quality of resulting performances and raters’ scoring decisions. Employing a mixed method (Green, 2007; Tashakkori & Teddlie, 2003), this paper aims to investigate the interlocutor effect in paired speaking task performances. Specifically, the paper examines whether and how the peer examinee’s proficiency level systematically influences the co-constructed nature of resulting interaction. For data, two
dyadic role-play tasks were developed to measure L2 English learners’ interactional competence for classroom assessment in an English for Academic Purposes setting. Additionally, analytical rating criteria with detailed descriptions were developed for stakeholders to make meaningful score interpretations and for ensuring raters’ reliable decisions. Examinees range from low-intermediate to advanced-level L2 English speakers (N = 102), consisting of 51 pairs for the dyadic role-play tasks. Six trained raters scored the examinees’ performances using the analytical rating criteria. For data analysis, a many-facet Rasch measurement was conducted using FACETS (Linacre, 2006) with five facets (i.e., examinee, rater, task, interlocutor, and rating criteria) to investigate whether the peer interlocutor’s proficiency influences raters’ scoring decisions. The FACETS analysis indicated that the interlocutor proficiency did not have a significant effect on averaged scores. However, the raters’ scores on sub-scale scores, particularly interaction-sensitive sub-scales, were influenced by the peer interlocutor’s proficiency. In order to further understand the nature of interaction elicited from paired speaking tasks, select examinees’ performances were qualitatively analyzed drawing on conversation analysis (CA) (Sacks, Schegloff, & Jefferson, 1974). CA findings indicated that the examinees oriented to completing the tasks at hand in general. However, the pairs consisting of low-intermediate and advanced level examinees showed distinct interactional features and sequential organizations compared to those with examinees with similar proficiency levels. Based on the research findings, implications for pairing examinees and assigning valid scores when administering paired speaking tasks and ways to utilize examinees’ speaking performances for teaching interactional competence in classroom will be discussed.

Producing actionable feedback in EFL diagnostic assessment

Hongwen Cai, Guangdong University of Foreign Studies, hwcai@gdufs.edu.cn

This study answers an updated call for relating second and foreign language diagnostic assessment to intervention (Alderson, Brunfaut, & Harding, 2014). It is observed that most existing practices in diagnostic language assessment do not lead to actionable procedures, whereas the effect of different types of feedback is typically investigated through group-level comparison in studies in second language acquisition, which does not result in any intervention plans that target the specific needs of individual learners. Clearly, the missing link between second language acquisition and diagnostic assessment is actionable feedback that is suited to the needs of individual learners. To explore how such feedback can be produced, an ethnographic study was conducted on 178 college freshmen in an EFL listening course in China. Information about their comprehension problems and how the problems are related to their learning experience was collected on multiple occasions in the classroom setting through the students’ self-assessment reports, their performance in in-class and take-home tests, interviews based on the students’ performance in the tests, and the teacher’s classroom notes. The data from the multiple sources were coded, frequencies of various decoding problems were counted, and students’ account of their learning experience was matched to the problems. With a focus on the decoding process, eight broad categories of problems were identified from the data: failure to recognize link-ups and weak forms, incorrect word recognition due to mispronunciation, ill adaptation to unfamiliar accents, failure to identify unfamiliar collocations, failure to recognize unfamiliar meanings of familiar words, failure to guess the meaning of new words, failure to identify set phrases, and failure to understand new structures. While most of these categories could find their counterpart in Field’s (2008) list of decoding processes, further analyses showed that
a large part of these problems could be traced to the students’ learning experience, especially the language input they had been exposed to, the instructions they had received up to the point of this study, and their habit of listening. On the basis of this exploration, a qualitative procedure for producing actionable feedback is proposed in the context of EFL listening comprehension. Two key elements of this framework are post-listening diagnosis of the learning process and relation of diagnosis to the learning experience of the learners. The post-listening diagnosis is based on the cognitive framework proposed by Field (2008), comprising decoding and meaning-building processes. Elements of the learning experience include the language input the learners have been exposed to, the instructions they have received, and their habits of listening (Norris & Ortega, 2000). In a nutshell, learners are encouraged to find out about their problems, trace the causes of their problems to their learning experience, and then make their specific decisions as to the choice of learning materials, habit of listening, and immediate steps to take.

C-tests outperform Yes/No vocabulary size tests as predictors of receptive language skills

Claudia Harsch, University of Warwick, C.Harsch@warwick.ac.uk

Johannes Hartig, German Institute for International Pedagogical Research, hartig@dipf.de

Placement and screening tests serve important functions, not only with regard to placing learners at appropriate levels of language courses but also with a view to maximizing the effectiveness of administrating test batteries. With the advent of computer-administered tests (CAT), adaptive testing becomes more and more attractive (e.g. Frey & Seitz, 2009). Here, reliable pre-tests become a prerequisite to optimize the CAT procedure. As stated by e.g. Bachman & Palmer (1996) or Read (2000), the test purpose determines all future decisions; with regard to placement, screening or pretesting, these purposes require a format which has been reported as a reliable and valid measure for screening and placement purposes, is simple and quick in administration and scoring, and poses little demand on the test takers while covering as many items as possible. We examined two widely reported formats suitable for these purposes, the discrete decontextualized Yes/No vocabulary test (e.g. Alderson & Huhta, 2005; Meara, 2005) and the embedded contextualized C-test format (e.g. Eckes & Grojtahn, 2006). Using Read’s (2000) and Read & Chapelle’s (2001) framework for assessing vocabulary as means of classification, the two formats operationalize opposing ends of the framework domains ‘discrete vs. embedded’ and ‘decontextualized vs. contextualized’. We employed these opposing formats in order to determine which format can explain more variance in measures of listening and reading comprehension. Our data stem from a large-scale assessment with over 3000 students in the German secondary educational context: the four measures relevant to our study were administered to a subsample of 559 students. Using regression analysis on observed scores and SEM on a latent level (Kunnan, 1998; Purpura, 1998), we found that the C-test outperforms the Yes-No format in both methodological approaches. The contextualized nature of the C-test seems to be able to explain large amounts of variance in measures of receptive language skills. The C-test, being a reliable, economical and robust measure, appears to be an ideal candidate for placement and screening purposes. With regard to the contextualized nature of the C-test format and its embedded items, positive washback for the foreign language classroom is to be expected (Qian, 2008) since solving C-test items draws on a range of processes and skills, all of which are needed to process language in context (Sigott, 2005). In a side-line of our study, we also explored different scoring approaches for the Yes-No format (Huibregtse, Admiraal & Meara, 2002). We found that using the hit rate and the false-alarm rate as two separate indicators...
yielded the most reliable results. As an additional benefit, these two indicators can be interpreted as measures for vocabulary breadth and as guessing factor respectively, and they allow controlling for guessing. This substantive interpretation offers benefits for the language classroom, giving feedback to learners and teachers alike on vocabulary knowledge and on the use of guessing.

How young children respond to computerized reading and speaking test tasks

Laura Ballard, Michigan State University, ballar94@msu.edu

Shinhye Lee, Michigan State University, leeshin2@msu.edu

Due to the move toward assessment development for young language-learners (e.g., TOEFL Primary and WIDA ACCESS) and the technological innovation that has led to a distinct shift to computerized testing, there is an urgent need for research that investigates young learners’ abilities to complete language tasks via computer-mediated platforms. However, relatively little is known about the ecological validity of computer-based assessment for young language-learner populations (especially recent immigrants and refugees) and how they process and view second-language tests (Butler & Lee, 2010; McKay, 2006; Rea-Dickins, 2000). There are unanswered questions, such as: are young learners able to complete tasks via computer? What affects their ability to complete such tasks? What effects does computerized testing have on young learners? How do young learners view computerized tests? In response to fill this gap, we report on an ongoing study that investigates forty 8, 9, and 10-year-old English-language-learner (ELL) children (10 each from China, Korea, Ethiopia, and Iraq) and twenty same-aged, native-English-speaking children’s performances on TOEFL Primary task, which will be computerized for the purposes of this study. We explore the extent to which the ELL children are able to complete the tasks compared to the native-English-speaking children and their attitudes toward computerized assessments by incorporating a wide array of analyses. The ELL children and the native-English-speaking children will complete two test sections: oral (six TOEFL Primary sample speaking tasks) and reading (seven TOEFL Primary sample reading tasks). We will videotape each child as he or she takes the test, which will be programmed on a Tobii TX300 eye-tracking computer to record eye movements. After the test, we will ask each child (in their L1) to draw a picture to show how he or she felt during testing (draw-a-picture technique; Carless, 2012) and to participate in stimulated recall about their test-taking experience in response to selected eye-gaze recordings. We will compute eye fixations and gaze duration calculations to explore the amount of attention paid to the different test components, which we can aggregate and analyze by test outcome. We will also code and analyze qualitative data using Nvivo 10 for (a) task-completion ability, (b) task effort, (c) ease or stress during testing, and (d) non-standard issues during testing. In our presentation, we will showcase the eye-movement records and qualitative data due in January 2015, which are expected to provide insight into the ecological validity of computer-based assessment for young language-learners. Specifically, we expect that the results will reveal the participants’ difficulties in completing the reading and speaking tasks due to barriers associated with computer exposure and proficiency. Thus, in the final part of our presentation, we will address the need to advocate a greater measure of technology focus in the classroom and computer integration policies in low-level curricula. We will also conclude with implications to testing programs and policymakers in the advancement and modification of language assessment policies regarding K-12 assessment to contribute to the development of more feasible, efficient, and interactive computer-based assessments for young ELLs (Jamieson, 2005).
Assessment is generally seen as one of the main challenges of blended learning. The question of assessing blended learning has been widely debated, with scholars such as Stanard and Basiel (2013) arguing that not only our understanding of what skills and knowledge we need to be effective language users has changed, and hence what we assess, but the way we assess has also changed. However, many researchers point to difficulties with assessing language achievements in an environment in which variables cannot comprehensively be controlled for (Lázaro, 2007). This paper addresses the issue of blended learning assessment with special attention to the information gathered from a specific assessment tool and its learning opportunities. Specifically, in this project, we will be looking at the challenges posed at conducting blended learning in network-limited areas with a technological gap and the impact on human lives of doing much with little, by implementing a virtual community on a limited intranet. We will discuss ELT blended learning assessments in the hands of both teacher and learner, as well as other stakeholders involved who may want to use the information gathered from assessments, and juxtapose these to the generally accepted approaches to assessment that have been developed either in or outside blended learning TESOL facilities. We argue that when feedback is transparent and assessments are carried out with responsibility and autonomy, there can be a positive impact on learning. In conclusion, this paper sheds new light on the neglected issue of getting accountability through assessing language achievements in a blended learning environment.

Test-takers’ use of visual information from context and content videos in the video-based academic listening test

Ruslan Suvorov, Center for Language and Technology, University of Hawaii at Manoa, rsuvorov@hawaii.edu

Several recent studies have explored whether the presence of visuals in L2 listening tests affects second language (L2) learners’ performance on such tests (e.g., Coniam, 2001; Ockey, 2007; Wagner, 2010). To investigate how L2 test-takers use visual information, researchers have analyzed self-reported data collected via post-test surveys (e.g., Coniam 2001; Cubilo, 2011; Suvorov, 2009) or verbal reports conducted during L2 listening examinations (e.g., Gruba, 1999; Ockey, 2007; Wagner, 2008). However, such use of the verbal report methodology is problematic because asking L2 test-takers to verbalize their cognitive processes while they are taking a listening test may negatively impact their performance. To address this limitation, the present study explored L2 test-takers’ use of visuals during video-based L2 listening assessment using cued retrospective reporting, which is a method for examining cognitive processes of the viewers occurring during their eye movements (Van Gog, Paas, Van Merriënboer, & Witte, 2005). In this study, 33 L2 learners, who had taken the Video-based Academic Listening Test (VALT), were shown video recordings of their eye movements captured with a remote eye-tracker during their completion of the VALT and asked to verbalize their cognitive processes at the time they had initially taken the test. Participants’ verbalizations were analyzed qualitatively in NVivo, which is qualitative data analysis software, to identify (a) the aspects of visual information that they had
focused on when watching two types of videos—context and content videos—used in the VALT, and (b) the aspects of visual information that they had perceived as helpful and as distracting for their test performance. The results revealed (a) differences between aspects of visual information that the participants focused on in context videos and content videos, and (b) differences between the aspects of visual information from the two video types that the participants perceived as helpful and as distracting. The implications of this study will be discussed with regards to the use of visuals for the development of L2 listening skills and the use of cued retrospective reporting and eye-tracking methodology for research on video-based L2 listening assessment.

**Comparing aural and written receptive vocabulary knowledge of the first 5k and the AWL**

**Brandon Kramer**, Momoyama Gakuin University, brandon.L.kramer@gmail.com

**Stuart McLean**, Graduate School of Foreign Language Education and Research, Kansai University, stuart93@me.com

This presentation investigates aural vocabulary knowledge and its relationship with written receptive vocabulary knowledge by providing preliminary validity evidence for two tests: 1) the Listening Vocabulary Levels Test (LVLT), an audio test designed to measure aural knowledge of the first five 1000-word frequency levels and lexis from the Academic Word List (AWL) (Coxhead, 2000), and 2) an equivalent written receptive form testing the same target words, hereafter referred to as the New Vocabulary Levels Test (NVLT). Secondly, this presentation will present and discuss the results of this investigation and the implications for vocabulary assessment and instruction. L2 vocabulary assessment literature makes it clear that there is a need for a widely available test that measures aural English vocabulary knowledge. Milton (2009) stated, ‘strangely, it is an aspect of knowledge that has attracted very little systematic interest from researchers’ (pp. 92-93), while Stæhr (2009) lamented that ‘no suitable test of phonological vocabulary size existed’ (p. 597). Milton and Hopkins (2006) investigated aural vocabulary size, finding that very proficient EFL learners frequently have much larger reading vocabularies than listening vocabularies, but the test used (A-Lex) is a size test and does not measure knowledge of respective frequency levels. Within the LVLT and NLVT, each 1000-word band is tested using 24 items created through the retrofit and redesign of previous Vocabulary Size Test (VST) items (Nation & Beglar, 2007). The 30 AWL items, however, were newly created using randomly selected target words and item specifications reverse-engineered from published descriptions of previous VST items. As in the previous VST format, each target word is first presented in isolation, then within a generic sentence designed to provide context and aid recall while not providing any hints for the correct answer choice. Quantitative analyses utilized several aspects of Messick’s validation framework to show that each test is a reliable measure while follow-up interviews and qualitative analyses indicated that each test measured the intended construct of aural vocabulary knowledge, that the formats are easily understood, and that the tests have high face validity. For those students who took both tests (N = 199), significant differences (p<.001) were found between their aural and written receptive knowledge of the vocabulary at the 1k, 2k, 3k, and 5k word frequency levels, with their knowledge of the written forms consistently outperforming their aurally presented counterparts. Overall the gap between the students’ two types of knowledge grew larger as frequency declined, suggesting that orthographic knowledge plays a larger part in remembering a word at those levels. Additionally, more proficient students showed greater discrepancy between the two scores than their less able peers, possibly signaling a lack of curricular balance between teaching orthographic and phonetic knowledge. In conclusion, a written vocabulary test is an inadequate substitute for a test designed to measure aural vocabulary knowledge. The quantitative and qualitative analyses support the
validity of the above tests as accurate, reliable, and distinct measures of each construct. The data supports the necessity of using construct-specific vocabulary tests when conducting high quality research.

**B2 or C1? Investigating the equivalence of CEFR-based proficiency level classifications**

**Johanna Fleckenstein**, Leibniz-Institute for Science and Mathematics Education, fleckenstein@ipn.uni-kiel.de

**Michael Leucht**, Leibniz-Institute for Science and Mathematics Education, leucht@ipn.uni-kiel.de

**Olaf Köller**, Leibniz-Institute for Science and Mathematics Education, koeller@ipn.uni-kiel.de

The Common European Framework of Reference (CEFR) was developed in order to provide a basis for mutual recognition and comparability of language qualifications. It is commonly used for the classification of performance in standardized language tests (Little, 2007). According to its self-description the CEFR can be used for specifying the content of tests and examinations, for stating the criteria to determine the attainment of a learning objective, and for describing the levels of proficiency in existing tests and examinations to enable comparisons across different systems of qualifications (Council of Europe, 2001). The question remains, however, whether CEFR-based proficiency classifications are really comparable. After all, most tests were developed or implemented for different contexts and target populations and some also differ in their construct of language performance (Fulcher, 2004). The validity and equivalence of such CEFR-based proficiency level classifications is the subject of critical discussion (Weir, 2005). Thus, as Charles Alderson put it: “How do I know that my Level B1 is your Level B1?” (Council of Europe, 2003, p. ix). While such critical evaluations of the CEFR are mostly theoretical in nature, there is a shortage of empirical linking studies that aim at analyzing foreign language tests to their compliance in terms of proficiency level classification. This study compares two assessment programs for students of English as a foreign language (EFL) that are mapped onto the CEFR: German National Assessment (NA) and the international Test of English as a Foreign Language (TOEFL). A sample of N = 3,639 German EFL students in the last year of upper secondary education were assessed by a short version of a released paper-and-pencil TOEFL (ETS, 1997) as well as a test that had originally been developed for NA in Germany 2009 (Köller, Knigge & Tesch, 2010). Student abilities were estimated by drawing plausible values (PVs) in a multi-dimensional, one-parameter (1-PL) Rasch model in ConQuest, Version 3.0 (Wu, Adams, Wilson & Haldane, 2007). PVs were recoded into a categorical variable according to the respective corresponding CEFR proficiency levels (A1, A2, B1, B2 and C1) based on cut-scores specified by Tannenbaum and Baron (2011) for the ITP TOEFL and Harsch, Pant & Köller (2010) for German NA, respectively. We found a systematic shift in the classifications of proficiency levels for the two scales: Students tend to score on a lower level on the TOEFL scale than on the NA scale. Thus, B2 on the one scale equals C1 on the other. The two classifications show an only moderate relationship of $\rho = .63$ (reading) and $\rho = .61$ (listening). These findings indicate that two EFL tests do not necessarily locate a student on the same CEFR level. Thus, the interpretation of a student’s actual language proficiency according to CEFR descriptors seems to depend largely on the individual test’s proficiency level classification. By bringing CEFR-linked tests together we can identify deficiencies in the operationalization of the framework for language assessment and standard-setting procedures. It can also contribute to the validation of policy-defined national educational standards.
Assessing the unobservable: Can AFL inform on nascent learning in an L2 classroom setting?

Christian Colby-Kelly, christian.colby@mail.mcgill.ca

It is known that L2 learning is not always accompanied by an observable increase in the accuracy of language produced (Lightbown, 2000). Several (Turner, 2012; Turner & Upshur, 1995; Young, 1995) have noted the difficulty this brings to the accurate assessment of L2 competence. Despite this limitation, it is important for L2 teachers to be aware of the unobservable, nascent learning of their students in order to take the best and most effective direction in subsequent lesson planning. An Assessment for Learning (AFL) methodology may be useful in helping teachers meet this challenge. AFL includes a diagnostic component to help both teachers and students better gauge ‘where students are in their learning’ for more effective lesson planning (Black and Wiliam, 1998). AFL also encourages students to be more aware of the learning process, and to take greater responsibility for determining their own learning progress. The present research will report on teacher and student perceptions of learning in a quasi-experimental investigation of an application of AFL in two classes of pre-university English for Academic Purposes in Quebec, Canada, as part of a larger study centering on the students’ learning of a challenging L2 grammatical feature. The pedagogical materials used included computer-assisted language learning (CALL), an online individual concept mapping (CM) exercise, and peer-group and teacher-class CM exercises. The data collection instruments included the concept maps produced, classroom observation field notes, transcribed group and class discourse, as well as teacher and student survey questionnaires. In addition, pre- and post-treatment tests of L2 production accuracy were administered to indicate trends. The data were analysed by mixed methods and the results triangulated. The qualitative survey data showed strong teacher and student perceptions that learning had taken place in the CALL and CM exercises. Evidence in the form of classroom observation field notes, audio-recordings, and audio transcriptions of group and class interactions, supports this view. The survey data included student perceptions of their learning processes during and following the AFL activities. The results of pre- and post-tests of grammatical production accuracy showed measurable gains in increased accuracy in the grammatical form usage in one of the classes but not in the other. Thus the triangulated results suggest that learning may have taken place in the two classes, and that measurable gains in production accuracy were demonstrated in only one of the classes under investigation. The overall results suggest that learner attestations may more completely inform on the status of internal and unobservable student learning than do production accuracy measures alone. These results have implications for the use of AFL in L2 classrooms to more accurately recognize nascent learning, as teachers build on and scaffold lessons learned, as a basis for future lessons. Using AFL, learners may be well placed to work with teachers in helping assess the unobservable, to inform on individual learning processes, to capitalize on learning progress, and to more accurately determine where students are in their learning, to better plan for future learning.
Empirically derived rating rubric in a large-scale testing context

Alex Volkov, Paragon Testing Enterprises, volkov@paragontesting.ca

Jacob Stone, Paragon Testing Enterprises, stone@paragontesting.ca

Adam Gesicki, Paragon Testing Enterprises/UBC, gesicki@paragontesting.ca

Rating rubrics are an integral part of a performance language assessment, such as The Canadian English Language Proficiency Index Program® General Test (CELPIP-G Test). The majority of the rating rubrics commonly used in the field are constructed theoretically (Knoch, 2009; North & Schneider, 1998). However, recently more attention has been given to empirically-derived rating scales (e.g., Fulche, 1996). We have re-designed the CELPIP-G writing rating rubric utilizing empirically derived data. This presentation reports on the project of empirical rating scale design, with a focus on methodology. Research Objectives The rating rubric revision project is pursuing the following goals: 1. Create a rating rubric with a structure that naturally emerged from the data. A data-informed approach should result in a more intuitive and clear rubric that facilitates rater training. 2. Achieve a better construct representation in the rating rubric by using the judgement of language assessment experts. Conceptual understanding of the functional language proficiency can inform the structure of the rubric at early stages. 3. Scale descriptors empirically capture raters’ perceptions of functional writing proficiency. Method 1. Descriptor Pool In order to achieve a better representation of the functional language proficiency construct, a large descriptor pool was created. The researchers collected 30 openly available rating rubrics. The focus was on non-academic second language writing. Supplemented by descriptors from the current CELPIP rating rubric and rater interview data, the descriptor pool included 800 descriptors. Following the methods described by Knoch (2009) and North and Schneider (1998), norm-referenced modifiers were removed, all the descriptors were encapsulated into stand-alone statements, and redundancy was minimized. Next, CELPIP writing construct was employed to discard the descriptors not directly applicable to the context. At this stage the descriptor pool included 300 unique, positively worded descriptors. 2. Q-Sort To obtain a better representation of the CELPIP language framework in the new rubric, an expert panel was organized. The panelists worked together to conceptually sort the descriptors into 5 conceptual categories. The panel also commented on the descriptors and suggested removing some less clear items. 3. DCL To rank descriptors by dimensions and analyze descriptor performance, another expert panel was organized. The descriptor pool included 152 statements describing test-takers’ performance. 10 panelists mapped descriptors to 100 samples of different CELPIP levels (2 to 12). Over 60% overlap on the samples and the descriptors allowed partial cross-over rating, keeping data integrity for MFRM (Linacre, 1989) modeling to produce usable results. Results 1. Compiling a large descriptor pool resulted in a wider representation of the functional language proficiency construct. 2. Q-sort results enabled to conceptually present the language framework in a form of descriptor categories. 3. Descriptor Checklist allowed to select better performing descriptor and rank them independently by dimensions. In sum, this presentation will introduce the emergent conceptual framework, the methodology through which the rubric was developed, and the MFRM analysis that determined the scaling of the newly developed rubric.

Chinese students’ test preparation for success on high-stakes English language tests

Jia Ma, Queen’s University, jia.ma@queensu.ca

Test preparation (TP) for high-stakes English language tests has increasingly been a research interest in washback studies.
(e.g., Alderson & Hamp-Lyons, 1996; Green, 2007). Mixed findings of TP effects (e.g., Elder & O'Loughlin, 2003; Liu, 2011) indicated the necessity to sequentially explore the nature of test-takers’ TP practices first, then examine the effects of TP to make legitimate interpretations of its effects on test performance, and further delve into test-takers’ perspectives on values of their engagement in TP in order to put the complete process of washback under empirical lens. This study, with a sequential multi-phase multi-method design, investigated nature, effects and perceived values of Chinese students’ test preparation for two major English tests in China – the TOEFL iBT and the College English Test (CET). The inclusion of Chinese students taking TP courses for these two tests provided the opportunity to reach a range of potential participants in terms of their cognitive (e.g., incentives), linguistic (e.g., English proficiency) and sociocultural (e.g., attitudes) characteristics, and in turn served as a contextual foundation to depict a fuller picture of Chinese students’ test preparation. Phase 1 used case study approach and collected multiple-source evidence (e.g., documents, interviews, classroom observations) from multiple stakeholders (e.g., students, teachers, administrative staff) to understand the TOEFL iBT TP practices at a TP centre in China. Phase 2 administrated a questionnaire to 532 undergraduate students in 8 universities in China who had taken a live CET and CET TP courses before test and collected their CET test scores. Drawing on Scriven’s (1998) framework of conceptualizing values from the perspectives of merit, worth and significance, Phase 3 unpacked how Chinese students (N=12) who had started their academic studies in a Canadian university perceived values of their TOEFL iBT TP courses. A one-hour semi-structured interview and a half-hour follow-up interview were conducted with each participant. Test orientation, linguistic emphasis, and sociocultural and cognitive engagements, which were derived by various incentives for test preparation, depicted the nature of Chinese students’ preparation for TOEFL iBT in a particular test preparation context. Result indicated that TP practices for CET, although with an unbalanced weight on test demands over English competence improvement, were not significant predictors of students’ CET test scores. Chinese students perceived values of engaging in test preparation as: access to structured instructions on both test orientation and English learning (merit); possibility of increasing test scores, and need to keep motivation and persistence (worth); and opportunity to be involved in communities formed by peers to prepare for both test-taking and academic applications (significance). These findings combined together, revealed nature, effects, and perceived values regarding Chinese students’ test preparation practices, and all of these depend on factors both within tests (e.g., stakes of tests) and test setting (e.g., social and educational functions of tests). This study contributed to our understanding of how washback functions at the interface of tests and test setting, an aspect of washback studies that needs further research (Cheng, 2014).

10:50-11:15am
Junior and senior high school EFL teachers’ practice of formative assessment: A mixed method study

Hidetoshi Saito, Ibaraki University, cldwtr@mx.ibaraki.ac.jp

The present study is an explanatory sequential design study (Creswell & Clark, 2011) of the current formative assessment (FA) practice of junior and senior high schools EFL teachers in Japan. The EFL teachers use FA to make instructional decisions, which directly affects students’ learning. However, there seems to have been surprisingly few systematic surveys of language teachers’ FA practice in Japan or in any other countries except for Cheng et al.’s trailblazing work (2004). The main purpose of this study is to describe how Japanese junior and senior high school EFL teachers implement classroom formative assessments. The study
comprises of two stages: A large scale survey and a qualitative follow-up study. In the first stage, a 40-item questionnaire was constructed by drawing upon a framework of FA by Wiliam (2010) and relevant sources, including a previous large-scale survey of classroom assessment (Cheng, Rogers, & Hu, 2004) and classroom evaluation (Genesee & Upshur, 1996). The questionnaire consists of items that cover four dimensions of FA—intentions/criteria, assessment method, feedback, and assessment purposes—and the teacher's account of actual FA use on a sample language activity. The questionnaire was piloted and refined through the feedback from 41 practitioners. A necessary sample size of 380 was determined by considering estimated population. The questionnaire was mailed to 512 junior high schools and 205 senior high schools across Japan by way of a stratified random sampling—the strata being 44 prefectures. The total of 732 teachers responded. The data were analyzed first by the Rasch Analysis. This revealed how frequently the teachers employ intentions, assessment methods, feedback, and assessment purposes. I then ran cluster analyses of the Rasch scores of the four FA dimensions along with years of teaching experience, type of school (junior or senior high school), hours of teaching, and class size as covariates. The three cluster solution turned out to be most explanatory and efficient. The four dimensions were all statistically significant contributors of clustering while none of the background variables were so. The three clusters were identified as “high,” “middle,” and “low” users of FA. Based on the Rasch logit scores of questionnaire respondents, one low, one middle and two high users of FA were recruited for the follow-up qualitative study. In the qualitative stage, statistical profiles of these four FA users were explored qualitatively by interviews, observations of their lessons, a follow-up questionnaire, and email exchanges. The results gave both supportive and negative evidence for the three cluster solution of FA users. That is, although salient difference in FA use among teachers seemed to lie in intentions and purposes, there seemed to be great overlaps between “high” and “low” FA users, thus making it difficult to draw clear demarcation between them. The present study reveals not only what kind of FA are most and least commonly used in Japanese junior and senior high schools, but patterns of typical FA use by the teachers. This has significant implications for assessment and testing literacy in in-service and pre-service EFL teacher education programs.

**Constructing a common scale for a multi-level test to enhance interpretation of learning outcomes**

Jessica Row-Whei Wu, Language Training and Testing Center, jw@lttc.ntu.edu.tw

Rachel Yi-fen Wu, Language Training and Testing Center, rwu@lttc.ntu.edu.tw

A criterion-referenced test may contain multiple levels. When a learner takes a higher level of a given test, his or her progress cannot be measured by directly comparing the scores from two different levels. The General English Proficiency Test (GEPT), a five-level criterion-referenced EFL testing system, was developed with reference to the English curriculum in Taiwan. The GEPT is taken by more than 500,000 learners annually for various purposes, among which assessing learning outcomes is the most common. Recently, an increasing number of schools have begun to encourage their students with better English proficiency to take the GEPT at a higher level. Under this circumstance, scores for more than one GEPT level are used in a single school setting. In other words, these schools compare the learning outcomes of students who do not necessarily take the same level of the GEPT. To establish more constructive relationships among teaching, learning, and assessment, it is desirable to create a common scale across all levels of the GEPT by vertical scaling (Kolen & Brennan, 2004). To this end, the present paper reports procedures for constructing a vertical scale for four GEPT levels (Elementary, Intermediate, High-Intermediate, Advanced; roughly equivalent to...
CEFR A2-C1, respectively) through the use of Item Response Theory. The study employed non-equivalent groups with anchor test design, and the test items were grouped into three testlets, each containing shortened versions of two adjacent levels to prevent data from being contaminated due to test-taker fatigue. The between-item multidimensionality model was used to analyze the score data from a total of 1,270 learners. According to the concurrent estimations of learners’ ability (θ) on the common score scale, the corresponding θ of GEPT scores across levels were estimated (in logit). Thus, scores from different levels can be compared on the vertical scale. Major findings include 1. A good fit with Rasch model was obtained. 2. An increase of difficulty across the GEPT levels provided internal validation evidence to support the hierarchical sequence. 3. A difference of 1.0 logit was found between the Elementary level and the Intermediate level, and between the Intermediate level and the High-Intermediate level. However, a wider gap of 1.5 logit was found between the High-Intermediate level and the Advanced level, suggesting that more learning is required for students to achieve the Advanced level. However, it is worth noting that vertically linking tests of different levels provides a weaker degree of comparability than horizontally linking or equating. Despite the constraints on the inferences that vertical linking can support, it is recommended that test-takers’ logit scores on the common scale be provided as a supplement to their GEPT scores when the test results are reported. In this way, the usefulness of the GEPT can be enhanced in order to better inform teaching and learning.

A students’ perspective in a high-stakes test: Attitudes, contributing factors and test performance

Ying Bai, University of Melbourne, yingb@student.unimelb.edu.au

Test takers/students, predominately passive receptors in the testing process, have rarely had their voices and opinions taken into account in test development. When social dimensions and consequences of test are included as evidence of validity (Bachman, 1996; Bachman and Palmer, 2010; Davies, 2008; Kane, 2001, 2002; Kunnan, 2000, 2003; McNamara and Roever, 2006; Messick, 1989; Shohamy, 2001), the test stakeholders’ position has to be re-evaluated. However, students/test takers’ views and their possible contributions to test development are still under investigation. This project seeks to investigate a high-stakes test (College English Test, CET) in China from students’ perspective. The CET, since its inception in the 1980’s, has aimed to play a positive role in promoting effective English teaching and learning commensurate with the College English Curriculum Requirements (CECR), which defines the requirements of each English language skill at tertiary level and gives special emphasis to English communicative skills. To achieve this goal, CET has gone through several major revisions in its development. According to the latest CET syllabus (2006, p.3), it aims to “accurately examine students’ integrated language skills and promote the implementation of the CECR”. This study is composed of two phases, with Phase I exploring students’ attitudes towards the test quality, test use and test preparation approach using qualitative data; and Phase II exploring the possible relationships between students’ attitudes, potential contributing factors to these attitudes and their test performance, using structural equation modelling (SEM) analysis. Twenty second-year non-English major students from two universities took part in a focus group interview in Phase I, with 2-4 students in a 1 to 1.5 hours semi-structured discussion. Students
demonstrated a mixed attitude towards test quality: they believe the CET in general is a valid test, but its major test format (multiple choices) and familiarity with test content, which could be acquired through sufficient test preparation exercises, may have had an effect on their test performance; a supportive attitude towards test use, but the reasons for such attitudes differ from the intended rationale. Their attitudes towards test preparation approach are largely influenced by the language policy at local level: they are more supportive of their English teaching when the policy focuses on improving English use competence; and less so when it focuses more on test results. However, due to the pressure of this high-stakes test, students’ English learning activities outside the classroom mainly focus on CET preparation. The questionnaire for Phase II, based on the interview results, was collected from 358 university students. The SEM analysis demonstrated that students’ English learning motivations and test-related learning experience positively affected their attitudes. Students’ attitudes towards test use and test preparation approach were positively and negatively related to their test performance respectively. Students’ attitudes towards test quality had no effect on their test performance. It seems that students’ perceptions can make a significant contribution in revealing the complexity of the interplay between testing policies at a macro level and their implementation in local settings.

**11:20-11:45am**

**What does the cloze test really test? A replication with eye-tracking data**

**Paula Winke**, Michigan State University, winke@msu.edu

**Shinhye Lee**, Michigan State University, leeshin2@msu.edu

**Daniel Walter**, dswalter@gmail.com

**Katie Weyant**, katie.weyant@gmail.com

**Suthathip Thirakunkovit**, Purdue, thirakun@purdue.edu

**Xun Yan**, Purdue, yan50@purdue.edu

One way to assess students’ language skills is through a Cloze test (Bachman & Palmer, 2010; Brown, 2005). To make a Cloze test, teachers can take a paragraph, delete certain words, such as every 7th or 9th one, and have students fill in the blanks. Teachers can also select a specific linguistic feature to delete, such as articles or modals. Over the years, while researchers have used Cloze tests to assess the proficiency of language learners (Fairclough, 2011; Horst, Cobb & Nicolae, 2005; Baltova, 2000), other researchers have debated what the construct of a Cloze test is (Carr, 2011; Purpura, 2004). That is, what does it measure? Reading comprehension, vocabulary or grammatical knowledge, or something else? In 2011, Tremblay (in a paper published in SSLA) advocated that a French cloze-test she created was a measure of reading comprehension and French proficiency. She suggested researchers could use her test to assess the French of any level of learner. Intrigued by her work, we had 22 individuals with different levels of French knowledge take her Cloze test online. Specifically, the participants were 4 groups of 4 to 5 students, each group from 1st through 4th year university-level-French classes, respectively; 1 non-traditional learner; 2 professors of French; and 1 native speaker. By
tracking test takers’ eye movements while taking the test, we were able to observe whether the participants read the passage, or if they focused their eyes on text solely around the blanks. We found that the test's average item discrimination was 0.345 when we set the test's cut score at below or more than 2 years of French instruction, and 0.341 with the cut at 3 years. These data support Tremblay's argument that the test can be used to segregate low-level learners from upper-level ones. But, we argue that the test may not be appropriate for all learner levels. We use interview data and eye-movement records to argue that the test construct (what is being measured) differs depending on proficiency level, with lower-level students demonstrating that they did not read (or were unable to read) the text in particular. We argue that Cloze testing is a discrete-point vocabulary and grammar test, depending on the missing text, and is not a test or task that normally involves reading comprehension in the traditional sense. Normal reading in the traditional sense normally involves reading comprehension in the traditional sense.

**Factors contributing to fluency ratings in classroom-based assessment**

Douglas Margolis, University of Wisconsin-River Falls, douglas.margolis@uwrf.edu

When language teachers focus on pronunciation, what should be prioritized to most help learners? The pronunciation literature is somewhat split between emphasizing prosody versus phonemic articulation. Many learners frequently express the desire to speak like a native speaker, suggesting a focus on both areas is necessary. But a growing body of literature repudiates the native speaker norm for both practical and ideological reasons, calling into question instructional choices and assessment approaches aimed at helping learners advance their pronunciation to the level of fluent and comprehensible speech that avoids spoiling learners’ second language interactions. For this reason, this study investigated the features of pronunciation most affecting perceptions of fluency and comprehensibility. The objective of the study was to examine exactly what features of student utterances contributed most to oral assessment scores and particularly the fluency and comprehensibility ratings. The researcher collected digitally recorded speech samples from eleven Taiwanese English learners and eleven Korean English learners in three task conditions: (a) reading a script, (b) comparing two pictures of houses, and (c) narrating a story. All speakers were university students living in either Taipei or Seoul. Then undergraduates were asked to rate the speech samples for fluency and comprehensibility. After the ratings were completed the voice samples were transcribed, analyzed, and coded for grammatical, phonemic, and prosodic errors, and then subjected to correlational and factor analyses to assess the types of errors most contributing to fluency and comprehensibility ratings and the degree to which task type influenced ratings. Vowel errors were aggregated into two categories: (a) tense-lax pairs and (b) other. Consonant errors were aggregated into three categories: (a) L1 variants, (b) L1 absent, and (c) other. Both Vowel and Consonant errors were also aggregated into a Phonemic category. Stress and intonation issues were also aggregated into one Prosodic category. Mean length of pauses in milliseconds and mean length of utterances were also calculated. Finally, total number of grammatical errors was included in the analysis, though these issues were not coded for finer grain analyses. The results suggest that prosodic issues most influenced the ratings, especially pertaining to the fluency scores. Nonetheless, phonemic errors also lowered fluency ratings. Grammatical errors may have also impacted on fluency ratings but the data in this study is too limited to make too certain of a conclusion. Moreover, significant differences were found between fluency ratings of the different tasks. Thus, this study suggests that best practices in classroom assessment of pronunciation dictate having multiple task conditions and that prosody should perhaps be weighted more than phonemic articulation. The results will be presented and discussed in terms of the implications for teaching, studying, and
Language assessment literacy in practice: A case study of a Chinese university English teacher

Yueting Xu, University of Hong Kong, traceyxu@hku.hk

Language assessment literacy refers to the knowledge and skills that language professionals need to possess in order to effectively assess student learning. It has received increasing attention in recent years due to the prominent role that it plays in student learning outcomes and the overall quality of education. Conventionally language assessment literacy is understood as a toolkit of knowledge and skills to be acquired and possessed by language teachers. Following this logic, it is assumed that once teachers are taught the basic principles, they would automatically be able to assess student learning effectively. However, assessment literacy is not a straightforward matter. Insights from teacher education research has shown that learning in any aspect of professional work is not merely the acquisition of theoretical knowledge and principles, but rather a complex process full of meaning negotiation and knowledge construction. Thus, the notion of language assessment literacy needs to be reconsidered if a constructive or interpretive epistemology is to be acknowledged. This paper aims to reconceptualize language assessment literacy for teachers on an empirical basis. As part of a larger research project which investigated assessment literacy among 891 university English teachers across China, it reports on a follow-up case study which explores a teacher’s (Linda, pseudonym) ‘assessment literacy in practice’, a key concept to be argued in the reconceptualization of language assessment literacy. Drawing upon multiple sets of data such as classroom observations, semi-structured interviews, and archival documents, it seeks to answer two research questions:

What is the language assessment literacy of the teacher? How is it developed and shaped in a particular context? The findings reveal five themes that comprise Linda’s assessment literacy: 1) heightened awareness of the importance of quality feedback to student learning; 2) structured plans and procedures to engage students in the assessment process; 3) strategic responses to unethical practices arising from assessment; 4) uncertainties surrounding grading; and 5) inadequate preparedness for test construction. These themes are discussed in relation to Linda’s prior learning and teaching experiences, her macro socio-cultural and micro institutional contexts, and relationships with her students, colleagues, and administrators. Based on these findings, the paper proposes a tentative framework for language teacher assessment literacy. It highlights the concept of ‘assessment literacy in practice,’ which is situated, dynamic, and interactive in nature, rooted in theoretical principles, filtered by individual conceptions of teaching, learning and assessment, and shaped by social, cultural, political, and institutional contexts. This reconceptualization attempts to advance the theory of language assessment literacy by framing the concept within the overlapping area between language assessment and teacher education and by arguing for a sustainable development mindset for language teacher assessment literacy. The framework is also expected to have implications for teacher assessment literacy development across contexts and in other subject areas.
3:00-3:25pm

Test use in the transition from university to the workplace: Stakeholder perceptions of academic and professional writing demands

Ute Knoch, University of Melbourne, uknoch@unimelb.edu.au

Susy Macqueen, University of Melbourne, susym@unimelb.edu.au

Lyn May, Queensland University of Technology, lynette.may@qut.edu.au

John Pill, University of Melbourne, tpill@unimelb.edu.au

Neomy Storch, University of Melbourne, neomys@unimelb.edu.au

The number of international students and local students whose first language is not English studying in English-medium universities has increased significantly in the past decade. Although most incoming students need to achieve a certain minimum English entrance requirement for admission to tertiary level study, little is known about the writing standards necessary for graduating students once they enter their respective professions, despite recent interest in English language exit testing from universities. This study gains a detailed insight into the changing writing demands from the last year of university study to the first year in the workforce of engineering and accounting professionals, and relates these to the demands of the writing component of IELTS, which is increasingly used for exit testing, although not expressly designed for this purpose. Data include interviews with final year engineering and accounting students and their lecturers as well as focus groups in which professionals reviewed both final year assignments as well as IELTS writing samples at different levels. The interviews focussed on students’ and lecturers’ perceptions of the writing demands both at university and in the workplace, the students’ perceptions of their preparedness for the language demands of their chosen profession, as well as the similarities and differences between IELTS language, university language and the perceived language demands of the profession. Lecturers of final year core courses were asked about common problems in students’ writing and about typical writing tasks in university courses. In addition, interviews were conducted with new graduates, employers of new graduates and members of professional bodies to explore their experience of workplace writing and its relevance to the university writing tasks and those used in IELTS. While all stakeholders agreed on the importance of ‘effective’ writing in both university and professional contexts, findings point to concern regarding the prevalence of group writing projects used for assessment in engineering degrees. When compared with the university writing tasks, a potential disjunct was apparent in terms of text types, the tailoring of a text for a specific audience and processes of review and editing involved in the workplace writing expected of new graduates. Stakeholders expressed a range of views on the suitability of the use of academic proficiency tests (such as IELTS) as university exit tests and for entry into the professions: in regard to IELTS, while some saw the relevance of both writing tasks, particularly in relation to academic writing, others questioned the extent to which two timed tasks representing limited genres could elicit a representative sample of the professional writing required in the specific contexts, thus raising the issue of general purpose versus specific purpose tests. The final phase, in which employers reviewed texts, showed that they were more able to comment on specific writing issues in relation to the profession-specific final year assignments than they were about the suitability of the IELTS writing samples. The findings are discussed in relation to indigenous assessment criteria and specific purpose language assessment.
Does accent strength affect performance on a listening comprehension test for interactive lectures?

Spiros Papageorgiou, Educational Testing Service, spapageorgiou@ets.org
Gary Ockey, Iowa State University, garyockey@hotmail.com

The construct of second language listening ability is increasingly viewed as the ability to comprehend multiple dialects of the target language, and as a result, a growing number of researchers and practitioners advocate use of a variety of accents for assessing second language listening ability. However, research indicates that the use of multiple accents, some of which are unfamiliar to many of the targeted test takers threaten test fairness because familiarity with an accent has been shown to relate to better comprehension (Adank, et al., 2009; Adank & Janse, 2010; Derwing & Munro, 1997; Gass & Varonis, 1984; Ockey & French, 20014). These studies have focused mainly on monologic discourse, i.e., a single speaker giving a talk, lecture or presentation. However, discourse researchers point out that most presenters monitor their speech and observe the reactions of their audience (Fox Tree, 1999). Because of this monitoring, monologues normally contain two types of discourse: main discourse, that is, the actual topic being discussed, and subsidiary discourse, that is, discourse concerned with the reception of the subject matter and monitoring the discussion (Coulthard & Montgomery, 1981). Academic lectures often contain such elements of interactivity, for example questions by the students and clarifications and further elaboration by the lecturers. It is not known whether the effect on performance found by previous accent studies also holds with these commonly encountered interactive lectures in a listening comprehension test, in particular when only the presenter's accent is unfamiliar. Investigation of interactive lectures is particularly important because, as found in the listening assessment literature, multiple perspectives by different speakers facilitate comprehension of aural input. If this is the case with interactive academic lectures, less familiar accents might be fairly used in tests containing such input. Using Ockey and French’s Strength of Accent Scale (2014), nine accent strengths were used to determine the extent to which comprehension of interactive lectures is affected by accent strength. All speakers were recorded giving the same lecture. A large sample of test takers were then randomly assigned to listen to the lecture given by one of the speakers (N = 21,726) and respond to six identical comprehension items and a survey designed to assess their familiarity with various accents. The test taker responses were further analyzed with the Rasch computer program WINSTEPS (Linacre, 2009) to investigate the relative difficulty of the items associated with the nine versions of the interactive lectures. Results indicated that comprehension of the interactive lectures was diminished after a threshold similar to that found by Ockey and French (2014) for monologic speech, but there was some evidence that this threshold may be slightly higher for the interactive lectures. The implications for the development of listening comprehension tests are discussed.

An Investigation into cognitive evidence for validating a visual literacy task in listening assessment

Luna Yang, School of Foreign Languages and Literature, Beijing Normal University, yanglvna@163.com
Zunmin WU, School of Foreign Languages and Literature, Beijing Normal University, wzm@bnu.edu.cn

Listening comprehension is now perceived as an interactive process instead of receptive process. Meanwhile, it is necessary to recognize the importance of visual literacy for the children of
Paper abstracts illustrated the expected and observed responses generated Item Characteristic Curves (ICCs) representations of photos and maps. Using IRT, the task of matching information with visual information and directions so as to accomplish required to listen to tapes, understand the In the integrative listening task, students are Compulsory Education English Curriculum at fourth level for G8 students set in the National objectives for listening skills and resource strategy listening task is designed in alignment with the from 50,000 to 100,000 G8 students annually. The concurrently from the analysis of data collected in 2013. The overall purpose is to provide the educational administration at different levels and middle schools with feedback information, as well as to better people’s understanding of assessment as learning and teachers’ assessment literacy concurrently from the analysis of data collected from 50,000 to 100,000 G8 students annually. The listening task is designed in alignment with the objectives for listening skills and resource strategy at fourth level for G8 students set in the National Compulsory Education English Curriculum Standards mandated by Ministry of Education of China in 2011. In the integrative listening task, students are required to listen to tapes, understand the information and directions so as to accomplish the task of matching information with visual representations of photos and maps. Using IRT, all the test items were calibrated to determine the difficulty levels in the form of item difficulty measures by performing the ConQuest. The generated Item Characteristic Curves (ICCs) illustrated the expected and observed responses of each item. Students of different proficiency levels (mastery and nonmastery) were invited to re-take the test, where their test-taking process was video-recorded. Then they were asked to recall what they were thinking and how they processed the audio and visual information and worked out the answer to each item. In addition, retrospective data from item writers of constructing these items were also collected. The qualitative data were transcribed and analyzed by NVIVO 10. The quantitative performances and cognitive process of students at different levels are compared, while qualitative data from both students and item writers are analyzed and discussed. The implications of the study on the item writing, and on the instruction of students’ visual literacy and listening proficiency are presented.

3:30-3:55pm

**Tackling the issue of L1 influenced pronunciation in English as a lingua franca communication contexts: The case of aviation**

**Hyejeong Kim**, University of Melbourne, hykim@unimelb.edu.au

**Rosey Billington**, University of Melbourne, rbil@unimelb.edu.au

English has been used as the default language in pilot-air traffic controller radiotelephony communication since 1951, and is therefore used as a lingua franca (ELF) in the international radiotelephony communication context. In 2003, concerns over the role of insufficient English proficiency on the part of non-native English speaking aviation personnel led to the establishment of Language proficiency Requirements (LPRs) by the International Civil Aviation Organization (ICAO). In the LPRs, six assessment criteria including Pronunciation, Structure, Vocabulary, Fluency, Comprehension and Interactions and associated rating scales and descriptors are specified as the basis for
any English proficiency tests developed to meet the LPRs. However, there is a lack of fit between the stipulated requirement and the actual communicative demands of radiotelephony communication from the viewpoint of the domain specialists, an issue that has been previously addressed and criticised (e.g., Kim and Elder, 2014, 2009; Knoch, 2014). Another challenge in the field of language testing is the lack of a means to easily incorporate additional features into descriptors of ELF communication. Focusing on the role and features of the Pronunciation and Comprehension criteria in radiotelephony exchanges between speakers with various L1 backgrounds, we provide a response to some of these challenges. This paper investigates the potential effects of L1 transferred pronunciations on the understanding of messages in radiotelephony communication. A radiotelephony discourse between a French pilot and a Korean air traffic controller in an abnormal situation was collected from the international air traffic control tower in Korea in order to explore these possible influences. Within the framework of grounded ethnography (Douglas, 2000; Douglas and Selinker, 1994), the discourse was analysed from the perspectives of aviation experts, three Korean airline pilots and three Korean air traffic controllers. Their feedback on the performance of two contributors in the discourse was analysed by a thematically based coding scheme. Observations that the experts made on pronunciation features present in the discourse were then supplemented by a close inspection of the phonetic detail of relevant sections of the interaction, using Praat (Boersma and Weenink, 2014). The features from these analyses were then compared with the Pronunciation and Comprehension criteria and associated proficiency descriptors specified in the ICAO rating scale. Results reveal that L1 influenced pronunciation is one of the major factors causing miscommunication on the part of the speaker, together with unfamiliarity with the pronunciation on the part of the listener. Findings also suggest the descriptors for the Pronunciation and Comprehension criteria in the ICAO proficiency rating scale do not reflect the ways the language users (i.e., pilots and air traffic controllers) deal with the issue of L1 influenced pronunciation, which is one of the challenges in the international radiotelephony communication context. Accordingly, it is concluded that a new conceptualisation is needed for taking up the real matters in the assessment of ELF in the aviation context as well as other English as ELF contexts at large.

Examining the effect of DIF anchor items on equating invariance in computer-based EFL listening assessment

Shangchao Min, Zhejiang University, msc@zju.edu.cn

Lianzhen He, Zhejiang University, hlz@zju.edu.cn

While differential item functioning (DIF) analysis has played a principal role in the evaluation of fairness at the item level (Kunnan, 2014), equating invariance concerns fairness at the test score level (Huggins & Penfield, 2012), a lack of which indicates that test takers with the same score on the scale yet belonging to different subpopulations obtain different expected test scores on the equated scale. Given the abundant literature on DIF analysis in language assessment (e.g., Elder, 1996; Pae, 2012; Ryan & Bachman, 1992), it is surprising that little effort has been devoted to examining equating invariance in language tests, especially considering the growing use of parallel test forms in high-stakes language tests in the past decade. Additionally, to our knowledge no empirical study has investigated the connection between DIF and equating invariance in the field of language assessment. To fill this gap, the present study aims to investigate the equating invariance across gender of a computer-based EFL listening test (CBLT), and in particular, the effect of DIF manifested in anchor items on equating dependence. The CBLT is part of the graduation requirements for non-English majors at a major Chinese university. The datasets used were item-level responses of 8,203 test takers, 60% male and 40% female, to 15 different forms of the CBLT. Anchor items were assessed for
both uniform and non-uniform DIF across gender in different forms using item-response-theory-likelihood-ratio test. The Stocking and Lord (1983) scaling method under IRT true-score equating methods was implemented via IRTEQ (Han, 2009). Equating invariance over gender subpopulations under three conditions were compared, in which there were uniform DIF anchor items, non-uniform DIF anchor items and no DIF anchor items. The findings showed that population invariance over gender subpopulation did not hold for some forms of the CBLT, as evidenced by the larger RMSD and REMSD values than a difference that matters (DTM) benchmark at 0.5 score points. However, the elimination of DIF anchor items, both uniform and non-uniform, would result in stronger population invariance in equating across all score levels. Most importantly, equating based on only DIF-free anchor items led to a more consistent pass/fail classification decision between the subpopulation and total population equating functions at the recommended CBLT cut-off score of 30, which is what really matters for the CBLT. The findings indicated that the presence of DIF anchor items indeed dampened population invariance in equating, verifying previous researchers’ conjectures (von Davier & Wilson, 2007) and simulation results (Huggins, 2014) with real-data examples. Theoretically, the present study, by combining the two aspects of measurement invariance and pinpointing the causes of equating invariance, might help develop a more holistic definition of fairness. Practically, it points to the need for language testing practitioners to study DIF in anchor items before conducting equating and provides them with guidelines on how to maximize score equity in equating.

Developing and validating achievement-based assessments of student learning outcomes in an intensive English program

Ryan Lidster, Indiana University, rflidste@umail.iu.edu
Sun-Young Shin, Indiana University, shin36@indiana.edu

Until recently, our Intensive English Program (IEP), like many, used a test of global proficiency to assess student readiness for advancement, despite the curriculum’s stated goal and the requirement of many language program accreditation agencies, including the Commission on English Program Language Accreditation, that students be advanced based on their level of achievement of student learning outcomes (SLOs). Typically, to measure SLO achievement, IEPs use course grades either alone or in combination with a standardized assessment tool (Jensen & Hansen, 2003). However, grades are composite measures which include many factors beyond SLO mastery and which are subject to variation in the means of assessment and rating severity across instructors. Meanwhile, score gains on standardized proficiency tests cannot be linked directly to SLO achievement. Thus, assessing SLO mastery in a reliable and valid manner presents a formidable challenge for language programs, but doing so is crucial for evaluating program effectiveness and demonstrating to stakeholders what students in the program actually become able to do through instruction (Norris, 2006). This study examines the development and validation of a set of in-house language assessments specifically designed to measure mastery of curricular learning outcomes in a multi-level IEP housed in a large, Mid-western University. There were three primary goals for developing a battery of level-based achievement-oriented assessment instruments for each level. First, developing assessments would cause instructors and administrators to refine and operationalize the curriculum’s stated goals and evaluate the appropriateness of its current design
and implementation (Byrnes, 2002). Second, the assessments would provide models enabling students and novice instructors to understand what is expected of successful students at each level. Third, performance data would provide valuable information on how students were actually performing at different levels over time, which in turn could inform curriculum revision (Sempere et al., 2009). In order to realize these goals, assessment development was designed as an iterative process, involving cycles of item writing, teacher and administrator evaluation, piloting, analysis, and revision. Content validity was examined at different levels, with separate teacher judgments on the appropriateness of the item's topics (especially important in a content-based instruction program), discourse structure, grammar, and vocabulary. Student performance was analyzed both using traditional CTT analyses test-externally and in comparison to course grades and separately collected teacher judgments of student achievement levels for individual SLOs. Results provide evidence that the tests corresponded well to other measures of SLO mastery, but more importantly, the act of explicating the connection between program-level assessment and classroom instruction has resulted in a clearer curricular focus. Feedback from both teachers and students suggests the tests were successful in promoting positive washback. Thus, considerable progress has already been made toward the first two major goals of this assessment, and tracking student performance over time will reveal how and where instruction reform is most warranted, underscoring the importance and value of greater coordination between teaching and assessment practices.

4:00-4:25pm

Promoting credible language assessment in courts: Two forensic cases

Margaret van Naerssen, Dr., margaret.vannaerssen@gmail.com

Users of language assessment in legal cases need to promote credibility for language assessment in courts by going beyond simply providing a score. Unfortunately, a judge might believe a test score is enough: “Looks scientific. It’s a test, isn’t it?” Two cases involving non-native English-speakers provide the legal contexts: murder and an undercover sting operation. Language assessment and linking evidence argumentation area presented. The audience has time to briefly become opposing attorneys, cross-examining the “expert witness” (presenter) on choices of language assessment and links to the legal question. Here are two situations commonly encountered. You are an applied linguist/ a language assessment expert. An attorney calls you, asking you to give an opinion on a client’s English proficiency. The attorney offers to send you a video of the police interacting with the client and asks you to say whether the client understood what the police said at the police station. You might be asked to use the interrogation video for your language testing or even to do it by phone, both also saving the cost of doing a live language assessment. Lacking audio recording, you might be asked to just test the client’s English and give an opinion on whether the client understood the police officer during a traffic stop. You need to be prepared to explain that it is not enough to simply assess a defendant’s language proficiency then jump to opining on whether the defendant understood. However, attorneys, even judges and jury members, may not realize the limits of a language test. Also, using a video to test language proficiency would not meet assessment standards. The objectives of this presentation are to illustrate 1. The legal/ court requirements placed on a language
testing expert; 2. The varied roles language assessment might have in a case in the US legal system; 3. The importance of linking language assessment to the legal question, through evidence argumentation; and 4. The importance of being prepared to answer the question: How do you know the person isn't just “faking” his/her language proficiency? Federal Rules of Evidence set the criteria for expert witness testimony in US federal courts and in most state courts, though some states still follow the earlier Frye standard. (Some other countries have similar rules.) A prospective expert witness should understand the rules determining whether the expert is allowed to testify and the acceptability of the language assessment methods. (A Rules handout is provided.) Judges are sometimes persuaded to accept assessment instruments based on reliability data. Experts should be prepared to argue for validity as the primary principle for assessing an individual’s communication abilities for a specific context. Insights from several disciplines can help build links between language evidence and legal questions. Experts may also need to quickly develop a supplementary task, taking care to ground it in theory, research and practice. Forensic linguists usually do not conduct traditional research: analyses are done on a single-subject basis. However, experts can draw on precedents and research from relevant disciplines.

High-stakes test preparation programs and learning outcomes: A context-specific study of learners’ performance on IELTS?

Shahrzad Saif, Université of Laval, shahrzad.saif@coli.ulaval.ca

Liying Cheng, Queen’s University, liying.cheng@queensu.ca

Mohammad Rahimi, Université du Québec à Montréal, rahimi.mohammad@uqam.ca

S

Standardized English language test scores are increasingly used worldwide by university admission offices and government officials for making high-stakes decisions regarding university admissions and immigration to English-speaking countries. This trend has, in turn, given rise to the institutionalization of test preparation activities, particularly in EFL contexts. Whereas previous research has established that high-stakes tests tend to affect the teaching and learning activities in the immediate classroom environment (Alderson & Hamp-Lyons, 1996; Alderson, 2004; Wall & Horak, 2006), there is no systematic evidence that test-preparation courses offered by test-preparation centers significantly raise test scores and, more importantly, English language proficiency particularly in EFL contexts (Green, 2007; Matoush & Fu, 2012). This study explores the nature of test-preparation effect on learners’ test performance and their language proficiency in an Iranian tertiary EFL context. The research questions addressed are (1) whether, and to what extent, test preparation enhances test performance, as shown by an increase in test scores, and (2) whether preparing for a high-stakes test also improves test-takers’ proficiency in English. Using a sequential exploratory mixed-methods design (Creswell & Plano Clark, 2011), the study was conducted in two consecutive phases in a major IELTS preparation center in Iran. The focus of Phase I was to understand the nature of test preparation activities, different stakeholders’
A mixed-methods investigation into the young learners’ cognitive and metacognitive strategy use in reading test

Gina Park, University of Toronto, gina.park@mail.utoronto.ca

Maggie Dunlop, University of Toronto, maggie.dunlop@mail.utoronto.ca

Edith van der Boom, University of Toronto, edith.vanderboom@mail.utoronto.ca

Eunice Eunhee Jang, University of Toronto, eun.jang@utoronto.ca

Comprehending academic reading materials requires a wide range of multiple cognitive skills and metacognitive strategies. Cognitive reading skills refer to mental processes and subconscious actions readers involve to comprehend reading materials, and metacognitive strategies pertain to regulative mechanisms that readers deliberately use to monitor their own reading (Pang, 2008; Sheorey & Mokhtari, 2001). As they both play a critical role in reading comprehension (Williams, 2002), learners who have a limited repertoire of reading skills and strategies may have difficulty comprehending academic texts with abstract or unfamiliar concepts (Oxford, 2002). While a considerable amount of literature has been published on strategy use in reading comprehension (e.g., Anderson, 1991; Block, 1992; Pressley & Aflerbach, 1995), there is little research on how young language learners process academic texts and how their psychosocial characteristics influence the mental processes during reading. The present study investigated the extent to which young learners with different learning profiles would implement different cognitive skills and metacognitive strategies. The study employed a developmental mixed methods research design (Maxwell & Loomis, 2003). In the quantitative phase, the cognitive reading skill profiles for Grade 6 students (n=90) were gained by applying cognitive diagnostic modeling (CDM) to the
population’s reading test performance, while the profiles of their learning goal orientations and perceived ability were created from the latent trait profiling and self-assessment. In the qualitative phase, think-aloud verbal protocols with students (n=14) were used to understand whether or not students with different profiles process academic reading materials differently. The results showed that although inferencing (e.g., predicting future events, making connections between the text and personal experience) is a highly sought strategy among young learners, its effect was subject to students’ reading skills mastery levels. For example, while high cognitive skill masters’ use of inferencing skills led to successful performance on reading tasks, low cognitive skill masters tended to recall irrelevant background knowledge and resulted in misunderstanding the text. The result also showed that learners’ goal orientations influence the reading strategy use. While performance goal-oriented students tended to use strategies associated with re-reading texts or test-taking strategies, such as eliminating distractors, mastery-oriented students tended to use planning/monitoring strategies and were less interested in confirming correct answers. Notably, mastery-oriented, low skill masters showed a high level of engagement in learning throughout the think-aloud session. The research concludes that a successful application of the inferencing strategy requires both basic textual comprehension and making connections, and without the former, the inferencing is not compensated for low skill masters. Moreover, given that the students’ goal orientations prompt different reading skills, it is essential to understand their goal orientations in order to help them develop and utilize diverse metacognitive strategies so that they can self-regulate their own strategy use.

Fri, Mar 20

10:20-10:45am

Aligning teaching, learning, and assessment in EAP instruction? Stakeholders’ views, external influences, and researchers’ perspectives

Talia Isaacs, University of Bristol, talia.isaacs@bristol.ac.uk

Carolyn Turner, McGill University, carolyn.turner@mcgill.ca

This exploratory case study examines the alignment between teacher and student perceptions and classroom practices in a Canadian pre-sessional English for Academic Purposes (EAP) program with respect to the assessment bridge—the link between teaching, learning, assessment, and curricular objectives (Colby-Kelly & Turner, 2007). Due to EAP teachers’ pivotal role in curriculum development, the research site was viewed as having a heightened potential for assessment activities (both formative and summative) to be viewed by stakeholders as integral to classroom activity and not merely conceived of as an afterthought detached from teaching and learning. The overarching goal was to investigate stakeholders’ perceptions of what “counts” as assessment, the purpose and value of tasks not directly linked to summative assessments, the role of pre-emptive and incidental feedback in gauging what students know and facilitating learning, and ways in which external pressures mediate classroom activity. Forty hours of field notes from classroom observations of two advanced-level intact classes (4 experienced teachers; 21 students) were inductively coded and analyzed using macro-level categories for: (1) the assessment bridge—instances of everyday assessment, (2) teachers’ and learners’ perceptions of assessment, (3)
formal classroom assessments, (4) feedback incidents, and (5) extraneous influences beyond the classroom. The observational data were triangulated with post-course interview data and content analyses of curricular materials. The results paint a rich and varied picture of classroom assessment practices, with each teacher relying on idiosyncratic strategies to facilitate learning (e.g., using abstract analogies; considering developmental readiness in feedback provision). Occasionally, students’ preoccupation with assessments that are consequential (i.e., count for marks) and awareness of the prestige value accorded to standardized (e.g., TOEFL) versus in-house tests conflicted with teachers’ promotion of the educational value of classroom activities, as was reflected in teachers’ frequent and often frustrated attempts to redress the focus back to formative components. Overall, students did not appear to subscribe to the view, as articulated by one teacher, that “everything that you’re doing is a part of learning,” revealing a complex power play between competing notions of what counts in the classroom. The results exemplify ways in which the teachers prioritized learning across all classroom activity, which is in line with current research investigating learning-oriented assessment (LOA) in classroom contexts (Turner & Purpura, in press). The data generate a snapshot of classroom stakeholders’ mindsets that demonstrated an expanded view of assessment (i.e., broader than assessment only being formal quizzes/tests) by the teachers, but not necessarily by the students. Teachers capitalized on “assessable moments” in an attempt to enhance learning, which provided evidence of the use of different forms of assessment situations (both planned and unplanned) to benefit further learning on an ongoing basis. This corresponds to an LOA approach where student performance information is used to guide and support the learning process. On the other hand, evidence from the student data has implications for work needed to create a learning-oriented culture in classrooms, where the role of an expanded definition of assessment is considered a vehicle for further learning by all stakeholders.

Diagnostic profiling of foreign language readers and writers

Ari Huhta, University of Jyväskylä, ari.huhta@jyu.fi
Charles Alderson, Lancaster University, c.alderson@lancaster.ac.uk
Lea Nieminen, University of Jyväskylä, lea.s.m.nieminen@jyu.fi
Riikka Ullakonoja, University of Jyväskylä, riikka.ullakonoja@jyu.fi

In recent years, language testing researchers have shown interest in diagnostic testing, which is an area that intersects both language testing and second language acquisition research. Diagnostic testing, however, requires a better and more detailed understanding of language abilities than is currently the case, and this has posed challenges to testers to define their diagnostic constructs both theoretically and operationally. This paper is based on a research project into the diagnosis of reading and writing, which studied two groups of Finnish learners of English as a foreign language: 14 and 17-year-olds. Both groups consisted of about 200 learners who completed a range of linguistic and cognitive tasks as well as replied to questionnaires about their background (e.g., reading and writing habits, attitudes to reading/writing) and motivation to use and learn English. The learners also completed several reading and writing tasks from an international English language examination and a previous research project. Writing performances were double-rated by trained raters and the ratings were analyzed with the multifaceted Rasch program Facets. Rasch analysis was also employed in the analysis of the reading test scores. The results reported here complement our earlier variable-centered findings which indicated that certain cognitive, linguistic and motivational/background measures could predict FL reading and writing in the above-mentioned groups. The current analyses build on the results of those measures and employ them in person-centered analyses that focus on similarities and differences among people. The approach adopted here is
the latent profile analysis (LPA) which seeks to identify homogenous subgroups of individuals, with each subgroup possessing a unique set of characteristics that differentiates it from the other subgroups. At the conceptual level, LPA is similar to cluster analysis but the key difference is that the validity of LPA results can be tested statistically. The software used to estimate the latent profiles was Mplus v. 7.2. The cognitive dimensions included in the analyses were: the efficiency of the working memory, phonological processing capacity, and ease of access to vocabulary in memory. The linguistic dimensions were the size of FL vocabulary, the ability to (linguistically) segment texts in L1 and FL, and L1 reading and writing. Background variables included learners’ frequency of using the FL and their motivation to use the FL. The results of the study contribute to the emerging research base on diagnosing FL proficiency and its development by providing insights into how FL readers and writers are similar vs different with respect to their cognitive, linguistic and motivational features. The findings may also allow us to characterize strong vs weak readers / writers more precisely.

Validation of a web-based test of ESL sociopragmatics

Carsten Roever, University of Melbourne, carsten@unimelb.edu.au

Catherine Elder, University of Melbourne, caelder@unimelb.edu.au

Catriona Fraser, University of Melbourne, fraserc@unimelb.edu.au

Pragmatics broadly describes language use in social settings (Bardovi-Harlig, 2013), and is commonly divided into pragmalinguistics and sociopragmatics (Leech, 1983). The former encompasses the use of linguistic tools for conveying relational meanings and accomplishing social actions, whereas the latter incorporates knowledge of social rules and norms about politeness, appropriateness, conventionality and typical trajectories of interactions (Arial, 2010; Mey, 2001). While pragmatic competence is widely acknowledged as a component of overall communicative competence (Bachman & Palmer, 2010; Canale & Swain, 1980; Purpura, 2004), not many tests exist that assess second language learners’ pragmatic ability, and most have narrow construct coverage. Tests of L2 sociopragmatics have traditionally focused on metapragmatic judgment of speech act politeness and/or appropriateness of production (Hudson, Detmer, Brown, 1992, 1995; Liu, 2006; Yamashita, 1996; Yoshitake, 1997) whereas pragmalinguistic tests assess comprehension of implied meaning and conventional production in social situations (Itomitsu, 2009; Roever, 2005). Recent tests have also evaluated learners’ ability to convey pragmatic meanings in extended role play conversations (Grabowski, 2009; Youn, 2013). However, most sociopragmatic tests have been limited to a small range of speech acts, and role play tests are less practical than non-interactive formats (Roever, 2011). In this study, we developed and validated a web-based test of sociopragmatics for second language learners of Australian English. Following an argument-based validation approach (Chapelle, 2008; Kane, 2006, 2013), we defined our target domain as everyday language use in social, workplace and service settings. We designed a web-based test delivery and rating system, which tested sociopragmatic knowledge by using nine Likert-type appropriateness judgment tasks, five extended discourse completion tasks where a short conversation was reproduced with one party’s turns missing, four dialog choice tasks asking test takers to judge which of two dialogs was more successful, and twelve appropriateness judgment and correction tasks (based on Bardovi-Harlig & Dörnyei, 1998) where test takers indicated whether the second pair parts of two-turn interactions were appropriate, and produced a more appropriate version where necessary. We ran the test with 447 EFL and ESL learners in Chile and Australia as well as 50 native speakers of Australian English. The test system scored receptive sections automatically, and productive sections were scored by two native speaker raters. We attained a reliability of Cronbach’s alpha = .82, and found significant, medium strength
correlations between sections as well as a four-factor structure. As expected, the native speaker group outscored both learner groups, and the ESL group outscored the EFL group. A factorial ANOVA showed that a model incorporating length of residence, proficiency, and amount of daily interaction in English best explained the data, which is in line with findings from previous research in sociopragmatics (Bardovi-Harlig & Bastos, 2011; Matsumura, 2003; Taguchi, 2011; Wang, Brophy & Roever, 2013). Our empirical findings support the generalization, explanation and extrapolation inferences of the argument-based validation approach. In terms of utilization, we concluded that our test is most appropriate for pedagogical and diagnostic decisions with limited utility for higher stakes decisions.

The concurrent and predictive validity of university entrance tests

Bart Deygers, KULeuven, Belgium, bart.deygers@arts.kuleuven.be

In order to gain access to higher education in Flanders, Belgium, non-native speakers are required to pass either PTHO or ITNA - language tests at the B2 level of the Common European Framework of Reference for Languages (CEFR). Every year, some 800 L2 students take such a language test before entering higher education. To date, little is known about the concurrent validity or the predictive validity of either test. In spite of their highly dissimilar constructs and operationalizations, both tests are assumed to measure language at the same level and to discriminate among candidates in the same way. The relationship of both tests to each other and to the target language use domain has never been the subject of rigorous study however. In order to determine the concurrent validity of PTHO and ITNA, 200 prospecting L2 students took both tests a week apart. The primary question guiding this leg of the study is a basic one: do both tests pass or fail the same students? Additionally, a principle component analysis was conducted to shed light on the relative importance of different criteria in each test. A multifaceted Rasch analysis serves to determine the item difficulty in both tests as well as the true ability of the candidates based on their performance in both tests. A common issue in predictive validity studies is the truncated sample problem. Since it is usually only possible to track those candidates who passed a test, it is impossible to determine how failed students would have fared. This study bypasses the truncated sample problem by following students who passed one test, but failed the other. The predictive validity study primarily is of a qualitative nature. A group of 20 students were interviewed on a monthly basis about their linguistic functioning at university and about the language functions they were expected to perform. A small subgroup of students was followed more intensively. Adopting an ethnographic approach, the researcher accompanied these students to class and observed which linguistic obstacles they faced at university. In order to determine the generalizability of the qualitative study, the whole population (N=200) filled out a questionnaire that tackled the same themes. The results of this study shed light on the validity of L2 university entrance tests and on issues of fairness in high stakes language testing. They show how testing constructs are sometimes very distant from real-life demands and how there is not necessarily a direct link between passing a test and functioning in that test' target language use domain. Finally, this study raises salient questions concerning academic language testing constructs and the extent of the inferences one can reasonable make, based on a language test score.
Integrating task-based language teaching and assessment: towards a unified task specification framework

Koen Van Gorp, Centre for Language and Education, KU Leuven, koen.vangorp@arts.kuleuven.be

During the past two decades task-based language teaching (TBLT) has gained considerable momentum as a pedagogy for L2 teaching (Van den Branden, Bygate & Norris, 2009). In TBLT students acquire language by performing authentic meaning-oriented language tasks that have a clear and motivating goal. Not dissimilar from TBLT, task-based language assessment (TBLA) subscribes to a “can do”-approach to language testing by assessing as directly as possible whether test takers can perform specific language tasks in particular, meaningful communicative settings. When applied correctly, TBLA should enable teachers to do more than acknowledge whether students have performed a specific task successfully. TBLA should enable teachers to provide their students with diagnostic feedback. To do this, teachers should be made aware of task specifications, of expected task performance and of task performance strategies so they can help learners improve their performance. In order for TBLA to reach its full didactic potential, it needs to rely on an assessment framework that generates rich information about in-class learning and teaching processes. Consequently, for instructional purposes and purposes of assessment, tasks should be conceptualized as a set of characteristics, rather than holistic entities (Bachman, 2002). A theoretically sound task specification framework enables the integration of different assessment components with task instruction and performance. Additionally, it helps to produce coherent evidence about language learners’ abilities to perform specific target tasks allowing all stakeholders to track and interpret important aspects of learner development over time. This paper addresses the theoretical and practical challenges to create a unified framework for Task-Based Language Teaching (TBLT) and assessment (TBLA). It discusses the task/test characteristics/specifications guiding the selection of task types and guaranteeing content relevance and representativeness for both pedagogic and assessment tasks. The development of such a set of task characteristics for a task-based language syllabus for both L1- and L2-learners in primary education is presented as a case. The task specifications framework was developed by a professional team of syllabus designers by integrating several relevant (re)sources: a needs analysis, official attainment standards, theoretical models of task complexity and empirical evidence regarding language skills development, task selection and sequencing. Relevant stakeholders (e.g. pedagogical advisors and teachers) were involved in every step of its development. Twelve teachers piloted the framework and provided feedback about its applicability on the basis of which the framework was finalized. We will present the framework and give examples of how teachers and students’ talk about tasks and task performances can support the language development of the students. The challenges of implementing such a unified framework for TBLT and TBLA will be presented based on the experience of teachers that have worked for ten years with it. Finally, the theoretical implications for TBLA and classroom assessment will be discussed.
What CEFR level descriptors mean to college English teachers and students in China

Yan Jin, Shanghai Jiaotong University, yjin@sjtu.edu.cn
Shaoyan Zou, Shanghai Jiaotong University, amandazsy@163.com
Xiaoyi Zhang, Shanghai Jiaotong University, zhxiaoyi1201@163.com

The Common European Framework of Reference for Languages (CEFR) in recent decades has gained full momentum in the field of language education. The very power of the CEFR comes in part from its level descriptors, claimed to be definite, clear, brief, and independent (Council of Europe, 2001, pp.206-7). These qualities of descriptors, however, are sometimes challenged in practice, possibly due to inaccuracy of terminology or insufficient consideration of context and theory-based validity (Alderson, 2004; Weir, 2005). With a few exceptions (e.g. Negishi, 2011; Paparougious, 2010), empirical studies have seldom focused on issues with the CEFR descriptors, especially when used by educational practitioners in a non-European context. The moderation of socio-cultural parameters, according to North (2000), is left largely to CEFR users (p.393). In a recent study of aligning China’s College English Curriculum Requirements (Ministry of Education, 2007) with the CEFR levels, we invited English teachers and students, in addition to experts, to make judgments about the alignment with a view to gaining an insight into what the CEFR descriptors mean to the two stakeholder groups. The study was also expected to provide methodological implications for the aligning procedure recommended by the Council of Europe (2009). The research questions were therefore: 1) how well do college English teachers in China understand the CEFR descriptors when used for aligning the Curriculum Requirements with the CEFR levels? 2) how well do college students in China understand the CEFR descriptors when used for self-assessment?

Survey data were collected from a valid sample of 114 teachers and 412 students. Follow-up interviews were conducted with 8 teachers and 15 students. The teacher-participants were required to map 52 CEFR descriptors (25 of reading, 24 of writing and 3 of global scale) at the B1, B2, and C1 levels onto the three levels of the Curriculum Requirements: basic, intermediate and advanced. The student-participants were required to assess their level of English proficiency by stating whether they could fulfil the task described in each of the 52 descriptors. Descriptive and Rasch analyses showed that both teachers and students achieved a better understanding of B1 descriptors than those at the B2 and C1 levels. Teachers’ perception of the difficulty levels indicated a blurred borderline between adjacent levels, especially between the B2 and C1 levels. Interestingly, the students, with a few exceptions, were able to assess their proficiency levels quite accurately using the CEFR descriptors. T-tests indicated differences in the perceptions of teachers with different educational backgrounds and teaching experiences. The interview data provided in-depth explanations for the difficulties experienced by the participants in understanding the CEFR descriptors. Recommendations were made to improve the transparency and accuracy of terminology in the CEFR descriptors on the one hand, and to equip users of the CEFR with a better knowledge of its descriptors on the other. It was also recommended that descriptors specific to the context of English language education in China, especially those at the advanced levels, be produced drawing on the experience of the CEFR.
Investigating a construct of pragmatic and communicative language ability through email writing tasks

Jin Ryu, Teachers College, Columbia University, cellest2@hanmail.net

Despite the importance of communicative language ability (CLA) in today’s globalized society, pragmatic ability, a significant part of CLA, has not been sufficiently and thoroughly investigated in the second language assessment field. Few studies have operationalized pragmatic ability within a CLA model and employed authentic tasks (Grabowski, 2013). To address such problems, the current study operationalized pragmatic ability within Purpura’s (2004) theoretical model of language ability and examined all five subcomponents of pragmatic ability in his model (i.e., contextual, sociocultural, sociolinguistic, psychological, and rhetorical meanings) through authentic email writing tasks. Eighty-two ESL learners in the University Settlement Society of New York participated in the research and were each asked to write four out of eight email writing tasks for a pragmatics test while performing the speech act of apology. The eight writing tasks were administered to each examinee in various four-item testlet versions so as to maintain connectivity during data analysis. The eight pragmatic tasks in the test varied in terms of three variables: the power and distance of the sociolinguistic relationship, and the context in the task (i.e., interactional vs. transactional) (Brown & Yule, 1983; Hudson et al., 1992, 1995). Since the present research had an experimental focus, some variables such as the speech act and degree of imposition were controlled for to enable more specific comparisons between the pragmatic sub-scores achieved on the tasks (Hudson et al., 1992, 1995). The collected data were scored by five native English speaking raters on the five pragmatic subcomponents, two grammatical subcomponents (i.e., grammatical form and semantic meaningfulness), and one holistic score of pragmatic ability (Purpura, 2004).

Many-facet Rasch measurement (MFRM) analyses were conducted, taking into account rater severity and task difficulty, and the true scores that were derived were used for all subsequent analyses. The findings from correlation analyses suggested a moderate amount of interaction between grammatical ability (GA) and pragmatic ability (PA) (r=.694). However, when the interaction was examined across different proficiency levels, only the low proficiency groups had a statistically significant moderate correlation for GA and PA. Further correlation analyses also revealed the distinctiveness of the five pragmatic subcomponents from each other. While the highest correlation between the pragmatic subcomponents was found between contextual and sociocultural abilities (r=.832), rhetorical ability correlated relatively weakly with the other subcomponents. Lastly, the findings from a multiple linear regression analysis revealed that most of the variability in the holistic score was accounted for by the five pragmatic subcomponents (R2=.912), but the source of the remaining variance remains a question for the overall PA construct. Four of the five pragmatic subcomponents except rhetorical meaning contributed almost equally to the variability of PA. In summary, the results provide strong evidence that one’s pragmatic ability should not be based on one’s grammatical ability alone, but should rather be diagnosed in terms of these five pragmatic subcomponents.
Modeling the relationships between construct-relevant and -irrelevant strategic behaviors and lexicogrammar test performance

Nick Zhiwei Bi, University of Sydney, zhbi6097@uni.sydney.edu.au

Test-takers’ strategic behaviors can be broken down into strategies which contribute to construct-relevant variance to test results (e.g., metacognitive and cognitive strategies) and construct-irrelevant variance (e.g., test-taking strategies). Most strategic processing research in language testing just concerns the effect of construct relevant strategies, while most studies do not look at the phenomenon of test-taking (also known as test-wiseness) strategies at all (Cohen, 2014, p.897). Furthermore, empirical research seldom investigates the combined influences of these two different kinds of strategic behaviors on language test performance. Therefore, more studies using empirical evidence to validate how construct-relevant and irrelevant strategic processes are related to test performance are needed. This paper reports on an empirical study which examines the nature of construct-relevant, i.e., metacognitive (planning, evaluating, monitoring) and cognitive (grammar, vocabulary, comprehending, retrieval, and memory) strategic processing and construct-irrelevant, i.e., test-taking (knowledge of test-formats, peripheral knowledge, elimination, clarification, checking) strategies and their direct and indirect relationships with lexicogrammar test performance through the structural equation modeling (SEM) approach. 416 Chinese intermediate level EFL test takers were requested to reflect on their strategic behaviors by answering a lexicogrammatical strategy use questionnaire immediately after they completed a lexicogrammar test. The results showed that both metacognitive and cognitive strategic processes were positively related to lexicogrammar test performance. However, cognitive strategic processing directly explained much higher variances (18%) of test performance than metacognitive strategic processing (1%) did. It was found that metacognitive strategic processing had an indirect effect ($\beta = 0.97 \times 0.43$; $R^2 = 0.17$) on lexicogrammar test performance via cognitive strategic processing. This indicates that metacognitive strategies alone do not influence test performance; rather the way this influence was accomplished was through metacognitive processing exerting an effect over cognitive processing, with both kinds of processing working together closely to impact on L2 test performance (e.g., Phakit, 2008; Purpura, 1999). With regard to construct-irrelevant strategic processing, test-taking strategies were found to have a significant, negative effect ($\beta = -0.13$) on language test performance, which suggests that the more construct-irrelevant strategies the test takers employ, the worse they might perform in the lexicogrammar test. It was also found that there was 83 % shared variance between test-taking strategies and metacognitive processes. This finding may suggest that the weak effect of test takers’ strategic processing on test performance may be resulted from their frequent use of test-taking strategies. In other words, the more test takers invoke test-taking strategies, the less they would use metacognitive and cognitive processing to facilitate their test performance. The finding from this study again suggested that strategic behaviors in test-taking tend to occur in combination with others. In order to use strategies to facilitate test takers’ performance, it depends on both the level of language learners (lower level students tend to invoke many test-taking strategies) and on the combination of the strategies with which test takers use them. Construct-relevant strategic behaviors seem to facilitate test takers’ performance, while construct-irrelevant strategies seem to hinder it.
Testing and assessment are often perceived as somewhat different, or distant, from teaching and learning – something that has to be done but should be quick and cause as little harm as possible, at individual, pedagogical and systemic levels. In particular, this goes for summative aspects of assessment, not least large-scale testing and examinations. Furthermore, there is often a sharp dividing line, drawn or experienced, between formative and summative functions of assessment; assessment for learning and assessment of learning are surprisingly seldom discussed within the same conceptual and pedagogical framework. What remains to be seen is whether the movement towards so called learning-oriented assessment will manage to encompass, clarify and operationalize both basic functions of educational assessment, namely to enhance learning as well as fairness and equity. The current study is set within the Swedish educational context, which has a long tradition of assessment at the national level. Mandatory summative tests, important for students’ continued education and life chances – and for their personal and academic self-esteem – are provided alongside extensive, formatively oriented materials. Building on an expanded view of validity, which focuses on the use and possible consequences of the procedures and products provided, the development of national language tests and other assessment materials is conducted using a collaborative approach. This aims to optimize the quality of the materials - with regard to learning as well as to fairness - but also to contribute to increased awareness and development among those involved. Consequently, different categories of teachers work together with researchers in developing and validating the materials. Moreover, in the different piloting rounds, students are interviewed and also asked to fill in questionnaires focusing on different aspects of the tasks presented. The opinions expressed are included in the overall analyses and decisions, together with other results from the pilot tests, including anchor items. The focus of the proposed paper presentation is on test-taker feedback collected from 15-year old students during regular, large-scale pretesting rounds of summative tests (n ≈ 400/task). The data derive from Likert scales and open responses. The responses have been analysed with a multifaceted approach, encompassing qualitative as well as quantitative methods, and are combined with achievement data. The presentation of results will include a tentative categorization of aspects of assessment perceived as positive and negative, respectively, according to students’ definitions and expectations, as well as analyses of retrospective, task-related self-assessment in relation to achievement. In the latter, certain interesting gender differences have been found. In addition, some comparisons between students’ and teachers’ perceptions of the same tasks will be presented. The practical and conceptual use of collaboration and test-taker feedback in test design, including the creation of construct-based profiles for presenting results, is described and discussed. Finally, some data from post-test questionnaire surveys are considered in relation to the pedagogical potential of assessment – for students’ learning as well as for teachers’ teaching.
Detecting evidence behind the college English curriculum requirements in China: A mixed-methods study

Wen Zhao, Northeastern University, dianawen@hotmail.com
Tan Jin, Northeastern University, alextanjin@gmail.com
Boran Wang, Northeastern University, brwang1972@hotmail.com

Detecting evidence behind the College English Curriculum Requirements in China: A mixed-methods study

English is a compulsory requirement for all college students in China, with the College English Curriculum Requirements (CECR, MOE, 2007) used as the guidelines for English teaching, learning and assessment. With reference to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR, Council of Europe, 2001), a tool for developing language curricula and assessment, China has no such a framework and the CECR for tertiary education has been criticized due to its lack of empirical evidence (see Yang & Gui, 2007). In this connection, the current study reports on a mixed-methods study, aiming to provide empirical evidence from two sources, ‘perception of experts’ and ‘performance of students’. The perception of experts was investigated via the establishment of a descriptor database (the first phase), while the performance of students was examined through the construction of a learner corpus (the second phase). The first phase was the establishment of the descriptor database. A pool of around 500 ‘can-do’ descriptors, representative of three CECR levels in terms of functional language abilities, was collected. Each descriptor was matched against a CECR level using qualitative ‘expert judgment’. Rasch model (see North, 2000) was subsequently employed to calibrate the descriptors quantitatively on a common scale of the framework. The second phase was the construction of a learner corpus. Since vocabulary tests are commonly used as indicators of language proficiency, a vocabulary test was first conducted to screen around 5,000 students from a national key university, with two dimensions of lexical knowledge measured (see Nation, 2008). A learner corpus was then established, including a unique collection of scripts written by the students screened across proficiency levels. It is worth noting that, to represent the English proficiency levels of college students in China, the corpus will involve more students from over 40 universities around the country. The collected scripts were judged by trained raters and assigned with relevant CECR levels. Linguistic features including lexical features, syntactic complexity, and coherence/cohesion, were measured and analyzed to distinguish the students across designated proficiency levels. Using the two sources of evidence from ‘perception of experts’ and ‘performance of students’, it’s anticipated that the project will develop an enriched CECR, contributing to the construction of evidence-based standards of English language proficiency in China.

Automated writing instructional tool for English language learners: A case study of WriteToLearn

Sha Liu, China West Normal University, liusha.sally@yahoo.com

Recent years have seen the significantly increasing popularity of automated essay evaluation (AEE) as an educational technology for both assessment and instructional purposes, despite initial rejections and lingering skepticism (Dikli, 2006; Xi, 2010). Particularly in the last decade, a handful of AEE systems have been further developed to accommodate the needs of English language learners (ELL) in classroom settings, including Criterion by Educational Testing Service (ETS), My Access by Vantage Learning, WriteToLearn by Pearson Education Inc., and Writing Roadmap by CTB/McGraw-Hill. In addition to the primary function of essay scoring, these
systems can also provide instant feedback on the quality of essays by flagging ungrammatical constructions (i.e. error feedback) and offering suggestions on discourse elements such as organization and development of ideas (i.e., discourse feedback). However, little research evidence can be found to support the instructional application of current AEE systems to ELL population. This study investigated the applicability of WriteToLearn to Chinese undergraduate English majors in terms of its scoring ability and the accuracy of its error feedback. Participants included 163 second-year English majors from a university located in Sichuan province of China, who wrote 326 essays from two writing prompts selected from the preloaded prompts in WriteToLearn. The whole set of essays were graded by WriteToLearn and four trained human raters, based on the analytic scoring rubric incorporated in WriteToLearn. Many-facet Rasch Measurement (MFRM) was conducted to measure WriteToLearn’s rating performance as compared to those of human raters. Of the sample, 60 randomly selected essays (30 essays per prompt) were also annotated by the human raters for the analysis of the accuracy of WriteToLearn’s error feedback in terms of its precision and recall. The MFRM analysis showed that WriteToLearn’s preloaded prompts and scoring rubric functioned well in evaluating Chinese undergraduate English majors’ writing ability. In addition, WriteToLearn was more consistent yet highly stringent relative to the four trained human raters in scoring essays and it failed to score seven essays. Given its scoring consistency, severity in rating and occasional failure to score essays, it is suggested that WriteToLearn be used as a second or third rater along with one or more experienced human raters, if it is to be used for the assessment of students’ writing ability. WriteToLearn’s error feedback, with an overall precision and recall of 49% and 18.7%, did not meet the minimum threshold of 90% precision for it to be a reliable error detecting tool (Burstein, Chodorow, & Leacock, 2003; Leacock, Chodorow, Gamon, & Tetreault, 2010). Furthermore, it had difficulty in identifying the frequent errors made by Chinese undergraduate English majors such as article, word choice, preposition and expression, etc. These findings extended previous AEE research on the efficacy of WriteToLearn in ELL classroom settings and provided suggestions for its application to such context. Currently, the implementation of WriteToLearn into ELL writing instruction may be more effective if it can be used as a supplement to rather than a replacement of writing instructors.

Investigating constructs of L2 pragmatics through L2 learners’ oral discourse and interview data

Naoki Ikeda, The University of Melbourne, nikeda@student.unimelb.edu.au

Over the past decades, developments in pragmatics focused on speech acts (Searle, 1969) and politeness (Brown & Levinson, 1987). Recent research has examined L2 pragmatic behavior through a discursive approach (Kasper, 2006; Roever, 2011). In language assessment, pragmatics is increasingly recognized as a key domain (Ross & Kasper, 2013) although research exploring the construct of L2 pragmatics remains limited (for exceptions see Grabowski, 2009; Youn, 2013). The present study attempts to shed further light on the construct of L2 pragmatics in speaking by examining learners’ oral discourse in response to open role-play assessment tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances. 47 participants in higher education with equivalent TOEFL iBT 59/ IELTS 6.0 or above completed six speaking tasks simulating university activities. Interview data was also collected to further explain learners’ oral performances.

The MFRM analysis showed that WriteToLearn’s preloaded prompts and scoring rubric functioned well in evaluating Chinese undergraduate English majors’ writing ability. In addition, WriteToLearn was more consistent yet highly stringent relative to the four trained human raters in scoring essays and it failed to score seven essays. Given its scoring consistency, severity in rating and occasional failure to score essays, it is suggested that WriteToLearn be used as a second or third rater along with one or more experienced human raters, if it is to be used for the assessment of students’ writing ability. WriteToLearn’s error feedback, with an overall precision and recall of 49% and 18.7%, did not meet the minimum threshold of 90% precision for it to be a reliable error detecting tool (Burstein, Chodorow, & Leacock, 2003; Leacock, Chodorow, Gamon, & Tetreault, 2010). Furthermore, it had difficulty in identifying the frequent errors made by Chinese undergraduate English majors such as
of how well the tasks reflected the real world academic; (c) self-assessment of pragmatic performance in relation to the construct. The construct was defined to include the ability to deliver pragmatically appropriate meaning depending on power relations and level of social distance with the interlocutor, and degree of imposition. The follow up interview questions were designed to identify possible construct-irrelevant factors (Messick, 1989) in the task administration. Discourse analyses revealed that several common traits identified in the literature were displayed in the participants' extended discourse. These included laying pre-expansion (Schegloff, 2007); mitigating imposition of the act; laying closing; and choices of linguistic resource (e.g., bi-clausal structure, modal verbs). The general findings are consistent with those in recent empirical studies (e.g., Al-Gahtani & Roever, 2012). The quantitative analysis revealed a tendency for participants to perceive their language use in the tasks to be similar to that in reality and that their performances were not unnecessarily affected by the administration procedures. Participants with higher proficiency especially perceived a stronger similarity between the tasks and the target language use domain. Also, participants at different proficiency groups self-assessed their language use differently, which is reflected in their task performances. Combined with some learners’ volunteered feedback on meaningfulness of the tasks, this study will discuss the constructs of pragmatics measurable and recognized as more important by L2 learners with relevance to the design of effective language assessment tasks to better assess L2 learners readiness for entry into the academic domain.

1:15-1:40pm

The construct of language assessment literacy as perceived by foreign language teachers in a specific context

Deena Boraie, The American University in Cairo, dboraie@aucegypt.edu

Assessment literacy (AL) or language assessment literacy (LAL) was conceptualized based on a deficit model and considered as a gap in the knowledge base of educators which should be addressed by professional development or training programs. This competence-based approach defined AL in terms of abilities or competencies to be acquired such as those listed by Popham (2009) as well as those found in several standards documents. In 2013, a special issue of the Language Testing journal devoted to AL/LAL contributed significantly to the conceptualization and definition of the LAL construct. The LAL construct is positioned in real life areas associated with the professional beliefs, perceptions and practices of educators and conceptualized taking into account the context or the community of practice it is located in (Taylor, 2013). This means that LAL must be conceptualized adopting a more complex and dynamic approach than the previous deficit model in the middle of tensions between the positivist psychometric and interpretivist socio-cultural paradigms of assessment and education (Scarino, 2013). The purpose of this study is to establish a contextualized definition of LAL in a specific language context by teachers of English (EFL) and Arabic (AFL) as a foreign language to adults. Semi-structured interviews as well as classroom observations were conducted for 6 EFL and 6 AFL teachers selected taking into account gender and teaching experience. The interviews focused on teachers’ knowledge, skills and attitudes about language assessment as well as how their LAL developed over time. The observations explored teachers’ classroom assessment practices. A
Can planning at the functional level affect test score?

Gwendydd Caudwell, British Council, Gwendydd.Caudwell@britishcouncil.org

Barry O’Sullivan, British Council, Barry.O’Sullivan@britishcouncil.org

Previous research focusing on planning strategies (e.g. Foster & Skehan 1996; Ortega, 1999; Wigglesworth) have concentrated on the impact on accuracy and fluency or complexity that planning strategies have on spoken performance. Following on from these, this study focuses on planning in what Weir (2005) refers to as an extended informational routine and examines more closely the interaction between planning strategies and the language functions used in spoken production. This study builds on the findings of a previous study (Caudwell, 2014) using functional analysis (O’Sullivan, Weir & Saville, 2002) which examined the effect of the reduction of cognitive load on the performance of young teenagers (13 to 15 year olds) taking a long-turn speaking task. The study highlighted the absence of a number of expected functions (e.g. staging) in the responses examined. Since any evidence of a limited range of language functions might well be perceived by examiners as reflecting a limited language range, and therefore impacting negatively on test score, it is clear that further exploration of the area is required. In order to work towards a more complete understanding of the relationship between functions produced and any effect on overall score, it is important to take into consideration the use made of candidates of the planning time typically built into this task type. In other words, it is important to know to what extent candidates consider specific functions when planning their responses to the task input. Over 100 participants ranging in ability level from A2 to C1 were asked to complete a long-turn speaking task as part of a larger, multi-task computer delivered test of English proficiency. In addition, they completed a questionnaire immediately post-test. This questionnaire focused primarily on the planning time included in the task, and was designed to gather evidence of any strategies used by the candidates. In order to present a multi-faceted perspective of the situation, data from 40 candidates at all levels from the completed questionnaires were compared to a functional analysis and to the scores awarded by trained raters. The findings allow us to draw meaningful conclusions as to whether the effective use of planning time can make a difference in overall score and language produced. These findings, as well as associated implications will be discussed.

Language constructs revisited for practical test design, development and validation

Xiaoming Xi, Educational Testing Service, xxi@ets.org

While theoretical approaches to defining language constructs continue to be debated, recent conceptualizations have converged on the interactionalist view, which highlights the important role of language use context (Chapelle, 1998; Chalhoub-Deville, 2004; Bachman, 2007). However, the practical implications of the interactionalist approach for defining and operationalizing constructs remain to be explored. This presentation will focus on explicating the interactionalist approach for the purpose of guiding the practical design, development and validation of speaking tests. In alignment with the interactionalist view, I argue for a multi-faceted approach to construct definition, consisting of two key elements: foundational and higher-order language skills and use contexts. Theoretical language ability models focus on...
delineating the foundational and higher-order knowledge and skills called for in language use activities (Canale & Swain, 1980; Bachman, 1990; Bachman & Palmer, 1996; 2010; Fulcher, 2003). Frameworks for language use have attempted to describe personal attributes of examinees and characteristics of the language use task and situation that shape language use (Bachman & Palmer, 2010). Characteristics of the language use task and situation essentially define the context of language use. Prevalent theoretical models attempt to articulate the language knowledge components called upon in language use activities in all modalities to facilitate generalization across them. I propose a model of oral communicative competence to provide an organizing structure for describing all underlying components relevant to oral communication, and to provide opportunities for highlighting salient components in this modality-specific model. In addition to knowledge components, this model includes processing components, and oral communication competences as evidenced in examinees’ speech, which form the basis for scoring rubrics. Construct definitions could include all or a selected subset of these skills depending on the purposes of a test, and both task and rubric designs need to be aligned with the construct definition. Given the critical role of context in interactionalist-oriented constructs, I describe a scheme to characterize key contextual factors of oral communication as layers, each of which adds specificity to constrain the contextual parameters of tasks: language use domains, speech genre, goal of communication, medium of communication and characteristics of the setting. I argue that while a model of contextual factors needs to be rich and all-inclusive, including all of the contextual facets would make the construct definition unwieldy and less useful in guiding construct operationalization. Therefore, test designers need to prioritize the inclusion of key facets of the language use context that characterize the target language use domain most meaningfully, and that are expected to account for variation in the quality and features of examinees’ responses. Moreover, test designers must decide which facets to control for and which ones to vary in test design. Appropriate inferences can be made to specify the contexts to which generalizations about language skills can be made. I will use examples to illustrate how the interactionalist approach could be applied to drive optimal test design and development and to guide validation, and discuss how priorities may be different for defining constructs for tests intended for summative vs. formative purposes.

1:45-2:10pm

Implementing formative classroom assessment initiatives: What language assessment literacy knowledge is required?

Ofra Inbar-Lourie, Tel Aviv University, ofrain@post.tau.ac.il

Tziona Levi, ORT Israel, z_levi@netvision.net.il

Formative assessment for learning theories which link teaching and learning with assessment currently form part of the stated policies of many central educational systems, disseminated via assessment tools and teacher professionalization sessions. Questions arise as to the effectiveness of such initiatives, i.e., if and to what extent teachers actually embrace, internalize and implement them and why. Findings from previous research on EFL teachers’ alternative classroom-based assessment practices place teachers’ assessment use in wide social and educational paradigms accentuating its complexity and pointing at the involvement of additional forces in the assessment decision-making process (Inbar-Lourie & Donitsa-Schmidt, 2009). Recent interest in language assessment literacy (LAL) reinforces this concept as it calls for acknowledging the role of additional stakeholders, such as policy makers and administrators, in the assessment process. Such broadening of the assessment circle necessitates a redefinition of the LAL required by the various stakeholders in order to fulfill their role in bridging the gap.
between policy and classroom practice (Taylor, 2013). Hence this research tried to shed light on the LAL components needed by the different parties involved in a teaching learning assessment initiative so as to inform future similar undertakings. The tool in question was an assessment of speaking kit developed by the Israeli Ministry of Education and the National Authority for Assessment and Evaluation for use by EFL teachers in junior high classes. It follows formative assessment rationale and comprises both teaching and assessment activities intended to promote speaking skills aligned with the national curriculum. Pilot research findings conducted at the end of the development stage, point at generally high levels of satisfaction of both teachers and students who used the kit. However, two years after its introduction the digitally available kit is hardly used and this research looked at the reasons why from the LAL perspective. The research sample includes Ministry officials, assessment and EFL experts who decided to create the assessment tool and took part in the process, as well as school officials (principals and administrators) who overlook the teaching and assessment processes in their schools, and EFL teachers. The research tools comprise semi-structured interviews and focus groups intended at eliciting views regarding the kit and its implementation. The data analysis follows the qualitatively inquiry tradition analyzed according to the eight dimensions set in Taylor’s LAL 2013 framework. Preliminary results point at the LAL knowledge components the stakeholders lack and thus need in order to disseminate the assessment initiative: Policy makers lacked understanding as to the implementation phase with resources allotted mainly to the creation of the tool, passing on responsibility to the school administrators who are unaware of the importance of formative assessment, of teaching language strategies and of the resources needed. Finally, teachers were found to lack understanding of formative assessment principles in general and knowhow in the teaching and assessment of speaking in particular. Implications for future formative classroom assessment endeavors on the policy and teacher education levels will be presented and discussed.

Investigating the impact of language background on the rating of spoken and written performances

Jamie Dunlea, British Council, jamie.dunlea@britishcouncil.org

Judith Fairbairn, British Council, Judith. Fairbairn@britishcouncil.org

This paper investigates the impact of the linguistic background of raters on the rating of productive skills in English as a Foreign or Second Language (EFL/ESL) assessment contexts. Previous studies have investigated differences between Native-Speaking (NS) raters and Non-Native Speaking (NNS) raters, but have focused either on the rating of spoken performances (e.g. Brown, 1995; Kim, 2009; Xi & Mollaun, 2011) or written performances (Hill, 1996; Johnson & Lim, 2009) separately. Rater language background in relation to speaking has also been addressed from the perspective of raters’ familiarity with a test taker’s accent (Winke, Gass, & Myford, 2012). The present study builds on this background and makes a particular contribution to the literature by addressing the rating of speaking and writing performances from the same candidates by the same raters, thus allowing for an examination of any differential effect of rater L1 background on the rating of these skills. The study utilized performances collected from a large-scale, international testing program administered via computer to adult EFL/ESL learners in diverse contexts. A total of 60 test takers’ performances on both the writing and speaking components of the test were selected for use in the study. Twenty test takers were chosen from each of three geographically and linguistically distinct backgrounds: Spanish, Japanese, and speakers of an Indian language. Twenty trained and experienced operational raters from the rating pool of the testing program were invited to take part in the study: 10 Native Speakers of English...
language proficiency (ELP) and progress in ELP development. For this accountability purpose, each state administers a summative ELP assessment. These ELP assessment results are also used to determine ELs’ readiness to exit from EL designation. Despite these high-stakes score uses, relatively little empirical research is available to examine the content and measurement qualities of the current ELP assessments. This study aimed to investigate the construct validity of three ELP assessments. Understanding the construct of state ELP assessments is of fundamental importance in making adequate inferences from scores about the types and level of English language proficiency that students have. While the current ELP assessments claim that they measure the academic English proficiency that students need to participate in K-12 school settings, it is not clear how this construct is being measured. Additionally, a positive, strong relationship between ELP and content assessment performance would provide evidence to support the construct validity of ELP assessments. Specifically, we addressed the following research questions: (1) What are the characteristics of academic and social English incorporated in three sample ELP assessments, and how are they represented across the assessments? (2) To what extent are the students’ academic and social language subscores on the ELP assessments related to their performance on their states’ content assessments? We obtained three states’ ELP assessments and the students’ scores on both the ELP and content assessments in English language arts and mathematics. The total sample size included 7,496 ELs (State A: 4,123, State B: 1,133, and State C: 2,240). We first conducted a linguistic content analysis of the items to examine the linguistic demands of each item. Our content analysis rubric included various categories such as types of vocabulary and syntactic complexity. One of the categories classified the items as social, general-academic, or technical-academic language items based on the degree of academic specificity in the item content. Descriptive statistics of the ratings were compared across the assessments. To examine the relationship between ELP and content assessment performance, a hierarchical

A review of the content and statistical qualities of three English language proficiency assessments

Mikyung Kim Wolf, ETS, mkwolf@ets.org
Molly Faulkner-Bond, University of Massachusetts, mfaulkne@educ.umass.edu

In the United States, K-12 public schools are mandated to annually assess and report on their English learners’ (ELs) English
linear modeling (HLM) analysis was performed, since students’ data were nested within schools and districts. The results indicate that the three ELP assessments contained different degrees and types of social and academic language items, both across assessments and across the four language domains (listening, reading, speaking, and writing). Compared to other assessments, State B’s ELP assessment was found to include more items contextualized in specific content areas such as mathematics and science, as well as more technical academic vocabulary words and more complex syntactic features. The HLM analysis results showed that the technical-academic language scores had a significant relationship with math assessment scores. We will present the detailed results of our linguistic and HLM analyses and discuss practical implications of the findings for the valid uses of ELP assessments. We will also discuss suggestions for classroom instruction regarding the types of English language skills that are measured in the assessments.

2:15-2:40pm

Development of innovative assessment tasks to measure spoken language abilities of young English learners

Youngsoon So, Educational Testing Service, yso@ets.org
Diego Zapata, Educational Testing Service, dzapata@ets.org
Keelan Evanini, Educational Testing Service, kevanini@ets.org
Jidong Tao, Educational Testing Service, jtao@ets.org
Christine Luce, Educational Testing Service, cluce@ets.org
Laura Battistini, Educational Testing Service, lbattistini@ets.org

In this presentation we will demonstrate the potential of using technology-assisted tasks to measure the English proficiency of young English language learners. The ‘trialogue’ system, a three-party conversational system between a student and two virtual characters that has largely been applied to teaching content knowledge in online tutoring environments (e.g., Cai, Graesser, Millis, Halpern, Wallace, & Moldovan, 2009), was adopted to develop language assessment tasks. One presumed benefit of using the system in language assessment is that it provides test takers with scenario-based tasks in which multiple language skills, such as listening and speaking, are naturally integrated. The integration of multiple language skills allows test developers to better simulate real-life language use situations. This authenticity of the assessment tasks enables a more direct interpretive link between test performance and success in performing real-life language use tasks. To explore the feasibility of a trialogue system for assessing spoken language abilities, a series of language assessment tasks was embedded within a scenario-based language use context. By design, the tasks incorporated real-life situations into the conversations and provided immediate feedback about the spoken responses in order to engage young students of English at the elementary school level. In each task, students interacted with two virtual characters, and the spoken responses were processed by an Automated Speech Recognition (ASR) system, the CMU PocketSphinx system (Huggins-Daines, Kumar, Chan, Black, Ravishankar, & Rudnicky, 2006). The ASR-generated text of the spoken responses was then processed by the Natural Language Understanding (NLU) and dialogue management modules in the AutoTutor system (D’Mello & Graesser, 2013) to provide an appropriate response to a student’s utterance. A small usability study was conducted in May 2014 with eighteen 3rd-5th grade English language learners in a public school in the USA. Results from the study suggested that the trialogue-based system is a viable method for measuring the English language constructs targeted in the tasks—the correlation between the system-based speaker-level task-completion scores using ASR output and human ratings was $r=0.80$. Moreover, student perception data, collected at the end of the study session,
generally suggested that students perceived the triologue-based tasks positively. For example, most (14 of 18) participating students reported feeling that the virtual characters understood their responses. The presentation will conclude with a discussion of the potential usefulness and current limitations of triologue-based assessment tasks for measuring English proficiency of English language learners with respect to target age groups, the range of constructs, and the types of expected responses.

**Analysis of norm-referencing modifiers as a component of rating rubrics**

**Alex Volkov**, Paragon Testing Enterprises, volkov@paragontesting.ca  
**Jacob Stone**, Paragon Testing Enterprises, stone@paragontesting.ca

Rating scales are an instrumental part of a performance language assessment, such as The Canadian English Language Proficiency Index Program® General Test (CELPIP-G Test). Recently more attention has been given to empirically derived rating scales, as opposed to ‘a priori’ theoretically created rubrics (Knoch, 2009). Knoch (2009) stresses that many existing rating scales have been criticized for using relativistic wording such as modifiers. However, not much research has been done in relation to modifiers that are commonly used to construct rating scale bands. Currently CELPIP-G is revising the writing rating scale by constructing an empirically-derived rubric. The present study employs Many-Facet Rasch Measurement (MFRM) (Linacre, 1989) approach to analyze to what extend modifiers affect the raters’ use of different descriptors and how modifiers may discriminate across different writing proficiency levels. Research Questions In their recent studies North and Schneider (1998) and Knoch (2009) suggest that normative modifiers should be removed before descriptors are empirically scaled. Scarce research in the domain of rating scale modifiers has pushed us to conduct a study to better understand the effects of different modifiers in terms of their power of discriminating between different proficiency levels. More specifically, the current study attempted to answer the following questions: 1) Whether the discrimination power of descriptors will be different when attached to different modifiers? 2) Will the same modifiers attached to different descriptors nonetheless be more discriminating at a specific ability range? Method 8 participants were given 30 writing samples and 15 different descriptors with normative modifiers (partially, sufficiently, and mostly). All the participants were experienced CELPIP benchmark raters and their task was to identify whether the descriptors applied to each sample response or not. All the writing samples were pre-scored in operational CELPIP settings and created a basis for ranking. MFRM (Linacre, 1989) was utilized to rank the modifiers. Results The analysis revealed good overall infit and outfit statistics (Infit MnSq .71 to 1.18, average 0.97; Outfit MnSq 0.70 to 1.26, average 0.76) which allows us to draw useful conclusions from the data. 1. Descriptors with different modifiers attached to them showed clear differences in performance on the writing ability scale. This suggests that modifiers are discriminating and have a major effect on raters' perceptions. 2. The same modifier attached to different descriptors would still cluster quite closely on the language ability scale. This implies that modifiers have a strong discriminatory power at a specific ability range and can provide invaluable information if their performance and parameters are carefully analyzed. In addition, the rater participants reported that they were able to detect a pattern of modifiers' distribution and started applying them more systematically. This finding suggests that using a limited number of modifiers systematically in rating rubrics can facilitate rater training and calibration. The evidence suggests that utilizing modifiers consistently directly affects raters' judgment and provides good discriminating capabilities. We will also suggest future empirical research into modifiers for rating scales.
The relationship between TOEFL and GPA: The case of Chinese students

April Ginther, Purdue University, aginther@purdue.edu
Xun Yan, Purdue University, yan50@purdue.edu
Joe Potts, Purdue University, jdpotts@purdue.edu

This study examines the predictive validity of the TOEFL iBT with respect to academic achievement as measured by first- and second-semester GPA of Chinese students enrolled in a large public university in North America. The TOEFL iBT scores submitted in students’ admission files and the first year GPAs of 1,738 students from mainland China, enrolled across three academic years (N2011=702, N2012=514, N2013=522), were examined. Correlations between GPA, TOEFL iBT total, and subsection scores were included in the analyses. The coefficients were also adjusted for attenuation using the Thorndike case 2 formula to correct for restriction of range. In addition, cluster analyses on the three cohorts’ TOEFL subsection scores were conducted to determine whether different score profiles might help explain the correlational patterns found between TOEFL subscale scores and GPA. For the 2011 and 2012 cohorts, strong and positive correlations were observed between speaking and writing subscale scores and GPA, and strong but negative correlations were observed for listening and reading. In contrast, for the 2013 cohort, only writing subscale scores were positively correlated with GPA. Results of cluster analyses suggest that the negative correlations in the 2011 and 2012 cohorts may be associated with a distinctive TOEFL iBT score profile: High on reading and listening (>27), but relatively low on speaking and writing (<20), and a total score above 90. Interestingly, this group had the lowest average GPA (<3.0) as compared with the other score profile groups that emerged in the cluster analyses. In the cluster analysis of the 2013 cohort, no comparable score profiles were present. Not only did the 2013 cohort not contain students with the distinctive TOEFL iBT score profile, in addition, all score profile groups achieved an average GPA above 3.0. Assuming an expected positive but weak relationship between TOEFL iBT subscale scores and GPA, the negative correlations between the reading and listening subsection scores and GPA for the 2011 and 2012 cohorts were surprising. These correlational patterns may be associated with the Reading/Listening vs. Speaking/Writing discrepant score profile of subgroups within the 2011 and 2012 cohorts. The findings underscore the need to consider subsection scores when evaluating ESL students’ language proficiency for both admissions and research purposes. Possible explanations for the uneven score profile and its consequences on academic achievement will also be discussed, along with the important implications for admissions decisions and provision of language support for Chinese students accepted into North American universities.
Testing as social practice: Canada and context in construct representation, test interpretation, and arguments for consequential validity

Janna Fox, Carleton University, Janna_Fox@carleton.ca
Beverly A. Baker, University of Ottawa
Michel Laurier, University of Ottawa, mlaurier@uottawa.ca
Dean Mellow, Simon Fraser University, dmellow@sfu.ca
Kaitlyn Begg, McGill University
David Slomp, University of Lethbridge, david.slomp@uleth.ca
Bruno Zumbo, University of British Columbia, bruno.zumbo@ubc.ca
Bryan Maddox, University of East Anglia

Discussant:
Carolyn Turner, McGill University

Canada’s multilingual diversity, its bilingual history, and interactions between Official languages (French and English) and indigenous languages, have led to a tradition of language assessment which is increasingly marked by recognitions regarding context: in construct representation, test performance, and assessment practices which impact teaching and learning.

Within the broader field of language testing, this more socially-informed view was reinvigorated by Messick (1989), who argued that construct definition was a value-laden process embodying both social and political meanings. Following Messick, a growing body of research has focused on the social consequences of test use, test impact (Wall, 2005), and washback (e.g., Cheng, 2005; Cheng, Watanabe & Curtis, 2004). Social theory (e.g., Bourdieu, 1991; Foucault, 1972; Vygotsky, 1987) has increasingly informed research in language testing (e.g., McNamara & Roever, 2007; Shohamy, 2001; Swain & Brooks, 2013) leading to new recognitions and prompting scholars such as McNamara (2007) to suggest, “arguably the greatest challenge facing language testing is the issue of the context in which the language testing is carried out” (p. 131).

Although currently, we find fewer proponents of the view that measurement is a purely cognitive enterprise which occurs “in a social vacuum” (McNamara, 2008, p. 417), it is still the case that many texts on testing and test development (e.g., Downing & Haladyna, 2006): 1) avoid explicit discussion of the local or social context within which tests act; 2) are largely detached from considerations of individual stakeholders, socio-cultural groups, indigenously drawn constructs, and the social identities of test takers; and, 3) only rarely examine testing practices through a broader social lens of fairness (Cole & Willingham, 1999; Kunnan, 2014), equity, and justice. If fairness is discussed in such textbooks, it is typically related to the merits of such psychometrically-driven testing formats as multiple-choice.

Within the Canadian context, however, discussions of fairness inevitably lead to the recognition that socio-cultural context is an inescapable element of testing practice. The present symposium brings together Canadian researchers who are informed by socio-cultural theory in investigating testing practices and their
washback, impact, and consequences, at both the micro and macro levels of context. Taken together, these papers provide a window on the unique contributions of Canadian research to language testing, and provide evidence that such research is indeed a response to calls in the language testing literature for a clearer understanding of context (see, for example, Chalhoub-Deville, 1995; Maddox, 2007; McNamara, 2007) in test development, validation, and interpretation. All of the researchers in this Symposium have taken an ecological approach (Fox, 2003), acknowledging socio-cultural context as a critical component of construct development, test interpretation, and consequential validity arguments.

The consequential validity of language tests for pre-service teachers in Quebec

Beverly A. Baker & Michel Laurier, University of Ottawa

This presentation focuses on issues of consequential validity in relation with the implementation of two language tests for prospective teachers in Quebec. The Quebec Deans of Education responded to a social demand with a language policy which included the design of a proficiency examination for all prospective teachers being trained in Quebec – either native or nonnative, in French or English. Most of the candidates are enrolled in programs in French and have to take the French examination (TECFEE). This exam has been created in 2008. The following year, an English exam (EETC) has been developed for candidates you will be working in English school boards. During each examination’s independent development process, differing interpretations of the same construct have led to two very different final products.

TECFEE is a handwritten examination. In Part 1, candidates are asked to write a 350-word text related to an educational issue. Raters must assess two facets of the production: Discourse and Accuracy (based on the number of grammar and vocabulary errors). Part 2 consists of multiple choice questions on grammar and vocabulary where candidates usually have to identify the correct form or definition.

EETC is a two-hour computer-based exam. Part A is an objective section consisting of sentence editing items and multiple choice items. Part B consists of two integrated reading and writing tasks which are aligned on authentic interactions with parents and colleagues and can be considered as an ESP language performance assessment. The performance is assessed holistically.

The social demand, the expected impact and the representation of teachers’ language competence may explain why these two assessment instruments appear to be so different. In addition to the comparability issue, we identified four concerns in relation with the impact of TECFEE which also applies to EETC.

• Under-representation of the construct
• Expectations that do not consider the test limitations
• Negative washback effect
• Barriers to access the profession

Assessing indigenous languages within socio-political contexts: The importance of local input

J. Dean Mellow, Simon Fraser University, & Kaitlyn Begg, McGill University

The languages of the Indigenous people of North America are remarkably diverse in terms of both distinct language families and regional variations in use. For these languages, local decision-making about the assessment of the language abilities that are valued and viable to achieve is necessitated by the limitations of published information about these languages, by the dynamic and variable patterns of multilingual language use, by the diverse types of teaching and learning contexts, and by the historical context of colonization with its difficult political relationships. In most cases, empirical evidence for the construct validity of assessment instruments (correlations or comparisons to pre-existing tests that have
been deemed valid) is not possible because such pre-existing tests do not exist for most of these languages. Content-related or logical evidence for construct validity primarily depends on the expertise of local speakers and educators, especially because language is so closely intertwined with identity and culture.

This paper discusses potential types of collaborations between local communities and language assessment developers, including a recent collaboration in the speech-language pathology profession between McGill University, the Montreal Children’s Hospital, and the Cree Board of Health at James Bay. This collaboration attempted to provide culturally appropriate speech-language services to northern communities through telecommunication assessment sessions and community resource development.

The paper also discusses the tensions in political situations in which the amount of local community input in educational policies and assessment is contested by local educators and federal politicians (e.g., challenges with Canadian Bill C33, the First Nations Education Act; and the U.S. federal policy “No Child Left Behind”).

**Integrating consequential validity evidence into the design and validation process for both large-scale and classroom assessment.**

David Slomp, University of Lethbridge

The increasing diversity of students in contemporary classrooms highlights the importance of developing assessment programs that are sensitive to the challenges of assessing across diverse contexts and with diverse populations. The issues of diversity and contextual sensitivity also make clear the importance of attending to the social consequences that result from the development and use of assessment tools/programs.

Within the measurement community, however, there has been considerable debate about the role consequential validity evidence plays within the broader validation process. On the one hand, assessment scholars such as Popham and Cizek have largely dismissed consequential validity evidence as being a pseudo form of validity. They also argue that often studies of consequential validity evidence focus too heavily on large-scale assessments while ignoring classroom assessments (which often carry more significant social consequences for students than do large-scale assessments). On the other hand, scholars such as Lane and Brennan have argued that consequential validity is an essential but too often ignored category of validity evidence. Lane further argues that one reason consequential validity evidence is missing in much of the validation research is that a systematic approach for collecting such evidence has not been developed.

This paper addresses both of these concerns. First, it presents a framework for systematically integrating considerations of consequential validity issues into each phase of Kane's validation model. Second, drawing on this model and fusing it with Huot’s concept of assessment as research, it presents both a process guide and a design heuristic based on this framework that can guide classroom teachers through a more rigorous assessment design process, one that compels them to attend to social consequences of their assessment design choices.

**Inviting the study of consequences by using ethnographic-psychometrics in operational testing programs**

Bruno D. Zumbo, University of British Columbia, & Bryan Maddox, University of East Anglia

Recently, Zumbo and Chan (2014) presented a systematic review of validation practices as a study of the scholarly genre of validation reports – its focus, style, orientation, and structure – and how this genre frames validity.
theory and practices across many disciplines and international jurisdictions. One of their starkest findings was that the study of consequences is consistently under-represented in validation practices. One of their recommendations is that it is important to expand the evidential basis of validity to include qualitative and mixed methods evidence as a way to invite the study of consequences. The purpose of this paper is to present an overview of our program of research that draws upon (a) an ecological model of item responding, (b) the Hubley-Zumbo (2011) framework of consequences in validation, and (c) what we have termed ‘ethnographic-psychometrics’ or ‘psychometric-ethnographics’, depending on your methodological orientation and proclivities. We will provide several cases from a forthcoming paper that demonstrate how the ethnography of testing situations, in an operational testing context, can combine with and complement psychometric analysis to identify and explore substantive hypotheses about test, item, and group performance. Our aim is to marry ethnography and psychometrics in a manner that does not privilege one over the other and hence we are going beyond the mere engagement of subject matter experts to aid in the interpretation of statistical findings after-the-fact. We will demonstrate how our method engages experiential experts in ethnography to both help inform as well as shape conclusions about the test performance. In short, the ethnography is not meant to simply provide explanations for the statistical findings but rather it meant to go beyond that to provide a richer sense of the ecology of the testing situation, and in so doing invite the study of consequences in test validation.

SYMPOSIUM 2

Thurs, Mar 19

10:20-11:45am (Mountbatten A)

Assessment for learning (AFL): teaching, learning and assessment in the writing classroom

Wenxia Zhang, Tsinghua University, wxzhang@mail.tsinghua.edu.cn

Mingbo Shen, Tsinghua University, shenmingbo01@163.com

Discussant: Jing Huang, China West Normal University, annahuang9466@sina.com

The goal of this symposium is to present how Assessment for Learning (AFL) Theories are implemented and researched in the target writing course and discuss some emergent issues such as how to develop students’ autonomy with all parties (teacher, the student writer, peer and machine) involved and how to set assessment criteria to guide and integrate teaching, learning and assessment in the classroom.

This symposium consists of four presentations presented by a classroom-based EFL writing instruction and assessment research team in mainland China from three perspectives- teaching, learning and assessment. The organizer and moderator, Prof. Zhang Wenxia, a full Professor in language testing and assessment at Tsinghua University in China will chair this symposium. Presenters are Associated Prof. Sun Haiyang and Shen Mingbo, two Lecturers, Huang Jing and Gao Manman and one current PhD student, Wu Yong. The first presentation is a conceptual framework to umbrella the other three presentations. It is a logical argumentation of the integrated teaching and assessment frame with Argument Use Assessment (AUA), which offers the theoretical backup for the design of the course (presentation
2) and the impact of the integrated feedback approach in EFL writing instruction (presentations 3 and 4). The second presenter is going to introduce how the writing course is designed and how the integrated teaching and assessment framework is manipulated in the course. The third presentation is mainly focused on how the integrated feedback approach is applied and its impact on students' writing conceptions and performance. The fourth presenter is going to discuss about how the integrated feedback helps develop the student writers' language autonomy. At last, the moderator will wrap up the four presentations, issues that may emerge and implications for future research.

An AUA validation study of the integrated teaching and assessment framework

Haiyang Sun

How to assess students' learning of a foreign language has become one research focus worldwide for language teaching practitioners and researchers. In order to overcome the limitations of the summative one-shot assessment framework in the Chinese College English learning and teaching context, a four-stage and multi-agent integrated teaching and assessment framework was proposed as the result of synthesizing the Assessment for learning (AfL) theories and the related social learning theories, as well as taking into consideration the current educational technology development and the researcher’s teaching experience. This presentation is to propose the design validity argument for the integrated teaching and assessment approach whose interpretations and uses can be justified. With the application of the theoretical framework of Assessment Use as Argument (AUA) put forward by Bachman, the backing for the inferences of the interpretive argument of the integrated Approach in this stage is mainly from theoretical analysis. The inferences, warrants, assumptions, and backing will be clearly stated to demonstrate the feasibility and viability of the integrated teaching and assessment approach on students’ performance and perceptions in order to explore a new way out to improve the teaching of College English writing in the new era.

The implementation of assessment for learning theories in a writing course

Mingbo Shen

This paper intends to explore how to implement the integrated teaching and assessment framework in course design and classroom management. The presenter also will talk about how student voice is heard in the course and the function of criteria in the course management. The presenter will first give a brief introduction of the design of the course which involves two writing instruction processes guided by the integrated teaching and assessment framework. The presenter will then investigate how the integrated teaching and assessment framework is applied and implemented in the writing classroom. Detailed information will be provided on what writing tasks and learning activities as well as assessment practices are involved, and how they are implemented in a 16-week EFL writing course. The design and development of task-specific writing assessment criteria will be specially introduced and discussed. A number of practical concerns and issues, such as how student voice is heard and the roles of assessment criteria, will be addressed.

The impact of the integrated feedback approach on students’ perceptions and performance

Jing Huang & Manman Gao

In order to overcome the limitation of the summative one-shot feedback approach, this paper attempts to propose an integrated feedback approach and further explores its impact on students' writing conceptions and their writing
The impact of integrated feedback approach on student autonomy

Yong Wu

The study was conducted to investigate how the Integrated Feedback Approach influenced students’ autonomy in EFL writing. In the first stage of the present study, the students in an English writing class at a prestigious university in Mainland China were the participants. The drafts they wrote and the feedback they received in two weeks were collected and analyzed. Secondly, stimulated recall data were collected when the students were asked to explicitly tell whether the revisions were influenced by the feedback given by AWE, the writing instructor, peers or other agents.

In the second stage, based on the quantitative analysis of students’ text revisions and qualitative analysis of stimulated recall data, some case study students were selected and interviewed to provide in-depth view of how the Integrated Feedback Approach activated and developed their autonomy in EFL writing. Preliminary results showed that the number of teacher-triggered revisions comprised the majority of the total revisions, and the students made more self-corrections than AWE-prompted revisions.
Roles and needs of learning-oriented language assessment

Jim Purpura, Teachers College Columbia, jp248@columbia.edu

Hansun Waring, Teachers College Columbia, hz30@columbia.edu

Liz Hamp-Lyons, CRELLA, University of Bedfordshire, liz.hamp-lyons@beds.ac.uk

Anthony Green, CRELLA, University of Bedfordshire, tony.green@beds.ac.uk

Kathryn Hill, University of Melbourne, kmhill@unimelb.edu.au

The principles and processes that are central to modern forms of formative assessment are typically absent or muted in assessment designed with summative purposes primarily in mind. In this project we have linked these principles and processes specifically to learning-oriented assessment (LOA) as elaborated by Carless, Joughin, Liu et al (2006) and an increasing body of work around LOA in language assessment. This symposium will comprise three linked presentations brought together by a moderator. It will present and discuss three perspectives on learning-oriented language assessment (LOLA; Green & Hamp-Lyons 2014), exploring the claim that assessment can serve as the ‘bridge’ between teaching and learning (Colby-Kelly and Turner 2007) through a particular focus on the contributing roles of the teacher/interlocutor and the learner(s) in facilitating effective language learning in learning-oriented language assessment events, be it in the form of large scale tests or the types of assessment which occur as part of routine classroom interactions (Hill, 2012; Purpura, 2014), and suggesting a way that this bridging may be supported through a consciousness of teachers’ assessment literacy needs.

The organizing questions for this symposium concern what the language interaction demands may be for learners to be effective in learning-oriented (language) and what assessment literacy demands are placed on language teachers if they are to move successfully into learning-oriented (language) assessment either in their own classrooms or in their roles as raters in the human rating of large-scale assessments. All learning events are situated in complex, social situations; when assessment becomes the dual and concurrent focus of learning the complexity of the learning-oriented language assessment (LOLA) event, taking place through a language interaction, becomes significantly more complex. The symposium as a whole looks closely at three specific yet varied assessment contexts and argues that whether assessment is informal and classroom-based or more formal and high stakes, whether the focus is linguistic or pragmatic, learners’ engagement in the dynamic process of learning to use the language effectively is essential; that teachers have a critical role to play in creating the conditions for learning; and that it is vital to understand and provide for teachers’ assessment literacy needs.

The first presentation works within a wide framework for thinking about learning-oriented assessment, focusing on one dimension in the framework to describe the interactional patterns present in a set of learning-oriented assessment episodes as teachers support learners in the classroom through the use of conversation analytic tools. The second presentation draws on a model of learning-oriented language assessment to describe the learning opportunities that might be present in the interactional possibilities of the speaking assessment of a large-scale language test. Complementing these two presentations, the third paper reports a taxonomy of what teachers actually do, and describes a self-assessment tool that helps teachers reflect on their own knowledge and needs for increased language assessment skills in the complex context of LO(L)A.
Examining the interactional patterns of feedback and assistance in spontaneous learning-oriented assessments and their effect on the dynamic processes of learning

Many assessment researchers (e.g., Turner, 2012; Hill & McNamara, 2012; Lantolf & Poehner, 2011; Purpura & Turner, 2012) have highlighted the central role that assessment plays in L2 classrooms and have expressed the need to relate assessment principles and practices to teaching and learning in L2 instructional contexts. In addressing this need, Turner and Purpura (2015) and Purpura and Turner (Forthcoming) have proposed an approach to classroom-based assessment, called learning-oriented assessment, in which learning and learning processes are prioritized while learners are attempting to narrow achievement gaps as a result of assessments. This approach focuses on how planned and unplanned assessments, embedded in instruction, are conceptualized and implemented from a learning perspective with relation to several interacting dimensions (i.e., the contextual, the proficiency, the elicitation, the interactional, the instructional, and the affective dimensions).

The current paper is concerned with the interactional dimension of LOA. It specifically examines the interactional patterns of feedback and assistance in unplanned, spontaneous assessments embedded in instruction, and then endeavors to describe how these patterns perform collaboratively in social interaction to contribute (or not) to the learners’ knowledge of and ability to use the passive voice to describe operational processes (e.g., desalination). One intermediate ESL class was videotaped from the beginning of a lesson to its culmination, when the results of an achievement test were returned (approximately 4 days). The video data were then uploaded onto Nvivo, transcribed, and examined through a conversation analysis approach. Instances of planned and unplanned assessments were then identified and examined iteratively and recursively through several lenses (e.g., interactional features, proficiency features, processing and learning features). This study shows a complex mix of patterns related to how spontaneous assessments, mediated through interaction, further the dynamic processes of learning to use the passive voice in these contexts.

Introducing opportunities for learning-oriented assessment to large-scale speaking tests

Drawing on recent developments in formative assessment and in broader concepts of learning-oriented assessment [LOA] (Carless 2003, 2007; Carless, Joughin, Liu et al. 2006; Joughin 2009), we describe a model of learning-oriented language assessment (LOLA) which we have applied to a set of speaking assessment standardization videos in order to understand whether and how LOLA opportunities might be inserted into a formal, large-scale exam. Further, we describe our findings about the ways in which official teacher support materials for this speaking test may, or may not, encourage the teachers who serve as speaking assessment interlocutors to put into practice aspects of LOLA through formative or dynamic assessment practices during the test event. The study suggests that putting LOLA into practice brings the personas of assessor and language teacher closer together, giving assessors the potential to implement teacherly skills, but that for this potential to become actual, some changes would need to be made to the current speaking test structure, the preparation materials, and the constraints imposed on interlocutors.

We have identified some opportunities for learning oriented practices to be inserted into the available resources for teachers, and for the strategies implied by the identification of these opportunities to also be included into teacher and learner preparation materials/resources. This study has shown that if effective LOLA is to occur in large-scale speaking tests, training of prospective interlocutors for the interlocutor role...
is critical, as is further development of the support materials for teachers and learners planning to take this test so that learners may take advantage of these new opportunities. With a view to advising the testing body on future development, a fully-fleshed programme of refresher training for interlocutors, and a programme of training materials development, trainer training, and ongoing observations of interlocutors in action are proposed. The resulting speaking assessment could be a model of the best in human engagement in live speaking performance assessment, but how far Lola insertions into the existing ethos and design of the speaking tests can in fact produce a more learner/learning-oriented approach to language test preparation will not be known until the attempt is made.

The implications of Lola for teacher professional development.

While the central focus in Learning Oriented (Language) Assessment (Turner & Purpura, forthcoming; Green & Hamp-Lyons, 2014) is on learners, learning processes and learning outcomes (Purpura & Turner, forthcoming p.6), teachers still have a critical role to play in creating the conditions for learning. However, teachers don’t necessarily have the skills and training to reflect in any systematic way on their own assessment practices, let alone on the theories of learning which underpin them or the contextual forces which shape them. This presentation will describe the development of a self-assessment tool designed to help language teachers to reflect on their language assessment literacy needs and develop their competence in Lola.

Development of the self-assessment tool was informed by an ethnographic study of what language teachers actually do (Hill, 2012), the principles of LOA (as enunciated by Turner & Purpura, forthcoming), existing taxonomies of teacher assessment practices and the literature on teacher assessment literacy as well as on good practice in classroom-based assessment more generally. It is based on a definition of classroom-based assessment (CBA) designed to account for the full spectrum of CBA practices, including the types of assessment occurring in the course of routine classroom interactions (Leung, 2005; Purpura, 2014). The self-assessment tool is designed to facilitate reflection on the nature of the assessments as well as the beliefs and understandings (about the subject, SLA, language teaching and assessment), which underpin them. It also explores the relationship between assessment and the relevant curriculum standards and frameworks and how prior knowledge, ability level, interest level and learning needs are taken into account. It emphasizes the learner perspective and agency in assessment and highlights the relationship between feedback, motivation and goal orientation. Furthermore, by including a specific focus on how practice is shaped by contextual factors at the local (classroom and institutional) level as well as the social and political level, the tool acknowledges the inevitable gap between policy and practice as well as the situated nature of CBA. Thus, it draws on Purpura and Turner’s (forthcoming) multiple dimensions to help teachers understand their own assessment literacy needs.
SYMPOSIUM 4

Thurs, Mar 19

4:45-6:15pm (Churchill Ballroom B)

New models and technologies for blurring the distinction between language testing and language learning

A. Jackson Stenner, MetaMetrics, jstenner@lexile.com
Barry O’Sullivan, British Council, barry.o’sullivan@britishcouncil.org
Jamie Dunlea, British Council, Jamie.Dunlea@britishcouncil.org
Todd Sandvik, MetaMetrics, tsandvik@lexile.com
Esther Geva, University of Toronto, esther.geva@utoronto.ca

There is a rapid convergence of new psychometric models (Stenner, Fisher, Stone and Burdick, 2013) and (Valsimer, Molenaar, Lyra and Chaudharg, 2009)); modern metrological thinking applied to education and behavioral science (Fisher, 2010); personalised learning platforms that blur the distinction between language testing and language learning (Hanlon et al., 2012); and cloud based delivery platforms that make these new technologies scalable and accessible around the globe. This symposium explores these intersections with a decided focus on illustrations and applications.

First, the symposium will introduce a Casual Rasch Model for language testing that represents a fusion of a substantive theory of text complexity with a Rasch Model. Major advances in educational measurement are evidenced by this approach and will be demonstrated from the context of individual-centered reading assessment. Second, the symposium will acknowledge critical issues in language testing when considering the perspective of the test-taker and explore how the concept of localisation may be applied to individualise feedback through a linking of complementary theoretical approaches to reading assessment. Third, the symposium will describe how the widely used descriptive labels of the Common European Framework of Reference (CEFR) can be combined with quantitative and qualitative measures to define reading text and task difficulty in a test of English proficiency. Efforts to integrate a Rasch-based measure of reading into specifications for an English as a Foreign Language (EFL) assessment will be examined in context for practical utility in language-test development and validation.

Finally, the symposium will explore ways in which the technical perspectives presented by the symposium can be integrated with and applied to day-to-day activities, which will be demonstrated by an online reading platform that fosters individualised learning. Special emphasis will be given to illustrating the cumulative benefits for educational systems moving toward a unified approach to reading measurement.

Causal Rasch models in language Testing: An application rich primer

Rasch’s unidimensional models for measurement are conjoint models that make it possible to put both texts and readers on the same scale. Causal Rasch Models (Stenner, Fisher, Stone and Burdick, 2013) for language testing fuse a theory of text complexity to a Rasch Model making it possible for a computer to response illustrate texts read by language learners during daily practice. Causal Rasch Models are doubly prescriptive. First, they are prescriptive as to data structure (e.s., non-intersecting item characteristic curves.) Second, they are prescriptive as to the requirements of a substantive theory for the attribute being measured i.e., item calibrations take values dictated by the substantive theory not the data. One consequence of this infusion of a Rasch Model with a substantive theory is that individual-
centered growth trajectories can be estimated for each reader even though no two readers ever read the same article or respond to a single common item. Rather than common items or common persons being the connective tissue that makes comparisons of readers possible, common theory is the connective tissue just as is true in, say, human temperature measurement where each person is paired with a unique thermometer. Thus, although the instrument is unique for each person on each occasion, a text complexity theory makes it possible to convert counts correct to a common reading ability metric in each and every application.

In this paper we will richly illustrate the above conceptual breakthroughs with applications. As an example, consider the implications for the measurement of a student responding to 10,000+ reader items per year embedded in her self-directed daily 40 minutes of reading practice. Additionally, we will illustrate how the text complexity theory can be tested by comparing the empirical text complexity measures based on reader responses with the theoretical text complexity measures produced by a computer based text analyzer.

Localising global assessment: Integrating the socio-cognitive model of test validation with applications of a causal Rasch model

When developing a language test, it is necessary to consider a range of elements. In order to ensure that these are systematically considered, it is necessary to conceptualise them in terms of a developmental model. This model, in turn should reflect a cohesive model of validity and validation. Within the context of the British Council’s English language tests, the model of validity considered most appropriate is a set of socio-cognitive validation frameworks proposed by Weir (2005) and updated by O’Sullivan & Weir (2011) and O’Sullivan (2011). The model sets out in detail the various elements of the test, from the perspectives of the test taker, the test task and the scoring system and has been successfully applied in a range of language test development and validation projects over the past decade (see O’Sullivan & Weir (2011) for an overview of these projects).

In considering the test-taker, two critical issues emerge. The first is the need to detail the model of language ability being tested, since the more explicit this model the more convincing any later validity argument. The British Council’s approach to testing reading, for example, involves writing tasks to reflect the underlying reading model and then accepting only those tasks which can be shown through their psychometric properties to meet the expectations of the model. The second issue is that of localisation (the appropriateness of the test for use with a specified population from a range of test-taker characteristics). The ultimate degree of localisation is the personalisation of the testing process. Though technology offers us innovative ways of reaching this goal, it is still some way off. However, one way in which personalisation of reading assessment is achievable is through the provision of individualised feedback. By linking the theoretical approaches of the British Council and Lexile we have begun to build a systematic understanding of how we can apply the concept of localisation/personalisation to a whole range of language learning and assessment initiatives. Offering personalised feedback based on reading test performance is just the first of what we see as a number of potential applications of the process.

Integrating measures of text difficulty and reading ability with a descriptive framework of language proficiency

The importance of comprehensive test specifications in test design is widely recognized. For tests of reading for English as a Foreign Language (EFL), specifications need to describe elements of both the input text as well as the cognitive processes test takers are
expected to engage in when completing the test task. When developing such specifications, we are now able to utilize a number of automated quantitative text analysis tools as well as a growing body of research on the relationship between the measures generated and empirical text difficulty. From a different perspective, descriptive frameworks of language proficiency have also seen rapid uptake over the last decade, with the Common European Framework of Reference (CEFR) increasingly used as a reference point for the description of language examinations. The CEFR is intuitively easy to apply through the use of can-do descriptors, which can be said to fulfill the function of Performance Level Descriptors in standard-setting terminology. At the same time, the illustrative scales of proficiency in the CEFR—always intended as user-oriented scales—lack the depth of description required for language-test development and validation.

This paper describes how a recently developed test of English proficiency has combined the widely used descriptive labels of the CEFR with both quantitative and qualitative measures to define reading text and task difficulty at different levels of the CEFR, and will provide empirical evidence of the relationship between these measures and task difficulty. The paper will then describe attempts to integrate the Lexile measure—a Rasch-based measure of text difficulty originally developed for use with native speakers of English in the United States and which places the reading ability of test takers and the reading difficulty of a text on a common metric—into the specification of the text. Test specification faces the challenge of teasing out a realistic and practical balance between comprehensive coverage and the use of a limited number of efficient and manageable measures, and this research aims to investigate the utility of using the Lexile measure in an EFL reading context to help achieve that balance.

Personalized learning: integrating testing and instruction in data-rich digital environments

The unfolding digital era is presenting unprecedented opportunities to enhance and personalise learning. The internet is bringing worldwide access to an ever-growing array of services, tools and content. Online learning activities can reach the spectrum of connected devices at any given hour around the globe. Cloud-based infrastructure is enabling rich data collection and imperceptibly fast data processing, which in turn is driving increasingly “intelligent” online experiences. More engagement with such technology over time can empower deeper personalisation with ever more sophistication. As a result, a simple activity such as silent reading—once a strictly passive task—can now be transformed into a highly interactive and responsive endeavor.

While general technology advances are bringing educational opportunities, the promise of personalised learning is dependent on the coherence and utility of measurements that fuel the system and empower the underlying analytics that drive user-centric features. The advances represented by Causal Rasch Models, applied within the context of an online platform that integrates testing and instruction with daily practice, presents a compelling example to consider. New analytic options open up when the data structure is small n (few to one person) but large P (many item responses observed over 5-8 years.) These new analytic methods resemble time series methods and can answer questions like the effects of summer vacation or daily practice on an individual student growth trajectory. These methods address the valid criticism that in general the study of inter individual variation to infer anything about intra individual variation is fraught with complications.

In this paper, we will situate key perspectives of this symposium within the dynamic of an educational system and add to this model the
further dimension of personalized, day-to-day learning. This addition will be exhibited in the form of a digital platform that harnesses trends in technology and advances in educational measurement. Exhibits using student data from around the world will be used to demonstrate the power of these new approaches in personalisation. Lastly, we will update our illustrative model of an educational system to show how common measures can flow freely throughout as a valuable new form of currency.

SYMPOSIUM 5
Fri, Mar 20
3:00-4:30pm (Mountbatten A)

Vocabulary in assessment: what do we mean and what do we assess?

Rob Schoonen, University of Amsterdam, rob.schoonen@uva.nl

Norbert Schmitt, University of Nottingham, Norbert.Schmitt@nottingham.ac.uk

Benjamin Kremmel, University of Innsbruck, Benjamin.Kremmel@uibk.ac.at

Scott Jarvis, Ohio University, jarvis@ohio.edu

Norman Segalowitz, Concordia University, Montreal, norman.segalowitz@concordia.ca

In many studies of language proficiency, vocabulary knowledge or vocabulary use plays a prominent role, and therefore the assessment of vocabulary is often a crucial part of these studies. The central role of vocabulary should not be surprising, because models of language processing/use always include a lexical-semantic component. However, the part vocabulary plays in language processing/use (be it reading, writing, speaking or listening) can vary considerably. For example, the requirements for adequate vocabulary knowledge in reading will be very different from the vocabulary skills needed in speaking. Nevertheless, vocabulary assessment is usually considered to be quite straightforward. Unfortunately, this sometimes leads to forms of vocabulary assessment that are misguided, or at least oversimplified.

Fifteen years after Read’s monograph Vocabulary Assessment, it is time to reconsider different types of vocabulary assessment and to investigate their validity for measuring different aspects of learners’ vocabulary knowledge. Based on cognitive models of language processing, we...
also have to question whether all relevant aspects of lexical-semantic processing are covered by the traditional measures of vocabulary knowledge.

This symposium aims to bridge the gap between cognitive models of vocabulary in language use, and various measures of vocabulary. Three complementary perspectives will be taken. First, vocabulary size will be addressed. Using vocabulary size in research or teaching suggests that we have a clear understanding of when a word is known. However, as the first presentation shows, different formats of vocabulary size assessment lead to different outcomes which potentially have consequences for language teaching. The second perspective concerns how well language learners employ the vocabulary knowledge they have, especially in language production. How do (traditional) measures of the lexical variability in language use (based on types and tokens) relate to language proficiency? Furthermore, do they allow valid inferences about language proficiency? Some alternative approaches will be suggested. The third perspective concerns how language learners access their vocabulary knowledge in online language use. One may know a lot of words, but still have problems in accessing that knowledge in conversation or reading/writing. Slow lexical-semantic processing clearly affects language performance. The questions become, then, to what extent do our vocabulary assessments take this accessibility aspect into account, and what are the best ways of measuring fluency in use?

Addressing these issues is an essential part of building validity arguments for the vocabulary measures that are so frequently used in language research and assessment. It may also shed light on new directions of vocabulary assessment. Additionally, valid assessment can inform test users about opportunities for learning and teaching.

The more, the merrier? Issues in measuring vocabulary size

Vocabulary tests are a key tool in the assessment repertoire of classroom teachers. They hold great potential for integrating learning and assessment as they could be used for diagnosis and identification of lexical strengths and weaknesses and tailoring instruction to the learners’ needs. Several tests have been developed and suggested for pedagogic use, such as the Vocabulary Levels Test (VLT) (Nation, 1990; Schmitt et al., 2001) or the Vocabulary Size Test (Nation & Beglar, 2007). Most of these have been commonly labelled measures of vocabulary size, suggesting that this is a well-defined construct, useful for pedagogic practitioners in their daily work with language learners. However, the interpretation of scores generated by tests of vocabulary size is not as straightforward as it seems. Size scores of learners are likely to be greater if only little depth of word knowledge is tested. Vice versa, learners will know less vocabulary to a greater depth. The paper will therefore argue that size figures need to be interpreted in conjunction with the degree of knowledge tapped into in order for this size estimate information to be useful for teachers and learners in pedagogic settings. The item format thus plays a crucial role as it directly affects size measurement. However, there has been little research into what different size test item formats show about depth of knowledge. This paper will report on a study that subjected four commonly-used item formats to empirical investigation. The informativeness of two recognition formats (VLT and 4-option multiple choice) and two recall formats (with and without a context sentence) was compared by administering 36 vocabulary items in the different formats in a Latin square design to native and non-native speakers of English. The word knowledge of the participants was afterwards verified in face-to-face interviews and written meaning recall measures, probing whether their scores derived from guessing, partial knowledge, or mastery of different word knowledge aspects of the target items. The results provide insights into the workings of the different
test formats, and will help both teachers and researchers interpret the resulting size figures stemming from vocabulary tests using these frequently-used formats.

**Variability in vocabulary use**

The diversity, variety, and variability of learners’ vocabulary use have been found in numerous studies to serve as useful and sometimes even powerful indicators of learners’ vocabulary knowledge and the quality of their language performance. The tools available for assessing lexical variability differ widely, but nearly all of them directly or indirectly evaluate the relationship between the abundance of lexical types and tokens in samples of language use, and are therefore fundamentally measures of either lemma or lexeme repetition. Although repetition is an important dimension of variability, a severe shortcoming of automated repetition measures is that they confound useful word repetitions with redundant word repetitions, and it is only the latter that reflect deficiencies in learners’ language abilities. A more promising approach to measuring lexical variability through automated means involves taking into consideration multiple other dimensions of variability in learners’ vocabulary use that affect the quality of their language performance. Such dimensions include lexical evenness, disparity, dispersion, and specialness. The purpose of this paper is to (a) describe what these dimensions are, (b) show how they affect human judgments of lexical diversity, (c) discuss useful methods of measuring these dimensions, and (d) show how such measures both individually and collectively correlate with ratings of language proficiency and writing quality.

**Fluency of vocabulary access and the assessment of L2 vocabulary**

When it comes to assessing the nature and extent of a person’s second language vocabulary knowledge the focus is often on the breadth and depth of that vocabulary, with distinctions made between receptive (or passive) vocabulary and productive (or active) vocabulary. Fluency, or the speed and ease with which a person can access and use his or her vocabulary, is also sometimes added as yet another dimension of vocabulary to assess. Fluency of vocabulary access, however, poses an interesting challenge. This is because measures of breadth and depth of vocabulary knowledge necessarily require people to access their vocabulary knowledge in order to fulfill the task demands of the assessment procedure. These measures, assumed to be distinguishable from fluency measures, may nevertheless implicitly reflect fluency of vocabulary access. It is important, therefore, to ask whether we should consider fluency of vocabulary access to be simply just another dimension of vocabulary knowledge (alongside breadth and depth) or to be a feature that inherently underlies all other measures of vocabulary. The presentation will address this question by considering a number of ways one can operationalize fluency of vocabulary access and how each of these may be related to other measures of vocabulary knowledge. Several different ways of thinking about fluency of vocabulary access will be considered, including speed of lexical access, efficiency of lexical access (automaticity), connectivity of vocabulary networks, processing implications of the so-called semantic/episodic memory systems distinction, the role of frequency versus context diversity in word retrieval, and retrieval phenomena related to transfer appropriate processing. These are discussed with the aim of formulating an approach to the assessment of vocabulary that integrates vocabulary knowledge and vocabulary accessibility.
The evaluation of school-based EFL teaching and learning in China

Zunmin Wu, Beijing Normal University, wzm@bnu.edu.cn
Luxia Qi, Guangdong University of Foreign Studies, qiluxia@hotmail.com
Baocheng Han, Beijing Foreign Studies University, hanbaocheng@bfsu.edu.cn
Hongyun Liu, Beijing Normal University, hyliu@bnu.edu.cn
Na Wei, Ministry of Education, wnlucy@yahoo.com.cn

With China’s curriculum reform in basic education having been in progress now for over a decade, it has become important for China to undertake a formal assessment of education quality that can provide information necessary both for informing educational policy-making and for curriculum development. The proposed symposium will report on a national assessment project in China designed to monitor the quality of the country’s basic education system.

The National Assessment of Education Quality (NAEQ) project was officially launched in 2009 by the Center for the National Assessment of Educational Quality, affiliated to the country’s Ministry of Education. NAEQ has set Grade 4 (10-year-olds) and Grade 8 (14-year-olds) as the target assessment groups, drawing on the experience of other international educational evaluation programs, such as NAEP, PISA and TIMMS. The assessment covers the major school subjects, such as Chinese, mathematics, science and English language education. It also includes the assessment of students’ well-being in schools and factors affecting students’ academic achievement, using questionnaire surveys of students, teachers and school principals. For example, in 2011 alone, NAEQ assessed 50 thousand students in 11 Chinese provinces (out of a total of 31). The results were reported primarily to the Ministry of Education but also to the provincial or regional educational administrative officials and teachers.

The proposed symposium links directly to the LTRC 2015 conference theme of the relationship between teaching, learning and assessment as it examines the following four areas: 1) English assessment in the NAEQ program and the contribution it makes to the country’s EFL teaching; 2) definition of the construct underlying the pen-and-paper test; 3) a baseline study of the test consequences of NAEQ English assessment; and 4) the statistical work supporting the system, including factors such as the sampling of test candidates, test score management, standards setting, and assessment results interpretation.

The most important issue has been the decision on test construct which set the basic principles for assessment development. The study of assessment consequences started along with the progress of the assessment program and is the first of its kind undertaken in China. The researchers went into classrooms and interviewed teachers in order to collect data for analysis. Study results provide valuable information on how a large-scale test, despite being a low-stakes one, can affect various stakeholders. The highly professional statistical support is a crucial part of the program, with its professionals coming from a psycho-educational research background. Take the sampling decisions, for example. They have to take into consideration a wide range of variables, relating to location, school and students. Thus the symposium connects to several of the LTRC 2015 subthemes, including the washback, impact and consequences of large-scale language testing.

It is hoped that the report of the NAEQ program and the pattern of its working mechanism can shed light on the process of assessing educational development at a national level and will make some contribution to the
broader field of language testing and assessment.

Assessing EFL learning in schools and factors affecting learning outcomes

The assessment of school students’ EFL learning is a part of the general scheme of the Center for the National Assessment of Education Quality (http://www.eachina.org.cn/eac/zxjj.htm 2006), whose mission is to establish an assessment system that monitors school learning outcomes and their well-being in order for the educational administration at various levels to make informed policy decisions. At the same time, assessment results impact on stakeholders beyond the school. English language learning is assessed according to the learning objectives specified in the National Curriculum Standards (MOE 2001) and it involves students at two school levels: Grade 4 and Grade 8. The Grade 4 students should have learned English for about 120 class hours and the Grade 8 students for more than 500. The assessment consists of a pen-and-paper test of English and questionnaire surveys of students, teachers and school principals. Taking into consideration the cognitive ability of young learners (11 years old), the Grade 4 test consists of only 20 items, divided between listening comprehension and reading comprehension. The Grade 8 test consists of 56 items measuring listening comprehension, reading comprehension, use of language knowledge and writing. The speaking test is designed for both groups.

The first English assessment was done in 2011 with approximately 50,000 students sampled from 11 provinces taking the pen-and-paper test and 900 students from 5 of these provinces doing the speaking test. Assessment results revealed that the proficiency level of the assessed population ranged widely from lower A1 to lower A2 CEFR levels. Key factors that contribute to these results are limited exposure to the target language and the varied quality of instruction received by students across the nation. It was also found that factors such as family economic situations, parental care, teacher-student relationship and sleep time affect students’ academic achievement. It was quite alarming to discover that in some rural areas Grade 4 students did not have English classes due to lack of teaching staff. The fact that when English is taught right from Grade 1 in some schools, students tend to confuse the English sound system and the Chinese pinyin is also cause for concern for curriculum designers and school teachers.

Construct definition of the English assessment in the NAEQ system

The purpose of this presentation is to report the construct definition and operationalization of the English assessment in the NAEQ system. Based on the National Curriculum Standards (2001), the pen-and-paper test aims to measure students’ ability to do things with English. However, it is not an easy job to define and operationalize the term “ability” for the test development because its meaning is equivalent to that of “competence”, “proficiency” or “skill” in the Chinese context. After examining the different uses of these terms and their inadequacies, the speaker, drawing on the latest theory on language use and research on the taxonomy of educational objectives, proposes a tentative thinking-based language ability framework for developing and using language tests. Within this framework language ability is defined as the ability of comprehending and expressing information that learners exhibit when they apply their knowledge of language and the knowledge of the world, and the strategies to perform language use activities in a variety of contexts. Language comprehension and expression abilities can be further classified into different subcategories. This thinking-based language ability framework was adopted in the development of the English assessment in the NAEQ system. The presentation concludes that language tests developed on this framework are in favour of the learners’ thinking skills development via language learning and it also provides the basis both for measuring learning and advancing learning.
A baseline study of the washback of the English assessment in the NAEQ system

This presentation reports on part of the results of a baseline washback study of the English assessment in the National Assessment of Education Quality (NAEQ) in China. Well aware of negative washback induced by high-stakes tests on teaching and learning, the designers of the NAEQ English assessment have tried to make it “low-stakes” in the hope of preventing negative washback. At the same time, they desire the test to facilitate learning if washback cannot be avoided. Will the test help to constrain negative washback and allow intended positive washback to occur? With this question in mind, we carried out a baseline study of washback, focusing on the nature of test stakes in light of dynamic system theory (DST), the intended washback by the test designers, and the status quo of English teaching and learning in the classroom. This is because the stakes of a test is the origin of washback, and a good understanding of the intended washback and classroom teaching and learning before the implementation of the assessment should serve as a point of comparison for a long-term washback study to be conducted later. The main findings concerning the stakes of the test and the intended washback are as follows. First, the stakes of any test can be dynamic, changing as the test system interacts with other systems and with elements in the context. Low-stakes tests can become high-stakes and vice versa. Second, the designers of the NAEQ have successfully kept it a low-stakes system and teachers and students do not know much about the NAEQ English test but have a positive attitude toward it at the moment. Third, the designers’ intended washback is that teachers will help students develop the ability to use English to communicate as they prepare for and interact with the test. Based on these findings, we suggest that the washback study should go on with the continuous assessment of the NAEQ to provide feedback to policy makers who can then take measures to help limit potential negative washback and encourage positive effects.

Sampling and scaling processes for the national assessment of education quality in China

Sampling students properly is of great importance to obtain quality data for analysis. This is because variables can be factors that affect the quality of education. Once data are collected, scaling participants’ performances in the English test can give a better picture of the quality of students’ learning outcomes. It enables the comparison between regions, school districts and schools so as to interpret what might be the reasons for the differences. The aim of this presentation is to explain the major sampling processes and scaling methods used in surveys conducted for the National Assessment of Education Quality.

The sampling precision requirement for English language assessment was a confidence interval of no more than ±0.1 standard deviation units at the 95 percent confidence level. The sampling design followed a stratified three-stage process. Firstly, regions (participating provinces, cities or school districts) were sampled according to probabilities proportional to student size (PPS) from the list of all regions in the national population containing eligible students, with provinces as an explicit stratification variable, and the proportion of urban students as an implicit stratification variable. For the second sampling stage, schools were sampled using PPS from the sampled regions of the first stage. In the third sampling stage, students were sampled using simple random sampling.

Student achievement was reported through scale scores derived from Item Response Theory (IRT) scaling. This approach allows the performance of a sample of students in a subject area to be assessed on a common scale when different students have been administered different items. The scaling model used in the National Assessment of Education Quality was the multidimensional random coefficients logit model (Adams, Wilson, and Wang 1997).

The presentation will describe, as an example, the sampling process for the 2011 assessment.
where 13,421 students were sampled from 446 schools of 140 regions which provided a guarantee that the assessed population would be representative of the Eighth Graders in general for that year. It will also describe the process of scaling of student scores and how scaled scores can be used to interpret education quality in relation to the various factors.

SYMPOSIUM 7
Fri, Mar 20
3:00-4:30pm (Rossetti)
Applications of automated scoring tools for student feedback and learning

Sara Weigle, Georgia State University, sweigle@gsu.edu
Paul Deane, Educational Testing Service, pdeane@ets.org
Ardeshir Geranpayeh, Cambridge English Language Assessment, Geranpayeh.A@cambridgeenglish.org
Masanori Suzuki, Pearson, masanori.suzuki@pearson.com
Lawrence Davis, Educational Testing Service, ldavis@ets.org
April Ginther, Purdue University, aginther@purdue.edu

Recent advances in natural language processing and other technologies have led to new developments in automated scoring and feedback systems to support language assessment and learning (Xi, 2010). Major international testing companies have made substantial investment in automated scoring and feedback systems, as evidenced by a special issue of Language Testing devoted to automated scoring and the recently published Handbook on Automated Essay Evaluation (Shermis & Burstein, 2013). While several recent publications have addressed the validity of automated scoring systems for language tests, an equally important potential use of these systems is to provide immediate, useful feedback to students and their teachers to support classroom learning. In this symposium, representatives from three major international test development organizations...
Combining product and process data: Exploring the potential for automated feedback in writing

This paper explores the potential for combining automated scoring and process analysis to provide feedback about student writing. Automated essay scoring (AES) systems typically use features that directly measure surface characteristics, such as sophistication of vocabulary, errors in grammar/usage/mechanics/style, and various indicators of organization and development. Attali and Powers (2008) argue, on the basis of a factor analysis, that such features reflect three underlying factors – fluency, accuracy, and vocabulary – which jointly characterize text production skills (critical for all writing, and all that novice writers use when they employ what Bereiter & Scardamalia, 1988, term a ‘knowledge telling’ strategy, in which they write from memory everything they know about a subject with little attention to addressing rhetorical or conceptual problems.) AES systems score the final written product, so they provide no information about planning, revision, or editing. On the other hand, these activities typically involve interruptions to the normal flow of text production, and as such, should be identifiable from a keystroke log. It should therefore be possible to gain a richer picture of student performance by examining a profile that combines AES product features with process features derived from a keystroke log. The paper analyzes data from an 8th grade timed writing assessment, exploring relationships between product features, process features (i.e., keystroke logs), and student scores. Paper #2 reports on the use of an online system used to provide diagnostic feedback within a learning-oriented, self-access, and/or teacher-directed assessment context. Paper #3 presents the current capabilities of a fully automated writing system that is used in both low-stakes, classroom-based assessments, and high-stakes assessments, describing how such automated writing systems are developed, and presenting validation evidence. Paper #4 presents research on the use of automated scoring of unconstrained speech, describing an initial exploration into making use of detailed speech data by including it in an augmented score report for a practice test of speaking that elicits unconstrained speech.

Learning-oriented assessment feedback: The case for automated assessment as evidence of learners’ progress

A major challenge for the past 25 years has been the need to individualize learning in classrooms - and in using the tools now at our disposal, to focus on learning and growth over time. This in Learning Oriented Assessment (LOA) often means using e-assessment during a course
of language learning. E-assessments allow learners to receive individualized feedback based on their performance, with recommendations directing them to remedial, development or extension exercises, depending on their test scores. As the learning program progresses, more classroom test data is collected and an electronic student profile emerges, and is saved in the form of an electronic portfolio. Formal e-records of achievement can be maintained by individual learners or for the whole class. This in turn has the potential to inform teacher planning, feeding back into the teaching cycle.

In this paper we report on the use of an online system used to provide diagnostic feedback within a learning-oriented, self-access, and/or teacher-directed assessment context, and address issues of construct validity, as well as of assessment use and impact. The system is an automated writing assessment tool which maps learners’ output to proficiency levels defined by external benchmarks and frameworks of language ability. Implicit in this mapping is the identification of positive and negative writing features related to topic relevance, organization and structure, language and style. We consider these features in terms of the degree to which they cover models and constructs of writing ability. These features can be weighted in different ways to maximize their predictive power within different language use contexts and for different L1 backgrounds. Learners can therefore receive overall and specific feedback that diagnoses the quality of their writing according to context, increasing construct validity.

Within a learning-oriented assessment context learners can, perhaps assisted by their teacher, repeatedly access the system and work on improving various aspects of their writing, referencing their earlier work as desired. Using questionnaires to collect feedback from teachers and learners, we show that the system can be a useful supplement to other modes of learning that can help to promote learners’ writing development.

Use of automated writing scoring system in classroom and high-stakes contexts.

Whether the use of a language test is for a low-stakes testing or high-stakes testing, the test-taker’s learning should ideally be enhanced as a result of the test event. One way to promote such learning is by providing feedback as quickly as possible while the test-taker’s experience with the test is still relatively fresh in memory. The importance of immediate feedback is a similar concept to the Output Hypothesis (Swain, 1993, 1995) in second language acquisition. In the context of language assessment, if the test-taker notices the gap between the task demands and the test-taker’s own ability to complete them during test-taking (i.e., output process), receiving scores immediately could help the test-taker plan how he/she can narrow the experienced gap. Therefore, automated scoring systems, in general, have the potential to help enhance the test-taker’s learning.

In this paper, the authors present the current capabilities of an automated writing system that is used in both low-stakes, classroom-based assessments, and high-stakes assessments. In the context of classroom-based assessment, the test-taker writes an essay in response to a prompt or reads a passage first, and then writes a summary of the passage. Once the learner’s writing sample is submitted, it is evaluated by the automated writing system immediately on traits such as content, organization, word choice, and sentence fluency. The learner then has opportunities to improve their writing based on the feedback and to submit a revised text for another round of immediate feedback.

Such an automated writing scoring system has also been embedded in a few different high-stakes tests. Written responses such as a passage reconstruction of a text the test-taker has read, and an email-style writing, have been scored on a few traits including vocabulary, organization, and voice and tone. Sufficiently high correlation coefficients (above r=0.90) between human scores
and machine scores have been demonstrated in validation studies.

The paper describes how such automated writing systems are developed, and presents validation evidence. The authors also discuss how this automated scoring system differs from other automated essay evaluation systems, because of the natural language processing technique called latent semantic analysis.

**Use of automated scoring technology to provide feedback on unconstrained speech**

Providing language learners with specific and accurate feedback on speaking ability is a challenging task for humans, but in theory could be a relatively easy task for machines. While a variety of automated systems have been devised to provide feedback on speaking performance, such systems generally rely on highly constrained speech elicited by tasks such as reading aloud. Automated systems for scoring unconstrained speech have been developed, but such systems are typically trained to produce holistic evaluations of test taker's language performance, with the purpose of approximating holistic scores awarded by human raters. Although such holistic scoring systems may incorporate a variety of language features, these features are combined into a single number that essentially discards the details of how the test taker performed. This presentation describes an initial exploration into making greater use of such detailed data by including it in an augmented score report for a practice test of speaking that elicits unconstrained speech. A draft version of the score report will be described, along with the considerations that went into selecting the information to be reported and the manner in which feedback was presented.

A holistic score reported in the current test results were supplemented with additional information including (1) sub-scores for two language domains that are mentioned in the scoring rubric for the test, and (2) measurements of individual language features obtained directly from the scoring engine. Domain-specific sub-scores were obtained from scoring models trained to predict human sub-scores for the language domains of delivery (fluency and pronunciation) and language use (vocabulary and grammar). Measurements of individual language features were selected based on the extent to which the measurement: (1) provided actionable information, (2) was understandable to test takers, and (3) discriminated between differing levels of performance. User perceptions of the clarity and usefulness of the expanded score report were also collected in a small-scale usability study and will be described. The presentation will conclude with a summary of design considerations for presenting automated feedback on unconstrained speech, as well as the current limitations of automated scoring systems for this purpose.
1. Assessing Chinese teaching assistants in the US

**Dingding Jia**, The Pennsylvania State University, duj144@psu.edu

International teaching assistants’ communicative competence is one of the most trending topics in studies of advanced second language speakers. Although many studies up to this point focused on the linguistic features of international teaching assistants, most of them were related to pronunciation and lexical variety. Few studies focused on their use of discourse markers. Discourse markers are not included in either international students’ previous English education in their home courtiers or assessment in the US, but they are truly important in classroom interactions. Discourse markers are important linguistic features because they may display the teacher’s stance and build rapport between the teacher and the students. This work examines the similarities and differences among Chinese teaching assistants at a large public American university. The first research question investigates the similarities and differences among Chinese teaching assistants in terms of the frequency of discourse markers. The second research question examines similarities and differences among Chinese teaching assistants with respect to the functions of discourse markers. The theoretical orientation of this research is guided by Schiffrin’s study (1988) of discourse markers. In particular, this study analyzed several of the most commonly used discourse markers, such as I mean, uh, so, actually, and okay. Chinese teaching assistants who are doctoral students in science, technology, engineering, and mathematics participated in the study. Their lectures were transcribed using conversation analysis conventions. Conversation analysis approach was used because it allowed the researcher to identify the functions of discourse markers. Results showed that in terms of quantity, so, okay and uh were the most frequently used discourse markers. With respect to functions, the Chinese teaching assistants primarily used discourse markers to manage their speech, but they did not use discourse markers very frequently to display attitudes or feelings. This study hopes to shed light on assessing, teaching, and training of international teaching assistants.

2. Development of WIDA early years language assessments: From assessment to instruction

**Carsten Wilmes**, WIDA, wilmes@wisc.edu

**Ahyoung Alicia Kim**, WIDA, akim49@wisc.edu

The WIDA Early Years suite of language assessments, designed for ages 2.5-5.5, aims to not only screen and identify incoming DLLs to early care and education programs, but also observe their language development over an extended period of time in everyday settings. Therefore, the assessments will have a strong formative characteristic. Due to the unique nature of early learners, a wide range of information will be collected from multiple sources, including parents/caregivers and practitioners. Detailed information obtained using the suite of assessments could be further used by practitioners for providing appropriate language support and instruction to DLLs. This poster describes the current work by WIDA to develop a suite of assessment tools to help practitioners understand, support, and monitor DLLs’ progress in language development over time. In both design and mode of administration, these assessment tools take into account that children, age 2.5 – 5.5 years, primarily learn language through the context of important relationships with significant caregivers during daily routines and play-based learning activities. The assessment
tools are designed to be used within a variety of early care and education programs and can be easily incorporated into existing routines and learning activities. In alignment with K-12 WIDA assessments, the WIDA Early Years Language Assessments will be standards-based and reflect the WIDA Early Language Development Standards. In addition, the assessments are to be used in conjunction with states' Early Learning Standards. Therefore, the assessments will measure the language of six standards: language of social and emotional development; language of early language development and literacy; language of mathematics; language of science; language of social studies; and language of physical development. In addition, the assessments will measure the two language domains—receptive and expressive—in three developmental levels—Level 1 (Entering), Level 3 (Developing), and Level 5 (Bridging).

3. Weighing the consequences of introducing an international large-scale English language test in Cuba

Jan Van Maele, KU Leuven (University of Leuven), jan.vanmaele@kuleuven.be

Roberquis Rodríguez González, Universidad de Oriente, translator-1@vlir.uo.edu.cu

Yoennis Díaz Moreno, Universidad de Oriente, translator-3@vlir.uo.edu.cu

Frank van Splunder, University of Antwerp, frank.vansplunder@uantwerpen.be

Lut Baten, KU Leuven (University of Leuven), lut.baten@ilt.kuleuven.be

The ILTA Code of Ethics admonishes that language testing professionals “should engage in a thorough evaluation of the likely consequences” of their projects (ILTA, 2000). Following Bachman & Palmer (2010), the “beneficence” of the intended consequences of a test and the “values sensitivity” of the decisions made on the basis of an assessment constitute two essential components of any assessment use argument. In this poster we investigate what these components could mean in view of the possible introduction of an internationally recognized and commercially available large-scale English language test in Cuba. Bearing in mind how new a phenomenon nationwide foreign language certification is in the country, the general question that we seek to answer is: What should and what can be done in order to maximize the chance that introducing the test would bring about good in Cuban society taking account of the (evolving) educational, sociopolitical, and legal environment? The authors are academics who have collaborated in long-term institutional university cooperation and network projects of the Flemish Interuniversity Council (Belgium) with universities in Cuba. In the Council’s country strategy for Cuba, English language learning and communication is listed as one of three transversal themes for all current and future interventions (http://www.vliruos.be/en/countries/countrydetail/cuba_3850/). Based on their experiences in a previous ten-year project in central Cuba, the authors concluded in 2011 that there was a need for establishing a regional testing center, noting the following key issues: (a) intercultural citizenship for Cuban academics, (b) sustainable local institutional development, (c) validity of administered tests, and (d) privacy and confidentiality issues. This need was reiterated in the external assessors’ final evaluation report (Carpenter & Vigil Tarquechel, 2013). Since then initiatives have been started up to establish testing centers in the context of a new project in eastern Cuba in view of the stated project goal of strengthening foreign language skills for intercultural and international academic purposes. It is expected that the introduction of an internationally recognized test at regional testing centers could contribute to increased opportunities for Cuban academics to study and work abroad and boost student motivation while providing an adequate answer to the previously identified key issues. At the same time, a number of concerns have been raised related to accessibility, administration, impact on instructional practice, and the question of how to
address some of the more elusive competences than the ones commonly captured in standardized testing, particularly the intercultural competences required for a successful academic mobility experience. The poster will report an initial answer to the general research question based on a stakeholder mapping exercise (identifying, analyzing, mapping, prioritizing) at project, institutional, and regional level that was conducted during the first two project years. Given that outward mobility for Cuban academics in the project concerned is almost exclusively Europe-bound, perspectives of European policymakers and students on academic mobility will also be included (Beaven, Borghetti, Van Maele & Vassilicos, 2013).

4. Investigating the effectiveness of formative assessment for an integrated reading-to-write task

Hyun Jung Kim, Hankuk University of Foreign Studies, hjk2104@gmail.com

Ahyoung Alicia Kim, WIDA, alicia.a.kim@gmail.com

In an academic setting, novice members, especially L2 writers, need to become aware of the expectation of the academic writing community. Despite the need for attention to these L2 newcomers, little has been researched on the development of their academic writing ability through continuous formative assessment. Furthermore, few studies have utilized reading-to-write tasks, which are one of the most prevalent task types required in the academic field. Therefore, the purpose of the current study was to investigate the effectiveness of repeated formative assessment on an integrated reading-to-write task at the graduate level. This qualitative case study examined ten TESOL graduate students at a Korean university, all of whom were in the beginning stage of preparing their master’s thesis. While taking a research methods course, every week for a total of six weeks, each student first read a journal article, selected upon each student’s interest. Then they wrote a summary and critique of the article as part of their classroom assessment. The instructor evaluated each summary/critique in terms of four dimensions of content, language, organization, and mechanics, and provided the individual students with detailed written feedback on the four dimensions. After the six week period, the instructor interviewed each student to understand their perceptions about the effectiveness of written feedback for the development of their academic writing ability. Qualitative analysis indicated that the students had diverse developmental patterns across the four dimensions. Students demonstrated quick improvement in the coherence and cohesion pertaining to the organization of their writing. However, they struggled to include appropriate content for summary and meaningful critique. Interview findings indicated that the students’ weakness in content was partly due to the nature of the integrated reading-to-write task. Overall, repeated writing practices and subsequent formative assessments helped the students improve reading skills as well as academic writing ability. The results provide pedagogical implications for the use of sustained formative feedback for L2 learners’ development of academic reading-to-write skills.

5. Linking learning to assessment: A critique and analysis of Duolingo’s English test

Elvis Wagner, Temple University, elviswag@temple.edu

Antony John Kunnan, Nanyang Technological University, akunnan@gmail.com

Recently, the Duolingo English test received a huge amount of publicity when it announced a 20-minute English language proficiency test that can be taken anywhere and by anyone with a smartphone and $20. Duolingo’s head of marketing referred to the language testing industry as “archaic” and “pretty much
a monopoly”, and stated that the Duolingo test “could produce equally representative results of a student or job seeker’s English proficiency” as the IELTS and the iBTOEFL. Is the Duolingo English test the comet that will wipe out our language testing dinosaurs? Or is this simply the case of a company with a large marketing budget trying to make a splash by ridiculing the (language testing) establishment? This work in progress tries to answer those questions by providing a critical overview of the Duolingo English test, the test tasks, how the test links to learning, and an exploration of possible washback effects. The presentation is directly related to the theme of the conference, in that we examine how Duolingo’s English test is connected to their teaching/learning program. Because the Duolingo test is very new, there is almost no published research on it (i.e., Vesselino & Grego, 2012), and a validity argument has not been made for this assessment. Therefore, this critique uses a number of conceptual models, including test useffulness (Bachman & Palmer, 1996), test fairness (Kunnan, 2000), and consequential validity (Messick, 1989, 1994), to critique the test in terms of its supposed reliability, practicality, construct and consequential validity, and fairness. Beyond this conventional critique, however, we employ Shohamy’s (2001) model of critical language testing to examine the Duolingo English test’s potential to have a disruptive influence (both positive and negative) on the language testing industry, and the possible washback effects (both positive and negative) of the test becoming widely used. The examination of its washback effects will focus not only on test-takers themselves, but also on other test users, including universities that might use the Duolingo for admissions purposes, and teaching and learning systems that exist in many countries to prepare learners for North American English-medium university admissions. This is a work in progress in that while our critique of the test will be complete by the time of the conference, we also are interesting in conducting an empirical study to collect evidence for the validity, reliability, fairness, access, and consequences of the test in order to examine its overall beneficial nature to the community.

6. Exploring the relationships among teacher feedback, students’ self-efficacy, self-regulatory control, and academic achievement in Chinese EFL context

Yongfei Wu, Queen’s University, yongfei.wu@queensu.ca

This study aims to explore the relationships among teacher feedback, students’ self-efficacy, self-regulatory control, and their academic achievement in Chinese EFL context. The existing research on classroom assessment has been largely practical, i.e., seeking successful assessment practices, but not theoretical (Brookhart, 2013). Without theoretical guidance, classroom assessment practices tend to be episodic and dispersed (Allal and Lopez, 2005), which may make our understanding of classroom assessment fragmented and inadequately explicated (Torrance, 1993). The research on feedback, the central aspect of classroom assessment, needs “considerably better theoretical grounding” (Wiliam, 2013, p. 207) so that it may play a powerful role in promoting students’ learning. Theories of self-regulation have been discussed in depth to develop and solidify the theory of classroom assessment, especially feedback (see Andrade, 2013; Nicol & Macfarlane-Dick, 2006; Wiliam, 2013). Yet empirical studies are rare that examine the relationships among the key aspects of feedback, self-regulation, and achievement. This study adopts a sequential explanatory mixed-method design with a quantitative and a qualitative phase. In the quantitative stage, a questionnaire will be designed to measure how students perceive teacher feedback, their self-efficacy for English learning, and their self-regulatory control. The survey will be conducted among approximately 1,000 EFL students at a comprehensive university in China. Students’ academic achievement data will be collected from the test scores in their English final exams. Descriptive statistics, correlation,
and structural equation modeling will be used to analyze the collected data to examine the interrelationships among feedback variables, self-efficacy variables, self-regulatory control variables, and academic achievements. In the qualitative stage, individual interviews (n=12) and focus groups (n=20) with students sampled from the quantitative phase will be conducted to compliment the relationships detected in the quantitative stage. A standard thematic coding process will be used and the co-occurrence of the codes will be identified to explore the relationships among feedback, self-efficacy, self-regulatory control, and academic achievement. The findings of this study potentially have both theoretical and practical implications. First, this study is interdisciplinary in nature, introducing cognitive psychology into the study on classroom assessment. This interdisciplinary inquiry makes a concerted effort with other researchers in repositioning classroom assessment, formative assessment in particular, within self-regulated learning in order to develop the theory for classroom assessment. The findings may provide nuanced in-depth understanding of the intertwined relationships between the core aspect of classroom assessment, i.e., teacher feedback, and the key components of self-regulation, i.e., self-efficacy and self-regulatory control, which may contribute to the development of classroom assessment theory. Second, the findings of this study have potential pedagogical implications. They may provide guidance for EFL teachers to avoid superficial adoption of formative assessment and capture its spirit, i.e., empower students to become self-regulated learners (Black & Wiliam, 2009).

7. Development of the Norwegian formative writing assessment package

Aasen Arne Johannes, Norwegian Centre for Writing Education and Research, arne.j.aasen@skrivesenteret.no

Lars Sigfred Evensen, Norwegian University of Science and Technology, lars.evensen@ntnu.no

Anne Holten Kvistad, Norwegian Centre for Writing Education and Research, anne@skrivesenteret.no

Siri Natvig, Norwegian Centre for Writing Education and Research, siri@skrivesenteret.no

Jorun Smemo, Norwegian Centre for Writing Education and Research, jorun@skrivesenteret.no

Gustaf Bernhard Uno Skar, Norwegian Centre for Writing Education and Research, gustaf.b skar@hist.no

This poster describes the development process of the Formative Writing Assessment Package (FWAP), which is funded by the Norwegian government. The FWAP is an online tool available for teachers in Norway. It consist of items, generic standards applicable to writing in a range of different contexts and exemplar texts associated with each standard. It also includes guidelines on how to use information from writing assessments when planning instruction. The purpose of the FWAP is to promote data-driven, formative writing instruction in Norwegian classrooms. An additional purpose is to establish a national norm community, i.e. shared meanings of what writing is and what it means to be a proficient writer. This has resulted in a database “bottom-up-design” in the development of scoring rubrics (c.f. Fulcher, 2012). Writing is defined as a key competency in the Norwegian National Curriculum. Writing is, therefore, taught across the curriculum, and all teachers share a responsibility for teaching writing within their subjects. The National Curriculum is informed by functionalist theories of writing in general.
and the theoretical model the Wheel of Writing (Berge, Evensen, & Thygesen, 2014) in particular. Commissioned by the government agency The Norwegian Directorate for Education and Training, the FWAP is developed by the Norwegian Centre for Writing Education and Research. A single writing assessment package is developed during a three-year cycle. In the 1st year, items are constructed and field tested on a small scale. A panel of a few teachers and test developers reviews the functionality of items by reading and marking scripts, as well as, reviewing the alignment between the items and the National Curriculum, subject syllabi and the theoretical model. Through this process a number of items are often dropped. In the 2nd year, items are piloted on a national scale with a representative sample of schools. In the 3rd year, four items are taken by 400 test takers per item. This data is used to continue to refine the scoring rubrics and identify exemplar texts. Consistent with the goal to establish a norm community, the rating procedure and development of the generic scoring rubric has a “bottom-up-design”. In the two final stages of the development cycle, the scripts are rated by two independent pairs of raters (i.e. 2x2), drawn from a national panel of 90+ trained teachers. To establish a culture of negotiation and eventually shared norms, the composition of pairs is altered so that two raters form a pair for only one rating session. Moreover, on the assumption that the panel of teacher-raters represents professional norms, the scoring rubrics are developed based on the 2x2-ratings, being thus entirely empirical (i.e. representing “actual” rather than “aspirational” standards) and coming from “below” rather than top down from the test developer.

8. Bridging the gap between large scale assessment and classroom assessment for learning

Phallis Vaughter, Educational Testing Service, pvaughter@ets.org
Pablo Garcia Gomez, Educational Testing Service, pggomez@ets.org
Trina Duke, Educational Testing Service, tduke@ets.org
Susan Hines, Educational Testing Service, shines@ets.org
Emilie Pooler, Educational Testing Service, epooler@ets.org

Assessment for learning (AfL) is an approach to formative assessment that is increasingly being explored in ESL/EFL contexts (Davis, 2007). Effective formative assessment encourages students to ask three fundamental questions about their learning process: “Where am I going?,” “Where am I now?,” and “How can I close the gap?” (Stiggins, 2005). AfL helps students answer these questions by using a few strategies that, taken together, provide the basic conceptual framework behind the approach. Some of the basic strategies are: - Providing a clear and understandable vision of learning targets - Using examples and models of strong and weak work - Offering regular descriptive feedback - Teaching students to self-assess and set goals - Designing lessons to focus on one aspect of quality at a time - Teaching students focused revision - Engaging students to self-reflection, letting them keep track of and share their learning AfL is a relatively new concept in ESL/EFL, few materials and concrete examples of possible implementations of the approach have been made available for teachers. Drawing on their expertise as ESL/EFL teachers, assessment specialists, and teacher trainers, the presenters have designed a professional development workshop that highlights AfL approaches. The workshop helps teachers
implement AfL through the use of large-scale assessment tools and practices, such as scoring rubrics, sample responses, descriptive information of responses, can-do statements, performance descriptors, and calibration processes and techniques. In this poster, the presenters will provide a brief introduction to AfL principles based on their extensive review of research on the topic. They will also provide a detailed description of the workshop content and the tasks in which participating teachers engage, such as: -discussing strategies to make test scoring rubrics student-friendly for use in the classroom - working with test-taker responses at different levels of proficiency -practicing scoring sample responses using scoring rubrics and benchmark responses as a reference -engaging students in self- and peer-assessment -practicing how to use assessment tools to guide focused revision Finally, the poster will describe how this workshop framework can serve as a reference in replicating the process with different content materials and for ELT courses of various kinds.

9. Maximizing feedback for learning: Investigating language learners’ differing cognitive processing when receiving computer-based feedback

Maggie Dunlop, OISE/University of Toronto, maggie.dunlop@utoronto.ca

Computer-based language testing is increasingly adopted by institutions required to serve large numbers of learners with limited human resources (Timmers & Veldkamp, 2011). The development of feedback systems by which learners can successfully reflect on and plan for learning is essential in these contexts, as the absence of an expert mediator such as a teacher requires learners to independently monitor their learning progress and goals. This process can be especially fraught for learners engaging in the complex, non-linear, process of learning foreign and second languages. However, although the language assessment field has paid much attention to the validation processes of test constructs and scoring methods (Messick, 1989; Weir, 2005), it is only beginning to explore issues of how test feedback is delivered to and used by learners (Chappelle, 2012; Jang, 2005). As testing becomes increasingly computer-based (cf. Criterion, DELNA, DIALANG, PTE, TOEFL iBT), understanding how to maximize the impact of computer-based feedback is imperative for language testers. This study addresses these gaps in our understanding of how learners negotiate computer-based feedback on foreign language tests by investigating ways in which learners of varying learning and affective backgrounds engage in processing computer-based feedback. The study adopts an experimental research design and utilizes mixed methods. Data is being gathered from adult immigrant English as a Second Language learners in Canada, and uses a reading test from the CELPIP, a Canadian test of English for immigration purposes. The reading test is being adapted to identify current learner abilities in key cognitive language skills. Adult immigrants to Canada who are currently studying English are invited to participate in this study. Study participants (n=100) receive their feedback report and plan their learning, all online. After participating in data collection activities, they download their report and learning plan to keep. Three data collection activities are taking place. The three activities are designed to collect multiple sources of data to inform the following research questions: How do students vary in their processing of different aspects of feedback reports? What are the relationships between processing feedback and using feedback for planning? What aspects of feedback reports and planning guidance facilitate most effective usage? After viewing their reports, all participants will complete a survey designed to draw out cognitive aspects of attention, processing, interaction, and application. Thirty participants will also receive their reports while using an eye tracking device, then complete a stimulated recall interview to elicit information on participants’ cognitive attention, processing and interaction. Finally, all participants are invited to participate
in a follow-up interview one month later to elicit information on ongoing processing, interaction and application. Data collection will be completed December 2014. The results of this study, which will be available by March 2015, will inform our understanding of the cognitive processes by which language learners engage with making meaning of descriptive, process-oriented computer-based feedback. These enriched understandings will assist educational institutions and test developers in developing meaningful, personalized computer-based feedback processes for learners, particularly where direct individual interaction with a subject expert is unavailable.

10. Consequences reconsidered: validating the new FSI speaking assessment

**Erwin Tschirner**, University of Leipzig, tschirner@uni-leipzig.de

**Christina Hoffman**, US Department of State, Foreign Service Institute, HoffmanCN@state.gov

**Olaf Bärenfänger**, University of Leipzig, baerenfaenger@uni-leipzig.de

The Speaking Proficiency Assessment of the US Foreign Service Institute (FSI) is considered by many the forerunner of most current oral proficiency assessments. The FSI scale has given rise to the ILR Level Descriptors and ACTFL Proficiency Guidelines and has influenced the Common European Framework of Reference for Languages (CEFR) among many others. In line with current calls to expand the notion of validity to include the impact and consequences foreign language assessments have, FSI decided to revise its speaking assessment on the basis of a needs assessment completed at US embassies and consulates worldwide. FSI, then, commissioned a validation study to have the new format validated independently. The validity study consisted of a side-by-side study of the old and new format of the speaking assessment across six languages: Arabic, Chinese, French, Russian, Spanish, and Turkish. A total of 41 speaking assessments following the new format were analyzed as well as a total of 36 speaking assessments following the old format. Old and new speaking assessments were administered to the same examinee for comparison purposes. These assessments were reviewed and re-rated by certified ILR testers who were native speakers of the language they assessed. In addition to providing ratings following two different protocols (FSI and DLI), the ILR testers filled in a 129 item questionnaire focusing on test administration and rating procedures, appropriateness of topics and tasks, examinee performance, and comparing old and new formats. Moreover, they participated in a focus group after completing the reanalysis of the assessments. The poster will describe the changes to the FSI speaking assessment and why they were implemented. It will also describe the results of the study to validate the new speaking assessment using Bachman/Palmer’s (2010) assessment use argument.

11. Validating a new instrument to test and teach sociopragmatic comprehension

**Veronika Timpe Laughlin**, Educational Testing Service, vtimpe@ets.org

**Ikkyu Choi**, Educational Testing Service, ichoi001@ets.org

Pragmatics has long been anchored in language competence frameworks. Instruments such as role-plays as well as written and oral discourse completion tasks have been used successfully to assess L2 pragmatic competence. However, while the majority of pragmatic tests are aimed at measuring productive skills, the assessment of receptive pragmatic skills is still an underrepresented field. This study introduces and explores the validity of the American English Sociopragmatic Comprehension Test (AESCT), a web-based sociopragmatics test consisting of 54 multiple choice discourse completion tasks. These tasks were designed to measure knowledge of speech
acts/implicatures (Section 1), routine formulae (Section 2), and culture-dependent differences in lexis (Section 3). The main purpose of this study was to investigate systematically the performance of the AESCT and explore internal, structural, and external aspects of construct validity as identified by Messick (1989). Therefore, the following research questions were proposed: (1) What are the AESCT’s psychometric characteristics? (2) Does the empirical factor structure of the AESCT reflect the construct? (3) Do students with high frequency of exposure to the target language score higher on the AESCT and its subsections than students with less frequent exposure? (4) Which types of target language input contribute to higher sociopragmatic competence as measured by the AESCT? To explore these questions, 97 university-level EFL/ESL learners (n=97) were asked to take the test along with (a) the Cambridge Placement Test (CPT) and (b) a background questionnaire about their exposure to different types of English language media. To address the internal aspect of construct validity descriptive statistics, G-theory, and item response theory were used. For the structural aspect, factor analysis was employed. Finally, to investigate the external aspect of construct validity, linear regression and multilevel modeling were used to analyze the relationship between the test’s total and section scores and different background variables. Findings provide preliminary evidence for the AESCT’s construct validity. First, descriptive statistics and item response theory showed that the AESCT provides reliable estimates of learner’s sociopragmatic knowledge. Moreover, correlations between the sections and factor analysis suggested that the three sections of the AESCT (measuring speech acts/implicatures, routine, and culture-dependent differences in lexis) are highly related and the test is essentially unidimensional. Finally, linear regression and multilevel modeling showed general English proficiency, living experience in the TL culture, and exposure to target language audiovisual media as features that were found to improve learner’s sociopragmatic comprehension. Surprisingly, frequent use of print media was not found to be a significant predictor for higher AESCT scores – an interesting finding given the large amount of text-based learning materials used in EFL/ESL instruction. In sum, the results reveal the reliable prototype of a test that can be used to assess students’ receptive pragmatic abilities. Moreover, they also provide ideas as to how the AESCT can be used in EFL/ESL education for the purpose of raising learner’s awareness of culture-specific sociopragmatic differences.

12. Low-level learners and computer-based writing assessment: The effects of computer proficiency on written test scores

Laura Ballard, Michigan State University, ballar94@msu.edu

In recent years, many large-scale, high-stakes tests have moved from paper to computer-based formats (e.g. TOEFL, IELTS, WIDA, Smarter Balanced). In response, education, technology, and language testing researchers have conducted comparability studies on these testing modes (Endres, 2012; Kim & Huynh, 2008; Kingston, 2009). While existing research on testing mode addresses the validity and ethics behind each mode, there is a lack of research examining their effects on specific populations of language learners; in particular, low-level ESL learners who potentially have limited computer exposure. To address this, I investigated: 1) if low-level learners perform differently on paper-based and computer-based writing assessments; and 2) if so, if score differences could be attributed to computer proficiency. Twenty-nine low-level English learners completed a paper-based and computer-based version of a university ESL writing placement test. The participants took an L1 and a L2 computer-proficiency test and responded to a questionnaire about computer habits. Results indicated no statistical difference between participants’ paper-based and computer-based writing scores; thus, computer proficiency had a negligible effect on writing scores. Students’ computer proficiency, however, mainly affected
their writing fluency, yet length was not taken into account in the rubric. Additional analyses revealed an interaction between writing fluency and writing-mode preference, a positive correlation between computer proficiency (in both the L1 and L2) and L2 proficiency, and a staggeringly low computer proficiency overall (17 WPM in the L1, and 15 in the L2). Importantly, the study revealed that student L2 computer proficiency lagged behind L1 computer proficiency until students reached an intermediate English-language proficiency, suggesting a lack of automatization in typing and letter encoding in novice- and beginner-level students. I use the results to discuss test-mode effects and computer-based test policies and protocols (time limits and mode optionality). I conclude by calling for computer processing instruction in ESL writing courses and more in-depth research on low-level L2-learner populations.

13. The role of visuals in listening test for academic purposes

Seyyedeh Saeede Haghi, University of Warwick, S.S.Haghi@warwick.ac.uk

Khosro Vahabi, Ozyegin University, khosro.vahabi@ozyegin.edu.tr

With the increase in the use of visual components in teaching listening comprehension, researchers in language testing began to investigate the role of these visuals in the testing of listening comprehension. Despite the growing body of research examining the inclusion of visual stimuli in listening tests, there are aspects which still invite further investigation. One apparent gap in the literature is the unequal scrutiny given to the two types of visuals used in academic listening texts. While many studies have assessed the role of one type of visual ‘context visuals’ in listening tests, very few have investigated the role of the other type, known as ‘content visuals’. In addition, a considerable amount of research has relied heavily on quantitative data obtained from comparative research design, which, though addressing the outcomes of experiments, has failed to properly explain the role of individual factors in those outcomes. This study therefore intended to broaden this area of research by investigating the role of these visuals in listening tests for academic purposes, focusing on both context and content visuals and using the mixed-method research approach. To this end, the effect of three types of delivery modes (audio-only, audio-video and audio-multimedia) on test-takers’ performance and note-taking practices in a listening test for academic purposes, as well as their perceptions towards the inclusion of these visuals, were surveyed and analyzed using a quantitative and qualitative mixed-method approach. The former approach included the analysis of test scores (making comparisons between the three groups), the content of the notes (both quantity and quality), and the responses to the Likert scale items in the exit questionnaire (the perceptions of test takers towards the three delivery mode). And the latter one consisted of the analysis of the open-ended items in the exit questionnaire. The results from the analysis of the primary and the secondary databases were then mixed in order to address the three research questions. Findings from this study revealed no significant difference in performance between the audio-only and content visual groups; however, a significant difference was noted between the audio-only group and the context visual group, with the audio-only group gaining significantly higher scores. These findings suggest that while content visuals seemed to have neither debilitative nor facilitative effects on test performance, the context visuals did appear to be inhibitive. The analysis of participants’ perceptions towards the inclusion of visuals also confirmed these results, with the content visual group showing a more positive attitude than the context visual group. These results also demonstrated that while the content visuals did not seem to contribute to, or take away from, test-takers’ performance in the test, they did seem to have significantly facilitated the efficacy and organization indices of notes. These findings make contributions to both testing and pedagogical aspects of listening for academic purposes, which are also identified in this study.
14. Factor structure of the test of English for academic purposes

Yo In'nami, Shibaura Institute of Technology, innami@shibaura-it.ac.jp

Rie Koizumi, Juntendo University, rie-koizumi@mwa.biglobe.ne.jp

Keita Nakamura, Eiken Foundation of Japan, ke-nakamura@eiken.or.jp

Factor structure of the Test of English for Academic Purposes Improvement in university entrance examinations to strengthen the positive relationship among language teaching, learning, and assessment is a recurring issue in Japan. In this vein, the Eiken Foundation of Japan and Sophia University, in collaboration with the University of Bedfordshire in the UK, recently developed the Test of English for Academic Purposes (TEAP), an exam for Japanese university applicants designed to measure four skills in an academic context: reading, listening, writing, and speaking. The test is aimed at levels A2 through B2 of the Common European Framework of References (Council of Europe, 2001). In contrast to traditional university entrance English examinations in Japan, which primarily measure reading and listening skills, the TEAP also includes writing and speaking in the hope that targeting a wider range of relevant and representative constructs in an academic setting will have a positive impact on learning at the secondary level. As the TEAP can be used in university admissions and the stakes are high, validation studies have been conducted, including testing for inter- and intra-rater reliability and the discriminability of contextual variables on the reading and listening sections (Taylor, 2014) and the writing section (Weir, 2014); also investigated have been stakeholders’ perceptions of university entrance examinations and the washback expected from the introduction of the TEAP (Green, 2014; Nakamura, 2014). One area of interest that has yet to be examined is the factor structure of the TEAP. This gap needs to be filled, because studies on factor structure show whether there are empirically supported relationships between the intended interpretation of scores and the constructs being measured (e.g., Messick, 1996; Sawaki, Stricker, & Oranje, 2009). Data were obtained from 100 Japanese EFL university students taking both the TEAP and TOEFL iBT—another test of English in an academic setting. We first analyzed the TEAP data to examine its factor structure; the best-fitting model was then tested with the TOEFL iBT to examine if both tests measured similar constructs. For both analyses, we used confirmatory factor analysis. First, for the TEAP, we analyzed three competing models: a one-factor model hypothesizing a unitary factor underlying the four skills; a correlated model hypothesizing correlations among variables; and a receptive/productive model hypothesizing one factor underlying the receptive skills (reading and listening) and one underlying the productive skills (writing and speaking). The results showed that the best-fitting model was the one-factor model. Second, this one-factor TEAP model was tested with the TOEFL iBT. The results showed a strong correlation between the two academic English proficiency factors, suggesting that both tests measure similar constructs in academic settings. However, negative relationships between reading and speaking measurement errors within and across the tests seem to reflect that the students were heavily focused on studying reading to pass university entrance examinations, leading to less time spent on speaking and thus that those who had higher reading proficiency likely had less speaking proficiency. Implications for the improvement of testing and teaching will be discussed at the poster.
15. The JLTA online tutorials for the promotion of language assessment literacy in Japan

Yasuhiro Imao, Osaka University, imao@lang.osaka-u.ac.jp
Rie Koizumi, Juntendo University, rkoizumi@juntendo.ac.jp
Yukie Koyama, Nagoya Institute of Technology, koyama@nitech.ac.jp

This poster reports on online tutorials created by the Japan Language Testing Association (JLTA), partially funded by a Workshops and Meetings Grant from ILTA. While enhancing language assessment literacy through face-to-face workshops is a useful endeavor, many teachers do not have time to attend them, yet still face a variety of difficulties related to language assessment every day. There are websites that cater to this need, such as the “Companion Online Tutorial” by the Center for Applied Linguistics (Malone, 2013) and the “Language Testing Resource Website” (Fulcher, 2014). These websites in English allow people interested to access information on matters such as how to compose good language tests and how to interpret test scores. As an organization dedicated to serve language teachers in Japan, we decided to serve local needs and create online resources in both Japanese and English. JLTA published useful materials on its website, such as web tutorials for language testing, which this presentation focuses on. We conducted a needs analysis using a questionnaire to investigate language instructors’ understanding of language testing and their related needs (Thrasher, 2012). The results, from approximately 60 language teachers from lower secondary to university level, showed that they felt their knowledge of language testing to be somewhat insufficient but that they were interested in a wide range of topics in language testing, although with a focus on basic practical issues over theoretical ones. These results led to the decision to prioritize basic topics and materials written in Japanese. We selected two groups of topics for the JLTA language testing online tutorial: (a) the basics of language testing and (b) test data analysis for test improvement and basic statistics of language education/test research. The first group of topics, which were the initial focus, included validity, reliability, practicality, test constructs, test specifications, and testing the four skills, vocabulary, and grammar. JLTA members were asked to contribute slideshow materials to the project—anatomized slides to help the main target learners (secondary school English language teachers) learn by themselves. Next, 17 English language instructors from elementary school to college level were asked to evaluate the online tutorial materials and give suggestions for revision. Overall, the tutorials were well-received by the reviewers, with some minor suggestions for revisions. The authors of the materials were requested to revise their materials based on this feedback. To allow access not only from PCs but also from tablet devices, whose use is increasing, the tutorial slides are in the JPEG image format; they were watermarked and a web slideshow was generated for each slide file using HTML and simple JavaScript. At the time of this writing, 11 web tutorials are available on the JLTA website (see https://jlta.ac/course/view.php?id=35). We will next create further web tutorials covering the second group of topics, improve the tutorials based on feedback from users, and continue to promote understanding of language testing among language education professionals in Japan.
16. Health care Finnish: Developing and assessing Finnish proficiency among health care professionals

Tiina Lammervo, Centre for Applied Language Studies, University of Jyväskylä, tiina.lammervo@jyu.fi

Marja Seilonen, Department of Languages, University of Jyväskylä, marja.a.p.seilonen@jyu.fi

Minna Suni, Department of Languages, University of Jyväskylä, minna.suni@jyu.fi

The project focuses on the development of second language proficiency assessment of health care professionals who have tertiary qualifications from outside the EU/EEA countries and have applied for professional authorisation in Finland. The topic is timely as defining the level of language proficiency that is required for employment in the health care sector and for ensuring patient safety is under wide discussions in many European countries including Finland. Currently second language proficiency of professionals from outside the EU/EEA countries is demonstrated through the Finnish National Certificates of Language Proficiency (FNCLP) intermediate level test of Finnish (levels B1-B2). The issue has been raised whether the minimum level B1 skills demonstrated in a test of general Finnish is sufficient for a health care professional in Finland, and some reforms have officially been suggested. On the other hand, employers have a key role as gatekeepers, since they are responsible for ensuring that the language skills of the hired health care professionals really meet the needs of their tasks at work. Employers have the right to require a higher proficiency level than B1. In practice, however, due to shortage of labour force, EU/EEA professionals may sometimes be employed even with relatively limited, undocumented L2 skills. The poster presents some key results of the qualitative analysis of health care professionals’ productive skills performance in L2 Finnish as demonstrated by the current language test. Results of this analysis will contribute to the design of an assessment experiment. The data of 25 medical professionals originate from the FNCLP corpus (levels B1-C2). The findings are discussed in relation to the participants’ background information, which includes participants’ migration histories, authorisation processes, language studies and self-assessments. The information is available from the FNCLP corpus and provided by the National Supervisory Authority for Welfare and Health (Valvira). The experiment of professional language proficiency assessment is conducted with reference to the National Certificates of Language Proficiency. In the experiment, productive professional language skills will be assessed in a tailored hybrid test version in which receptive skills will be tested in general Finnish. The experiment will thus provide information on, for instance, how a performance based language assessment could complement the language proficiency profile of a professional. Much of the professional language learning takes place on the job, and the test construct and criteria should reflect this (see e.g. Elder et al. 2012). In the future, a module of professional language proficiency could be included in demonstrating adequate healthcare qualifications. This might have a wash-back effect on the development of field-specific language education and an impact on the authorisation and recruitment processes.

17. Looking beyond scores: Validating a school-based high-stake speaking assessment

Li Liu, Renmin University of China, refilane@gmail.com

Guodong Jia, Renmin University of China, gdjia@ruc.edu.cn

The current study focuses on a school-based high-stake oral proficiency assessment, Test of Oral Proficiency in English (TOPE) at Renmin university of China. TOPE is the first attempt in mainland China to assess the oral communicative
proficiency of college students who have undertaken compulsory English study, based on the College English Curriculum Requirements (Ministry of Education, 2007). The TOPE adopts a face-to-face format, consisted of three major sections: 1) Reading aloud and retelling; 2) Presentation and discussion and 3) Impromptu speech and Question & Answer. With the aim of determining the overall validity and acceptability of the assessment, the designed study is intended to explore following two research questions: 1) To what extent does the TOPE test elicit intended language functions? 2) Is there any evidence from candidate performance that validates the scale descriptors? Students’ speaking performances in real testing situations were video-recorded and transcribed. As the actual assessment are conducted by college teachers, views and professional judgments of teacher-assessors regarding the testing and rating procedures were also consulted by using questionnaires and interviews. To analyze the data, language function analysis was conducted for each speaking task and instances of functions derived from O’Sullivan et al (2002) were counted. In addition, discourse analysis of students’ speech performances was performed in terms of salient features characterized each level. Results of different proficiency groups were compared to identify to what extent each of these features differs between adjacent proficiency levels. Overall, the study led support to the validity of the TOPE. Implementing the school-based assessment is highly challenging in mainland China, issues involved in the process of development and validating the test including teacher-assessor motivation, assessment management and implementation, assessment for learning were also discussed.

18. Assessing low educated L2-learners – issues of fairness and validity

Cecilie Hamnes Carlsen, Vox - Norwegian agency for lifelong learnnig, cecilie.carlsen@vox.no

Assessing low educated L2-learners – issues of fairness and validity When constructing language tests, one of the many factors constructors need to take into account, is the group of test takers for which the test is intended. A test for young learners needs to be constructed differently from a test for adult learners, and a test, which may yield valid scores for one group, may not be suitable for another (Bachman 1990). One test taker characteristic, which poses specific challenges in language assessment, is the candidates’ educational background, or to be precise: the lack of such. Adult L2-learners with no or very poor literacy skills form part of the immigrant population. As more and more European countries make language requirements for entry to the host country, for permanent residency or citizenship as well as for entrance in the work marked, low educated learners form an increasing part of the test population as well (Extramiana, Pulinx & Van Avermaet 2014). Test fairness is defined in different ways in the test literature. Most professionals seem to agree that it should be free of bias, and all candidates should have and equal opportunity to demonstrate their language skills (Shaw & Imam 2013, Kunnan 2014). Fairness is closely linked to validity. To be valid, a test should measure as much as possible of the construct in question, and as little as possible of other skills and abilities. The latter is referred to as construct-irrelevant variance, and it is a matter of particular challenge when assessing learners with limited school background and prior experience with testing (Allemano 2013). The aim of this paper is twofold: Firstly, I will present the development of a new CEFR-based test of Norwegian as a second language in which the concern for the low-educated learners has been highly prioritized from the start. In the presentation I will highlight specific measures that were taken in the test format and
test system to accommodate to the needs of this group of test-takers. I will also show how specific attention to the group of low-educated learners necessitates heavy contextualization, adapted test formats, extensive use of pictures, avoidance of hypothetic situations where possible etc. Secondly, I will present the test results of this particular group of test takers as compared to the results of learners with more school background in order to demonstrate that if constructed with low educated learners in mind, it is indeed possible to construct a fair and valid test for this group, giving everybody a fair chance to demonstrate their language abilities. This is a matter of existential importance to the individuals it concerns, and an ethical responsibility of policy makers and test constructors.

19. Developing an online, automated system to provide diagnosis, feedback and tracking of academic English writing ability

Alan Urmston, Hong Kong Polytechnic University, alan.urmston@polyu.edu.hk
Michelle Raquel, Lingnan University, michelleraquel@ln.edu.hk

In Hong Kong, there has been renewed emphasis placed on academic literacy at tertiary level in both English and Chinese, especially in terms of English writing (Lockwood, 2013). The support for developing students’ academic writing skills, however, is limited to formal English language classes in the first and second year and support provided by self-access centres (e.g. writing clinics). What is needed is a cost-effective, easily accessible system that will allow students to have their writing analysed for weaknesses, provide students with useful feedback on how to improve these weaknesses, as well as some form of tracking that can measure improvement. The Diagnostic English Language Tracking Assessment (DELTA) was developed to assesses the proficiency in reading, listening, grammar and vocabulary of NNS students at English-medium universities using a web-based platform. Students receive a report which shows their overall proficiency level (DELTA Measure), Component Skills Profile and Component Diagnostic Reports. From the reports, students know which of the four components they are relatively weaker in and the sub-skills that caused them unexpected difficulty. Given the importance of writing at tertiary level, a writing component for DELTA was deemed as being essential and so this has been in development since 2012, with the features described above, i.e. diagnosis, feedback and tracking in an online, automated system. This poster describes the development of the DELTA Writing Component. It will provide information on task and rubric design, the utilisation of Lightside Labs as its automated essay scoring engine, and the design of an automated feedback system. The poster will also report on the results of the trial of this system at five universities in Hong Kong, in particular, students’ perception about effectiveness of this system as a tool to enhance their writing ability. It is envisaged that the addition of this Writing Component will further enhance the effectiveness of DELTA as a diagnostic and tracking tool that will support students’ development of academic literacy at the tertiary level.

20. Rubric development through the lens of a can do philosophy

Megan Montee, Center for Applied Linguistics, mmontee@cal.org
Tanya Bitterman, Center for Applied Linguistics, tbitterman@cal.org

What values underlie test constructs, and how are these values realized within a test? Tests are powerful decision-making tools, and a consideration of the values inherent in assessments can lead tests developers to a more careful and thorough consideration of potential consequences. As McNamara & Roever (2006) have argued, language testers must consider the social dimension of assessments as an essential
R ubrics are the link between performances on a test and the test score, and rubric descriptors mediate how raters and other stakeholders understand examinee performances. Test developers face a challenge in creating rubric descriptors that are meaningful to raters and grounded in the underlying test construct. In reference to scoring scales, North (2000) has described this challenge as “trying to describe complex phenomena in a small number of words using incomplete theory.” This poster describes the development of a rubric and scoring descriptors for a computerized assessment of academic English language proficiency for students in U.S. grades 1-12. In this project, a single rubric was developed to score the oral performance of students ranging in age from approximately 7-18 years old. Researchers faced the challenge of creating a single scale that would be interpreted in relation to the cognitive development of examinees. In developing the rubric and related scoring materials, researchers addressed the following questions: How are language proficiency levels realized across various ages and developmental levels? What criteria do raters need to understand and apply the rubric across grade levels? The development process for the rubric drew from both empirical and theoretical approaches. Researchers analyzed examinee performances within and across grade levels in order to understand how rubric features are realized at by examinees at different grade levels. The research team investigated the following features of examinee performances: comprehensibility, fluency, linguistic complexity,
language forms, and vocabulary. Researchers also analyzed grade-level standards for language and content knowledge development as well as theory related to second language acquisition and cognitive development. Draft rubric descriptors were then applied to performances across grade levels and raters provided feedback that was used to further refine descriptors and supporting materials. The result of this project was a single rubric to use across grade levels as well as a set of rubric interpretation guides. The rubric interpretation guides describe how the rubric should be interpreted at different grade levels. Finally, the results of the project informed the design of self-access rater training materials. In the poster, we present the rubric development process, discuss how the analysis of examinee performances affected rubric descriptors, and consider implications for approaches to rubric development.

22. Assessing future learning: Testing an ability to identify definitions for technical vocabulary in introductory textbooks

**Daniel Isbell**, Northern Arizona University, daniel.r.isbell@nau.edu

This paper reports on an innovative measure of prospective ESL university students’ abilities to develop knowledge of new terms presented in a text. When students enter an American university, they typically take a variety of introductory courses. In these courses, information is commonly (if not primarily) provided to students through textbooks, where they begin learning foundational concepts and technical vocabulary of a field. This technical vocabulary accounts for a considerable portion of words in textbooks (Chung & Nation, 2003), and much of it considered beyond the realm of L2 instruction (Read, 2000). However, introductory textbooks are notable for providing in-text definitions (Carkin, 2001). Assessing whether or not L2 students have sufficient ability to connect in-text definitions to technical vocabulary is thus potentially valuable information. Existing direct measures of L2 vocabulary primarily focus on existing vocabulary knowledge (Read, 2007), and common indirect measures involve the robustness of vocabulary in spoken/written production. Neither of these methods would appear to account for the learning of unknown technical vocabulary through reading. The present study aimed to 1) design a test which measures the ability to correctly identify in-text definitions for technical vocabulary and 2) compare two item formats for delivery of the test. To those ends, a test called the Vocabulary through Reading Test was developed. An excerpt from an introductory biology textbook was chosen as input, and technical terms supported by in-text definitions were substituted with nonwords to control for existing vocabulary knowledge (K = 10). In a short-answer format, examinees (N = 158) read the text and wrote definitions for each of the nonwords. Responses were scored dichotomously based on meaning. A subset of responses was scored by two human raters, who had acceptably high agreement. This allowed for scoring to be completed by a single rater, yielding an internal consistency (Cronbach’s alpha) of .79. The test was found to effectively make distinctions among examinees (M = 5.12, SD = 2.84), and items were within desirable ranges of facility and discrimination. However, while human scoring was acceptably consistent, it was also labor intensive. To explore the issue of item format, a multiple-choice version of the test was developed using responses from the short-answer version to create keys and distractors for each item. 147 different examinees took the multiple-choice version, which also effectively made distinctions among examinees (M = 5.48, SD = 2.23), though its internal consistency was somewhat lower (Cronbach’s alpha = .61). Mean scores for the two groups were not significantly different, suggesting that the two versions of the test provided equivalent information. Items in both formats generally performed similarly, as did first-language examinee subgroups. Further analysis revealed some difference in one of the two subconstructs, which may suggest a small degree of construct
irrelevant variance due to format. In terms of practicality, the multiple-choice version was more easily and quickly scored. These results should be of interest to higher education stakeholders and of relevance to language testers interested in item formats and the integration of reading and vocabulary measurement.

23. Comparing dimensionality assessment approaches using empirical data: Investigating the dimensionality of CELPIP-G reading test

Michelle Yue Chen, The University of British Columbia / Paragon Testing Enterprises, mchen@paragontesting.ca

Background Determining test dimensionality structures serves important purposes (Tate, 2003). The widely used passage-based reading comprehension items, however, poses challenges to the determination and interpretation of the reading test dimensionality. Previous research showed controversial findings about whether the dependence of items led to multidimensionality (e.g., Schedl et al., 1996; Zwick, 1987). Many approaches have been used to assess dimensionality. The most common parametric approaches include eigenvalue rule (Kaiser, 1960), scree plot (Cattell, 1966), Velicer’s minimum average partial (MAP) test (Velicer, 1976), parallel analysis (Horn, 1965), and difference chi-square (Schilling & Bock, 2005). Besides, several nonparametric are also used in the literature, including HCA/CCPROX (Roussos, 1995), DETECT (Zhang & Stout, 1999), and DIMTEST (Nandakumar & Stout, 1993). However, these different approaches do not usually lead to the same conclusion. Purpose Given that few studies have provided comprehensive comparisons and discussions of these commonly used approaches. This study attempted to conduct an empirical comparison of these methods in real language test contexts. The research questions we aimed to answer include: 1) For each form of the reading test, do different dimensionality assessment approaches produce consistent results of the test dimensionality? 2) How do the items cluster and how to interpret the results? Whether there is a consistent pattern between the two forms? Method Two forms of the reading test in the Canadian English Language Proficiency Index Program (CELPIP-G) Test, which consists of 4 passages with 38 passage-based items, were used as examples in this study. The CELPIP-G reading test was designed to assess test-takers’ ability to engage with, understand, interpret, and make use of written English texts. Similar to many other testing programs, multiple equivalent forms were constructed under the same test blueprint in CELPIP-G. This study randomly selected two forms of the reading test to be examined, which allowed us to compare the results from different forms. Eight commonly used approaches in assessing dimensionality were compared. These approaches are 1) eigenvalue rule, 2) scree plot, 3) MAP test, 4) parallel analysis, 5) difference chi-square, 6) HCA/CCPROX, 7) DETECT, and 8) DIMTEST. Results As expected, these different dimensionality investigation methods do not produce exactly identical solutions for the same test. Eigenvalue rule and scree plot methods tend to suggest a larger number of underlying dimensions. The results were similar across the two forms. Item clusters were found, but the pattern of how the items were clustered together was not clear based on the two forms analyzed in this study. The essential unidimensionality was observed on both forms, which supported the current unidimensional item response theory (IRT) based scoring procedure employed by CELPIP-G. Although a dominant factor was found underlying the reading test, several minor dimensions were detected as well. Differential item functioning (DIF) analyses didn’t flag any item, which suggested that these minor dimensions may not be caused by DIF items. In other words, these minor dimensions can also be meaningful. More studies are needed to understand these important but usually be ignored minor factors in reading assessment.
24. A healthy context: Navigating language access policies and regulations to test the language abilities of Canadian hospital employees

Lauren Kennedy, Second Language Testing, Inc., lkennedy@2lti.com

Kevin Joldersma, Second Language Testing, Inc., kjoldersma@2lti.com

Marie-Soleil Charest, The Ottawa Hospital, mcharest@toh.on.ca

This presentation is co-authored by representatives from a language test provider and a major Canadian hospital system in a bilingual English-French environment. Within Canada a series of federal and provincial policies and regulations govern English and French language access policies for patients, staff, and administrators. Coupled with multiple and sometimes conflicting guidelines and recommendations from internal and external organizations, historical internal practices, and changing community profiles, Canadian hospital administrators have a difficult task navigating language access and language assessment policies within hospitals. The first part of this presentation will describe the federal and provincial language access policy and language assessment regulatory environment that hospitals must navigate in order to remain in compliance. It will also review the internal language policy and assessment policy decisions one major Canadian hospital has made to best meet the needs of its community. The planned, unplanned, and hoped for effects of requiring and promoting staff bilingualism within the hospital on patients, staff, administrators, and the community will also be presented. In the second part of the presentation, the combined policies and regulations’ washback on the language assessment of hospital staff will be examined. The authors will describe test design decisions that were made to address particular policies, regulations, or assessment contexts. Challenges encountered during the operationalization of the Linguistic Assessments will be presented to exemplify conflicts and their resolution. The authors will also show how the Linguistic Assessments are linked to the job tasks of staff within a hospital, and how this linking has facilitated the acceptance of the assessments within the hospital community. The creation of language assessments for bilingual or multilingual populations imposes unique challenges on test designers and stakeholders to ensure balanced collaboration, access, and representation by all relevant administration, regulatory, and language groups at key points during creation, analysis, and administration phases. This paper will conclude with recommendations for policy makers, test administrators, and test developers for working with one another in high-stakes and high profile contexts.

25. Yes/No vocabulary testing for assessing class achievement

Toshihiko Shiotsu, Kurume University, toshihiko_shiotsu@kurume-u.ac.jp

Yes/No Vocabulary Testing for Assessing Class Achievement
Yes/no vocabulary testing is typically considered or used for obtaining rough estimates of language learners’ vocabulary size, but it has not been much utilised in language classrooms for other purposes. This presentation reports on an effort to utilise yes/no vocabulary testing in university EFL classes as an instrument for (1) providing the students with opportunities for self-monitoring of their learning, (2) facilitating their vocabulary reviews and their target word prioritisation processes, and (3) obtaining an index of their achievement. In an EFL course involving a reading component, a list of words appearing in the course materials were presented to the students for vocabulary reviews and preparations for subsequent yes/no testing. The vocabulary items on the list were sorted according to frequency for prioritisation purposes. Students were made aware that the yes/no testing would involve many more words
as question items compared to other types of vocabulary testing but they would not only be asked to provide self-report of their knowledge but to complete a post-yes/no knowledge test on a subset of words to adjust their scores. The above was implemented in four different groups differing in academic background, level of motivation, and proficiency. Quantitative and qualitative data were collected to evaluate the performances of the tests and explore the students’ perspectives on the approach. There are indications in the data of fairly high correlations between the yes/no test results and student performances on independent measures of their achievement, and positive attitude held by a large proportion of the students involved.

26. A validation study on the newly-developed Vietnam standardized English proficiency test

Phuong Tran, Vietnam National University, p.tran76@gmail.com  
Hoa Nguyen, Vietnam National University, hoadoe@yahoo.com  
Trang Dang, Vietnam National University, trangdangthu411@gmail.com  
Minh Nguyen, Vietnam National University, hmmnh102@gmail.com <hmmnh102@gmail.com>;  
Lan Nguyen, Vietnam National University, lanthuy.nguyen@gmail.com  
Tuan Huynh, Vietnam National University, thonhontuan@yahoo.com  
Ha Do, Vietnam National University, habong54@yahoo.com  
Huu Nguyen, Vietnam National University, maihuu@yahoo.com <maihuu@yahoo.com>;  
Fred Davidson, Illinois University, fgd@illinois.edu

This work-in-progress will present the results of the development and validation of the new Vietnamese Standardized English Proficiency Test (VSTEP), an important intended outcome of the National Foreign Language 2020 Project, which aims to enhance the English foreign language proficiency among Vietnamese people to prepare for the country’s economic growth and regional integration. The VSTEP is a four-skill test which aims to measure students’ English language proficiency from B1 to C1 levels, which are the expected English levels for post-secondary English learners according to the CEFR-VN (an adapted version of the CEFR to suit the Vietnamese context). The issues presented include: the psychometric properties of the VSTEP four sub-tests, the linkage between the items and the CEFR-VN, and the linkage of the items with the original CEFR. Implications will be presented regarding the capability of the VSTEP as a measurement tool and its potential impact on the teaching and learning of English in Vietnam when the test is officially used to test hundreds of thousands of test-takers annually.

27. Towards building a validity argument for grammaticality judgment tasks

Payman Vafaei, University of Maryland, payman.vafaei@gmail.com  
Illina Stojanovska, University of Maryland, ilina.stojanovska@gmail.com  
Yuichi Suzuki, University of Maryland, yuicchi0819@gmail.com

The current study, which is a validation study on grammaticality judgment tests (GJTs), aims to rely on advances of L2 assessment in the area of validation research to contribute to one of the challenging issues of SLA research which is measuring implicit vs. explicit L2 knowledge. The results of the present work may be used as backing for several inferences in an interpretive argument for interpreting GJTs’ scores. GJTs are among the most commonly used assessment tools in SLA field. For this reason, GJTs have been subject to many validity investigations. Several
studies have concluded that manipulating GJTs’ time condition may turn them into measures of either explicit or implicit knowledge i.e., untimed GJTs are measures of explicit knowledge and timed GJTs are measures of implicit knowledge (Bowles, 2011; R. Ellis, 2005; R. Ellis & Loewen, 2007; Loewen, 2009). However, all of these studies relied on tradition approaches to validation, and rather than building an argument for and against the interpretations that can be drawn from scores of different types of GJTs, they only provided criterion-related evidence for their claims. In many instances, the validity of the criterion against which GJTs were compared was itself under question. In addition, many of the previous studies suffered from methodological and statistical flaws such as small sample sizes or not testing rival models when confirmatory factor analysis (CFA) was used. Contrary to the previous GJT validation studies, we hypothesize that applying time pressure on GJTs do not turn them into measures of implicit knowledge, and timed GJTs measure automatized explicit knowledge at best. The current work is an improvement compared to the previous studies in the following aspects: A) to show that GJTs are too coarse to measure implicit knowledge, we operationalized the use of implicit knowledge in terms of observing when error in the stimulus is registered without awareness, or with minimized level of explicit knowledge involvement. We attempted to measure the implicit knowledge by employing two psycholinguistic methods called the Self-Paced Reading Task (SPR) and Word Monitoring Task (WMT). We also have a measure of meta-linguistic knowledge as a measure of explicit knowledge. Through comparison of performance of participants on GJTs of different types with more robust measures of implicit and explicit knowledge, we will establish better criteria for validity arguments. B) We also aim to provide evidence for our hypotheses by including two measures of cognitive aptitude for implicit (Serial Reaction Time) and explicit (LLAMA F) learning. Previous studies (e.g., Granena, 2013b; Suzuki, 2013) showed that cognitive measures may provide us with evidence for the nature of underlying knowledge participants rely on to perform on different assessment tasks. C) We will avoid the previous methodological and statistical flaws. We will have a large enough sample size (N= 100), and test several CFA rival models. Presentation of the current study, which is at the data collection stage, and receiving feedback, will assist the researchers to implement improvements in the design, materials and analyses and conclusions of the study.

28. Développement d’un outil d’évaluation du rendement en lecture pour les militaires canadiens anglophones

Geneviève Le Bel Beauchesne, Défense nationale du Canada, genevieve.lebelbeauchesne@forces.gc.ca

Sara Trottier, Défense nationale du Canada, sara.trottier@forces.gc.ca

L’Académie canadienne de la Défense, qui s’occupe de la formation linguistique des membres des Forces armées canadiennes (FAC) grâce à son propre curriculum de français, a créé un test de rendement de la compréhension de l’écrit afin d’évaluer les militaires canadiens anglophones terminant leur formation en français. La création d’un tel test s’imposait puisque le contenu du précédent contrôle de rendement était compromis par une dizaine d’années d’utilisation. De plus, il n’était pas spécifiquement en lien avec le curriculum de français des FAC. Ce nouveau test de rendement s’inscrit dans la lignée des tests traditionnels à questions à choix multiple (QCM), donné en format papier-crayon. Les items sont constitués de textes suivis de questions ou d’énoncés à compléter et de quatre choix de réponse (la clé et trois leurre). La définition du construit s’appuie sur une revue de littérature portant sur l’évaluation de la compréhension écrite en L2 en situation de test à choix multiple. Pour assurer la validité du contenu, un examen minutieux des textes du curriculum a été réalisé afin de les classifier selon leur type et leur contenu (lexique, nombres de mots, nombre de paragraphes, densité propositionnelle, etc.).
L’emploi de techniques issues de l’analyse de texte par ordinateur (ATO) a appuyé l’examen. Cette analyse a permis de spécifier les caractéristiques des items à créer. Ceux-ci ont été produits par une équipe de professionnels des domaines de l’éducation et de la linguistique. Ils sont issus de matériel authentique (articles de journaux militaires, rapports professionnels, récits, etc.) et ont été révisés par deux comités : pédagogique et militaire. Les items ont été validés à l’aide d’une méthodologie portant sur la théorie classique des tests. Près de 80 volontaires appartenant à la clientèle-cible ont participé à l’étude de validation des items. L’indice de facilité, la corrélation item-total et les patrons de distribution des réponses ont permis de retenir les items les plus performants. De plus, des données sur la perception de la difficulté des textes et des questions ont été systématiquement recueillies pour chacun des items ce qui a permis de montrer une corrélation entre la perception de la difficulté et la difficulté réelle des items, assurant ainsi une excellente validité apparente. Une deuxième étude de validation a été menée auprès de 73 volontaires issus de la clientèle-cible afin de mesurer la cohérence interne du test, à l’aide de l’alpha de Cronbach (0,78) et de fixer un seuil de réussite. Ce seuil a été déterminé par la méthode des groupes opposés. Le résultat du test en compréhension de l’écrit de la Commission de la fonction publique, auquel tous les participants ont été soumis, a été utilisé comme mesure étonnam, ce qui a permis de maximiser la sensibilité et la spécificité de l’outil. Ce test, développé sur une période de trente mois, permet de mesurer adéquatement le rendement en compréhension de l’écrit des membres des FAC anglophones terminant leur formation en L2. Il contribue ainsi indirectement à assurer le bilinguisme au sein des FAC.
1. The (re)construction of 4th year pre-service teachers’ practices about formative assessment and its impact on their learning to teach process

**Alexandre Mesquita**, Université de Sherbrooke, Alexandre.Alves.Mesquita@usherbrooke.ca

This study examines 4th year pre-service English as a second language (ESL) teachers’ lack of confidence to evaluate their pupils and its impact on their learning to teach process. Exploring and understanding how pre-service ESL teachers construct their knowledge about evaluation throughout their teacher education program allow teacher educators to make any necessary adjustments or improvements in their own courses and programs in order to provide the best preparation in terms of evaluation and allow pre-service teachers to graduate feeling prepared to evaluate. The theoretical underpinnings for this research are studies about a) pre-service teachers’ prior beliefs, b) practicum courses and c) formal education courses (teaching methods and formal evaluation courses). The data will be collected through the following methods: open-ended questionnaires, semi-structured interviews, stimulated recall and narratives. Preliminary results indicate that a series of contextual factors influence beginning teachers’ self-confidence and sense of preparedness to evaluate. Among these factors, we can mention Quebec’s recent reformed curriculum that has been implemented in the schools, including a competency-based socio-constructivist approach to teaching and learning that involved many changes to the way teachers were asked to evaluate pupils and the cooperating teachers’ uncertainty around evaluating with this new approach.

2. New directions: Investigating new tasks for integrated skills assessment

**Lia Plakans**, The University of Iowa, lia-plakans@uiowa.edu

For the past ten years, interest has increased in testing and teaching second/foreign languages skills (speaking, listening, reading, writing) in combination rather than as separated skills. A variety of assessment tasks have been developed and validated for integrated skills testing (Ascension, 2005; Gebril, 2009; Gebril & Plakans, 2009; Huang, 2012; Plakans, 2009) providing ample evidence for making inferences about the ability of using English for Academic Purposes (EAP). However, research and development has focused on using these tasks to assess performance-based skills, i.e. speaking and writing. Second language testing has not considered integration in the other direction, such as writing-to-read tasks or tasks that are bi-directional or more iterative, such as writing-reading-writing tasks. The proposed research study will explore these types of tasks as potential means to assess second language literacy in academic English. The study in progress investigates writing-reading or iteratively integrated skills tasks (writing-reading-writing), to consider how these new measures elicit the abilities considered a part of the second language literacy. The proposed research will explore whether these tasks tap ‘discourse synthesis’ that has been proposed as a construct for integrated L2 literacy assessment (Ascension, 2005; Plakans, 2009; Spivey, 1990; Yang & Plakans, 2012). In addition, there is potential to evoke aspects of an EAP literacy construct not captured by current tasks, such as scaffolding knowledge, reading comprehension, and vocabulary knowledge. This study is researching integrated skills tasks to determine if they provide useful information.
about language learner’s abilities, and if they can be supported by validity evidence for use in assessing English for academic purposes. If such tasks can be used effectively, language programs and testing companies may consider adopting them for assessment. Furthermore, such tasks can be used as a research instrument to study an area important to second language literacy: the connection between reading and writing abilities. Currently, a literature review is underway, and after the review phase, specifications for two assessment types will be designed: (1) a writing-to-read assessment and (2) writing-reading-writing iterative assessment. Once tasks have been outlined, their development will entail selecting topics and texts, developing prompts, building a rating scale, and piloting the tasks and scale. The results from the last step will be analyzed with revisions to the tasks and rating scale. Once a set of integrated tasks/assessment have been developed, the next stage is to investigate issues related to validity. The study will use both qualitative (think alouds and interviews) and quantitative (performances and scores) data collection and analysis to explore questions related to the products of these tasks as well as the process. The following questions will be investigated: 1. Are scores on writing-to-read, reading-to-write, and iterative reading-writing tasks related? 2. How are constructs of L2 reading and writing elicited in the new tasks? 3. Is there evidence of discourse synthesis processes in completing the new tasks? 4. What are the shared and divergent processes in test takers’ composing of new tasks?

3. Is true comparability possible?

India Plough, Michigan State University, ploughi@msu.edu

Is True Comparability Possible? The current state of foreign language education in the United States has been characterized as a “language crisis” (Berman, 2011). More and more colleges and universities are eliminating or lowering their foreign language (FL) requirements for admission and/or for graduation. Additionally, as entire FL programs are dismantled, a degree in certain foreign languages is no longer an option for students on some American campuses. In contrast to this trend, a relatively new residential college within a large land-grant university in the United States remains committed to making the study and use of a world language an essential part of its curriculum, as evidenced in its degree requirement of proficiency in a world language. In 2012, the college conducted a comprehensive review of its language proficiency program with an eye toward revisions that are built on the same methodological principles and thus contribute to the inherent interdependence of the processes of teaching, learning, and assessment. Two significant changes came out of that review and are currently underway: the adoption of a Culture and Languages across the Curriculum (CLAC) pedagogical approach and the development of a local test of language proficiency. The performance-based test of speaking proficiency employs a paired format and is intended to assess intermediate speaking proficiency in the more commonly taught as well as the less commonly taught languages. This Work in Progress focuses on the development of the proficiency scale. Using an intuitive method (Fulcher 2003), linguistic and interactional features of a preliminary scoring rubric have been assembled from level descriptors used in internationally accepted proficiency scales (e.g., ACTFL OPI, CEFR, CLB, ILR, TOEFL iBT). During test trials, examiners provide feedback on these features and also assist with revising level descriptors; in this way, empirical development of the scale is in progress. Test trials in English, French, German, and Spanish have been completed. Test trials in Turkish and Korean begin late Fall 2014. Preliminary analysis of transcripts of completed test trials as well as feedback from native speaker examiners and participants highlight the difficulty of using the same scale for multiple languages, some of which are quite diverse. The intricate and dynamic array of interactional and linguistic features that are important to effective communication in one language, culture, and context are not necessarily equally important, if at all, in another language, culture, and context. Specifically, initial analysis
of data in the current study brings into question the place traditionally held by syntactic complexity [as measured by Analysis of Speech Units (Foster, Tonkyn, & Wigglesworth, 2000)], in construct definition and related scale descriptors. Issues to be explored with LTRC colleagues include: How is test comparability maintained when adapting a test for multiple languages? When empirical evidence dictates variations in level descriptors across languages, at what point has the construct that the test is intended to assess been changed?

4. Professional language competency, the CELBAN and professional licensure

**Stefanie Bojarski**, Queen’s University, stefanie.bojarski@queensu.ca

The Canadian Nursing Association has raised concerns that as large numbers of Canadian nurses approach retirement age, a “critical shortage of nursing professionals” could result in Canada (Epp & Lewis, 2009, p. 286). The College of Nurses of Ontario (CNO) requires Internationally Educated Nurses (IENs) who are seeking membership to show evidence of language competency by successfully completing either the CELBAN or IELTS exams (CNO, 2013). Researchers of testing validity recommend an approach to “obtain[ing] additional information from test-takers for validating large-scale assessments” (Cheng & DeLuca, 2011, p. 104). Therefore this study will implement a series of focus groups and interviews immediately following the completion of the CELBAN in Ontario in order to analyse assessment experience, thus contributing to a wider understanding of construct validation from the perspective of test-takers. Given that the CELBAN is a relatively new testing format, with widespread use for high-stakes decisions (a component of certification and licensure), further research validating IEN-test-taker responses to construct representation is critical to our understanding of the role of competency testing in the certification and licensure of IENs. A semi-structured interview guide adapted from DeLuca, Cheng, Fox, Doe and Li (2013) will be used to direct data collection in both focus group and interview settings. Focus groups immediately following the completion of the reading, writing and listening section of the CELBAN will be conducted with IENs who are required to complete a language assessment (IELTS or CELBAN) as a component of their certification with the CNO. Individual interviews will be conducted with IEN test-takers immediately following their individual speaking assessments. Focus groups and interviews will be audio recorded and transcribed verbatim. Transcripts will be coded abductively via the theoretical framework derived from the Cheng & DeLuca (2011) study of test taker responses to English language proficiency exams as well as the 2013 study of the TOEFL iBT (DeLuca et al.). This research hopes to add to the critical understanding of what is tested by the CELBAN and the role of large scale language testing in the certification and licensure experience of IENs.

5. Using audio-only and video tasks for a computer-delivered listening test

**May Tan**, Canadian Defence Academy, Department of National Defence, may.tan@forces.gc.ca

**Nancy Powers**, Canadian Defence Academy, Department of National Defence, nancy.powers@forces.gc.ca

**Ruth Lucas**, Canadian Defence Academy, Department of National Defence, ruth.lucas@forces.gc.ca

**Roderick Broeker**, Canadian Defence Academy, Department of National Defence, rod.broeker@forces.gc.ca

The English Language Test Section of the Canadian Defence Academy is in the early stages of developing a new listening proficiency test for NATO military personnel which will be computer delivered. It will be a multilevel test developed to meet the specification of the NATO
Standardization Agreement 6001 language proficiency descriptors. This test will replace the older version of the listening test in which candidates listen to recorded texts and mark their responses on an answer sheet. In line with current ideas about the importance of authenticity and interactivity in defining construct validity (Bachman & Palmer, 1996; Ockey, 2007), this new listening test incorporates audio and video listening tasks that are representative of the target language use domain (Messick, 1996; Wagner, 2014). As part of the development process, candidates will be interviewed using a verbal feedback protocol to see how they perceive older version of the test versus the new computer-delivered trial version. Their reactions to audio-only versus video listening tasks within the new test will also be recorded. A Benchmark Advisory Test (BAT) for listening proficiency is available to NATO members by the Bureau for International Language Coordination. This is the recommended test against which other listening tests should be validated. However, the BAT uses audio only in its test tasks and constrains the number of times candidates get to listen to a text: the higher the level of difficulty, the fewer the number of replays allowed. At this point in time, the test development team is grappling with a few issues and is interested in initiating discussions around some of the concerns outlined below: - How can this newer version of the test be validated against the benchmark which uses audio-only tasks? Would a rigorous item development process and expert opinion (via a modified Angoff method) in determining the suitability and difficulty level of items suffice? - In contrast to traditional listening tests where students have very little control over the input, time and item sequence (no pausing, no replay, time for each item is pre-set, candidates must access items in linear order etc.) candidates may be allowed to control various aspects related to input and time with this new test. For each of these aspects, how much is enough for optimal candidate performance? How do these changes impact item and student performance? What are the consequences for construct validity?

6. Validity and validation in the Singapore context

Yun-Yee Cheong, The Institute of Education, University of London, ycheong@ioe.ac.uk

The most important consideration in assessment is validity and validation, the nature of which is complex and heavily contested. This thesis attempts to deconstruct the evaluation process into a series of more manageable questions grouped into three categories, namely the micro, meso and macro-level. Designed (with guidelines from Cambridge Assessment), set, scored, calibrated and offered twice a year (June and November) in Singapore, the Singapore—Cambridge GCE O-Level Chinese Language Examination (GCE 1162) has the largest number of candidates among all mother tongue language papers. At the micro-level, this ex post facto study intends to determine the degree of validity in using the GCE 1162 Paper 2 (Reading) for the purpose of demonstrating Chinese language reading achievement and proficiency. Newton and Shaw’s Neo-Messickian Framework (2014) will be employed to highlight critical areas for analysis, with measurement objectives being the main focal point of my research. The research design is then framed by the investigative steps proposed by Weir (2005). Special attention will be given to the cognitive and contextual aspects of validity evidence. At the meso-level, the discussion is taken forward by examining the evidence amassed within the broader context of Singapore. Since Messick (1989), the impact and washback effect of test use on institutions and society has been a significant component of test validation. There is, however, little discussion about how society can influence and shape the process of validation itself. The relationship is bidirectional rather than linear. The specific social, political, cultural and educational environment in which validation occurs will inevitably determine its feasibility and meaningfulness. The meso-level of analysis will therefore consider how these contextual factors can guide, steer or hinder validation practice, thereby strengthening (or weakening) our claims to validity. The micro and
meso levels are subsequently superimposed onto a theoretical plane. A fitting validation process will depend equally on an in-depth understanding of the prevalent validation frameworks today and the theories that underlie them. Most of the influential validation frameworks were conceptualised in Europe and the USA, calling their generalizability and transferability into question. The potential utility of different frameworks in the Singapore context will therefore demand investigation. The macro-perspective will not be a major focus of the research process; although, it is likely that useful insights will be gained from the attempt to apply a range of different frameworks. Discourse around assessment in Singapore has been dominated by issues of bias and fairness, assessment load and balance of assessment forms. Although various assumptions regarding validity are embedded in these matters, informed discussion about validity and concerted effort devoted to validation has assumed too low a profile. The first of its kind in Singapore, this study hopes to encourage genuine and open dialogue among persons involved with or invested in the examination process, including policy owners, specialists, administrators, students, teachers, parents and many others. It will also be of interest and relevance to researchers with an academic interest in educational testing, especially those examining the application of western validation frameworks in global settings.

7. Translation tests in China: Is there any opportunity for assessment for learning?

Manman Gao, Anhui University, gaomm03@gmail.com

Yuanyuan Mu, University of Science and Technology of China, yuanyuan_mu@hotmail.com

The last 20 years witnessed the flourish of translation tests in China, including CATTI, NAETI and SFLICE. Due to the increasing need of translation talents, more and more undergraduates and postgraduates have participated in these translation tests. Take CATTI (China Accreditation Test for Translators and Interpreters) as an example, the test takers of CATTI increased almost 10 times from 2003 to 2009. In the year of 2008, CATTI started to play a substantial role in the certification of MTI program (Master of Translation and Interpretation), that is, the postgraduates of MTI program in a number of universities can only graduate if they take the test and reach a certain level. By doing so, the government attempts to apply the test criteria to translation instruction. This may present an opportunity to promote students’ translation skills through assessment. Nevertheless, few empirical studies have been done on the washback effect of translation tests in China, not to mention tapping their potential in improving learning. In the pilot study, interviews with 20 test takers were conducted in 2014 to investigate their test motivation, perception, and preparation. Regarding test motivation, students admit that they are eager to obtain a certificate for future job-hunting, but many, especially those who are strongly interested in translation want to use this test to examine their English or translation proficiency, and more importantly, improve their skills. Most of interviewees took the test in the junior year when the curriculum burden was lighter. So they signed up for the test since they believed the preparation process could enhance their language and translation ability. Before the test, most of them spent at least one month in intensive preparation, which in students’ words remarkably improved their English and translation ability. Also for English majors, the intensive practice pushed them to acquire an in-depth understanding of those translation strategies taught in class. However, it is worth noticing that students paid no attention to the translation criteria, as they had no access to them. Since students tend to regard preparation for translation tests as one learning opportunity, the next step aims to incorporate AFL into translation instruction in class, to be specific, to make formative use of the summative translation tests. Case study will be conducted in two translation classes of MTI postgraduates who are required to take CATTI. Drawing on the test tasks and rating criteria of CATTI, the two teachers will every other
work-in-progress abstracts

Each month, two teachers and four groups of students will be invited for interviews. The study attempts to explore effective ways to integrate assessment with learning in translation instruction, as well as possible adaptations of translation tests. Over the course of the past two decades, the quality of the English components of four-year technical and vocational education joint college exams in Taiwan has improved to a remarkable degree, shifting emphasis from discrete linguistic knowledge to real-life language application. Nevertheless, English tests such as these have been criticized due to the fact that they tend to focus on the evaluation of reading skills and utilize multiple-choice questions, and thus are unable to assess the communicative competence of students. In view of this criticism, beginning in 2015, there will be an essential change to the administration of the joint vocational college English exam, with the addition of a non-multiple-choice writing assessment. On this English test, controlled writing tasks (e.g., vocabulary assessment tasks, ordering tasks, translation tasks, and short-answer and sentence completion tasks) will account for one-fourth of the total score. It is therefore the goal of this study to explore the washback effects that this major change to this high-stakes test will have on vocational high school English teaching and learning. This study is being undertaken for the following reasons: 1) the lack of empirical washback studies on high-stakes “writing” tests in the field language assessment, 2) minimal extant washback research on the interaction of tests, teaching, and learning, and 3) insufficient knowledge of whether washback varies from the different timings of any investigation. To bridge this research gap, it is the intent of this study to investigate particular aspects of washback effects (Shohamy et al., 1996; Watanabe, 1997; Messick, 1989; Cheng, 2005; Green 2007) including: 1) general and specific washback; 2) positive and negative washback; 3) strong and weak washback; 4) intended and unintended washback; and 5) long-influence and short-influence washback. One example is whether or not the aforementioned writing assessment influences the practices of both teachers and students in the course and therefore improves students’ communicative competence in real-life writing. Another example is the extent to which the test format, test task, the weight of the test, and its purpose influence writing instruction and learning. In addition, whether or not students’ proficiency levels, their years of study, and their teachers’ professional development play a role in generating the five aspects of washback as mentioned above, and what these aspects look like, are all areas worthy of further exploration. Instruments utilized in this study include questionnaires, interviews, and classroom observations. Sixty English teachers and 600 business students of different years at three vocational high schools in Taiwan (the required entrance scores of which are ranked top, medium, and bottom) are being recruited. The findings drawn from this study will generate a better understanding of the complexity of the effects of washback on teaching and learning. Furthermore, this research will provide teachers with valuable information necessary for preparing students for tests and for facilitating English instruction and learning objectives so as to elicit positive and eliminate negative effects. Finally, test developers can utilize evidence drawn from this study as one key component of test validation.

8. The washback effects of the high-stakes writing assessment in Taiwan

Yi-Ching Pan, National Pingtung University, yichingpan@yahoo.com.tw

Over the course of the past two decades, the quality of the English components of four-year technical and vocational education joint college exams in Taiwan has improved to a remarkable degree, shifting emphasis from discrete linguistic knowledge to real-life language application. Nevertheless, English tests such as these have been criticized due to the fact that they tend to focus on the evaluation of reading skills and utilize multiple-choice questions, and thus are unable to assess the communicative competence of students. In view of this criticism, beginning in 2015, there will be an essential change to the administration of the joint vocational college English exam, with the addition of a non-multiple-choice writing assessment. On this English test, controlled writing tasks (e.g., vocabulary assessment tasks, ordering tasks, translation tasks, and short-answer and sentence completion tasks) will account for one-fourth of the total score. It is therefore the goal of this study to explore the washback effects that this major change to this high-stakes test will have on vocational high school English teaching and learning. This study is being undertaken for the following reasons: 1) the lack of empirical washback studies on high-stakes “writing” tests in the field language assessment, 2) minimal extant washback research on the interaction of tests, teaching, and learning, and 3) insufficient knowledge of whether washback varies from the different timings of any investigation. To bridge this research gap, it is the intent of this study to investigate particular aspects of washback effects (Shohamy et al., 1996; Watanabe, 1997; Messick, 1989; Cheng, 2005; Green 2007) including: 1) general and specific washback; 2) positive and negative washback; 3) strong and weak washback; 4) intended and unintended washback; and 5) long-influence and short-influence washback. One example is whether or not the aforementioned writing assessment influences the practices of both teachers and students in the course and therefore improves students’ communicative competence in real-life writing. Another example is the extent to which the test format, test task, the weight of the test, and its purpose influence writing instruction and learning. In addition, whether or not students’ proficiency levels, their years of study, and their teachers’ professional development play a role in generating the five aspects of washback as mentioned above, and what these aspects look like, are all areas worthy of further exploration. Instruments utilized in this study include questionnaires, interviews, and classroom observations. Sixty English teachers and 600 business students of different years at three vocational high schools in Taiwan (the required entrance scores of which are ranked top, medium, and bottom) are being recruited. The findings drawn from this study will generate a better understanding of the complexity of the effects of washback on teaching and learning. Furthermore, this research will provide teachers with valuable information necessary for preparing students for tests and for facilitating English instruction and learning objectives so as to elicit positive and eliminate negative effects. Finally, test developers can utilize evidence drawn from this study as one key component of test validation.
9. Improving Cree students’ reading comprehension through formative assessment practices and explicit teaching.

Maria-Lourdes Lira Gonzales, Université du Québec en Abitibi-Témiscamingue, maria-lourdes.lira-gonzales@uqat.ca

The Cree School Board [CSB] has jurisdiction for elementary, secondary and adult education for members of the Cree Nation of Eastern James Bay and other people resident in one of its nine communities. Numerous reports have documented serious shortcomings in almost every aspect of service delivery of instruction in Cree (level of understanding of bilingual models of education, language development and acquisition in general, effective literacy approaches, and curriculum development). These shortcomings have an impact in the performance of the CSB students who have remained well below the Canadian norm in all academic areas. According with the CSB annual report (2013-2014), 15% of Grade 6 students achieved a Stanine 4 on the Canadian Achievement Tests (CAT), indicating that they are meeting end of grade expectations. This is a decrease of 3% from the 18% who were achieving at the expected level in 2012-2013. The target set for 35% has not been met for 2013-2014. Therefore, there is still work to do in order to close the gap between Cree student achievement and that of other students in Canada. The goal of the present study is to improve Cree students’ reading comprehension – and therefore CAT results- through formative assessment practices (Black et William, 2009; Hattie et Timperley, 2007; Isaacs, 2001; Rodet, 2000) and explicit teaching (Gauthier, C. et al. 2013; Tardif, J., 1997). Three research questions are posed: 1. How to lead teachers to develop reading strategies in English as a second language (ESL)? 2. Which are the types of feedback teachers provide to enhance learning during reading strategies lessons? 3. What are the self-competency perceptions before and after the teacher training sessions? This collaborative action research (Desgagné et Al. 2001; Desgagné, 1997; Maxwell, J. A., 2005; Creswell, 2014) will take place in two Grade 6 classes. The participants will be two ESL grade 6 teachers. Data will be collected through class observation (Laperrière, 1997), video and audio recordings (Flick, 2002) as well as semi structured interviews (Vermersch, 2006). This study presents three phases; each of them has the same structure: (1) participating in a training session, (2) applying in class what has been learnt during the training session and (3) receiving feedback on their class performance.

10. An investigation into alignment of Lexile scores and analytical rubrics

Roxanne Wong, City University of Hong Kong/University of Jyväskylä, elroxann@gapps.cityi.edu.hk

Carl Swartz, MetaMetrics, cswartz@lexile.com

The Diagnostic English Language Tracking Assessment (DELTA) is a low-stakes multicomponential online diagnostic assessment designed specifically for undergraduate students in the Hong Kong educational context. Students are given a detailed report on their overall proficiency and performance in reading, listening, grammar, and vocabulary. The DELTA currently does not have a writing component. One of the key features of the DELTA writing is that it will be using an Automated Essay Scoring system. This is essential as scoring of writing papers is labor intensive as well as expensive. Between 2012 and 2014 a writing component has been designed and trialed. Approximately 1500 university level students in Hong Kong participated in the writing pilot. The pilot test used MetaMetrics the EdSphere Lexile Framework for writing, an online platform where students can immediately receive a Lexile score on their work. The Lexile program was chosen because like the DELTA, it has a built in tracking function that can record and report on students progress. While the students are given a Lexile score upon completion of their writing, it is
not enough for diagnostic assessment. According to Alderson (2005), “A crucial component of any diagnostic test must be the feedback that is offered to users on their performance. Merely presenting users with a test score is quite inappropriate on diagnostic tests.” In a joint project with MetaMetrics and one school district in the United States, a similar test was conducted with 539 students in upper secondary school. These tests were then scored simultaneously by members of the DELTA team, staff from the MetaMetrics corporation, and teachers from the school district. The scoring instrument was specifically designed for the DELTA and has the same general domains as the EdSphere self-assessment for writing. This study investigates the use of the DELTA rubric for scoring, and the initial phases of feedback design using a combination of the Lexile system and the scoring rubric. This study aims to look at the reliability of the scoring instrument as used by a variety of stakeholders in a cross-national context. The instrument designed for DELTA is an empirically based analytical scale which has six domains, namely: Task fulfillment, Academic register and stance, Organization, Semantic complexity, Grammatical accuracy and Vocabulary. It was initially designed for formative feedback purposes. Weigle (2002) indicates that Analytic scales are more appropriate for L2 learners to develop their writing ability as did Knoch (2011). The main purpose of this study is to investigate the ability to provide students with diagnostic feedback using analytical scales. The Lexile score provided to students has no easily discernible function for most students and therefore, the rubric is being trialed to determine the usefulness for diagnostic feedback purposes. The steps taken to ensure reliability and trialing of the rubric in this context will be discussed.

11. Improving language teaching and learning through language testing: A washback study on the Hanyu Shuiping Kaoshi (HSK)

Shujiao Wang, McGill University, shujiao.wang@mail.mcgill.ca

The rapid growth of China in the global economy has increased the need for the Chinese language. As a result, the number of learners studying Chinese as a second/foreign language (CSL) has reached 40 million all over the world. Hanyu Shuiping Kaoshi (HSK), literally “Chinese Proficiency Test,” also known as the “Chinese TOEFL,” is the central national standardized test of Chinese language proficiency for non-native speakers, and it plays a vital role in certifying language proficiency for higher education and professional purposes. Despite the significance and the claimed high validity and reliability of this test (Chen, 2009; Luo et al., 2011), very few empirical studies related to its validity have been conducted, in particular, with regards to the washback effect (Huang, 2013; Huang & Li 2009). The research questions of the on-going study are: 1) What are the perceptions of CSL teachers and learners towards the HSK content, use and impact? 2) What are the relationships between learners’ test-taking expectations, test preparation practices and test outcomes? 3) What is the nature of and intent of washback from the HSK and 4) Within the educational and societal context, how do HSK stakeholders and score users interpret and react to washback? These questions are addressed through a mixed methods sequential explanatory design to provide a holistic account of the phenomena under investigation (Creswell, 2009; Creswell & Plano Clark, 2011). In Phase 1, participants are 300 HSK test-takers who register for the HSK test and 30 CSL teachers from 5 countries. The survey questions include the learning/teaching strategies and the perceptions towards the test design, test preparation, test expectation and test use. In Phase 2, after taking the HSK test, the test-takers are asked to report their test scores by category.
Quantitative analysis in these 2 phases includes descriptive statistics and item-level exploratory factor analysis, confirmatory factor analysis and structural equation modeling. In Phase 3, individual semi-structured interviews will be conducted with 10 HSK stakeholders from various education (e.g., student, teacher, institution), government and business contexts on their test score interpretations and uses. In the end, both types of data sources will be cross-examined and synthesized. This research is the first washback study on the HSK from the perspective of multiple stakeholders at both the micro (classroom) and macro (society) levels. It obtains multiple perspectives on washback effects of high-stakes tests, not only provides findings that are applicable to pedagogical and methodological issues of CSL teaching and learning, but also contributes a more comprehensive model to enrich the existing knowledge in washback literature. Last but not least, this study has social and significant implications for the reform of the HSK and score users.

12. Lexical performance and rater judgements of English language proficiency: Implications for teaching, learning, and assessment

Scott Roy Douglas, University of British Columbia, Okanagan Campus, scott.douglas@ubc.ca

This report outlines an ongoing project exploring lexical validity and standardized English language testing in which lexical evidence extracted from test-takers’ responses to speaking and writing prompts on the CELPIP-General Test is gathered to contribute to an understanding of what is represented by that test’s scores. Lexical validity is conceptualized as the uncovering of evidence supporting the recognition that vocabulary elicited by a standardized test’s prompts unfolds as expected against that test’s rating scale. The CELPIP-General Test test-takers are scored for each component of the test on a scale from 0 to 12. The CELPIP levels correspond to assigned Canadian Language Benchmark (CLB) level equivalencies. The current research project is a timely response to the ongoing need to contribute to evidence supporting the overall validity of high stakes standardized English language proficiency tests that are used to provide evidence of English language abilities needed for high stakes immigration and citizenship decisions. It also points the way to incorporating appropriate vocabulary targets for candidates preparing for the CELPIP-General Test. For the current study, a corpus is being built of approximately 200 speaking samples and 500 writing samples from the responses to the CELPIP-General Test. Quota sampling is employed to gather from test-taker samples available as even a number as possible of samples from each of the different CELPIP levels of proficiency. To analyse breadth of vocabulary usage, lexical profiles are generated for each individual oral and written text using digital corpus analysis tools (Cobb, 2014). Profiles consist of measures of lexical frequency based on the percentage coverage of text by the General Service List (GSL) (West, 1953), the Academic Word List (AWL) (Coxhead, 2000), text not covered by either of those lists, and the percentage text covered by the British National Corpus (BNC) (Cobb, 2014). To analyse depth of vocabulary usage, Vocabulary Error Ratios (VER) are calculated based on the detailed coding of the oral and written texts. The VER is calculated for each text by the following formula: (total vocabulary errors/total words)*100. On compilation of the measures of lexical sophistication in the corpus as a whole, data are correlated using the product moment correlation coefficient (Pearson r) and scatter plots are generated for the relationships under investigation as a preliminary analysis. Next, hierarchical multiple regression analysis is employed to determine the predictive value of the measures of the independent variables of lexical performance and the dependent variable of the determined CELPIP levels of proficiency. This analysis is carried out to determine which of the lexical variables under analysis offer the most predictive value for CELPIP rating scale performance. Once
the relationship between measures of lexical sophistication and rater judgements of test-takers’ performance in comparison to the CELPIP levels of performance is explored, the current study turns to the implications for teaching, learning, and assessment. In particular, implications connected to how vocabulary in use unfolds over the CLB descriptors can be postulated along with the concomitant implications for teaching and learning within the CLB framework.

13. EFL teachers’ professional knowledge of assessment

Hossein Farhady, Yeditepe University, Istanbul, hossein_farhady@yahoo.com

Kobra Tavassoli, Karaj Branch, Islamic Azad University, Karaj, kobitavassoli@yahoo.com

Many scholars have recently addressed the importance of teachers’ professional knowledge of assessment from different perspectives (Inbar-Lourie, 2013; Malone, 2013; Popham, 2009, 2011; Stiggins, 2005, 2008; Taylor, 2009). At a minimum level, it is often referred to as “assessment literacy”, and defined as “the knowledge, skills and abilities required to design, develop, maintain or evaluate large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice” (Fulcher, 2012, p.125). Despite some good quality research on the topic (Çakan, 2004; Razavipour, Riazi, & Rashidi, 2011), findings are not conclusive on the extent to which teachers’ professional knowledge of assessment would exert positive effect on students’ learning and achievement. This paper is a progress report on a three-phase research project designed to investigate teachers’ professional knowledge of assessment and its effect on teacher-produced classroom tests as well as students’ achievement. The first phase of research was an attempt to conduct a needs assessment and investigate teachers’ perception of the importance of their professional knowledge of assessment in an EFL context. A modified version of the professional knowledge of assessment needs questionnaire, developed by Fulcher (2012), along with some open-ended questions on teachers’ needs was administered to 143 EFL teachers in Iran. Preliminary analysis showed that the majority of EFL teachers considered the major topics in language testing as either important or essential to be included in language testing courses in teacher education programs. In the second phase, EFL teachers’ claims about their professional knowledge of assessment will be checked against their real performance on a professional knowledge test. Further, the effect of certain factors such as teachers’ age, gender, experience, and academic degree on their professional knowledge of assessment will also be investigated. Qualitative data will be collected from first observing the classes of teachers who participated in the first stage of research and then interviewing them. The purpose is to investigate the extent to which they apply their assessment knowledge to their classroom assessment and whether there is a relationship between teachers’ professional knowledge of assessment and the quality of their tests and students’ achievement. In the third and final phase of the study, the procedures will be extended to non-EFL teachers to compare and contrast their assessment professional knowledge with those of EFL teachers. The purpose is to inform the education community that all teacher education programs may need to reformulate and refine their treatment of assessment and pay more attention to improving teachers’ professional knowledge in all disciplines. Further, comparative studies will be performed in some other countries such as Turkey, Malaysia, and UAE to investigate the levels of professional knowledge of teachers in different countries. The findings are expected to have contributions to our understanding of the significance of teachers’ professional knowledge in general, and in language education in particular. They are also expected to offer guidelines for reforms in planning in-service trainings for teachers in all major disciplines.
14. The washback of ENEM EFL test in Brazil

Flávia J. de Sousa Avelar, University of Campinas, SP; CAPES Foundation, Ministry of Education of Brazil, Brasília – DF 70.040 – 020., avelarfj@yahoo.com.br

Matilde V. R. Scaramucci, University of Campinas, SP, matilde@scaramucci.com.br

The washback of ENEM EFL test in Brazil This work-in-progress session will discuss preliminary results of a study which aims at investigating the washback of the EFL exam which is part of ENEM (Exame Nacional do Ensino Médio or High School National Examination), implemented by the Brazilian Ministry of Education in 1998. Initially proposed as a diagnostic test to be taken by students leaving high schools to assess the quality of education at this level, the exam had its function and status changed in 2009, and became a mandatory high-stakes entrance examination for high school students who intend to enter federal universities. The change was justified as a means to improve the quality of education and at the same time to make access to public university easier. Entrance exams (or “vestibulares”) have had a long tradition in Brazilian educational system. The ENEM comprises 180 multiple choice questions divided in four areas: Natural Sciences (Biology, Chemistry and Physics), Human Sciences (History, Geography and Philosophy), Mathematics and Languages (a written essay in Portuguese and either English or Spanish reading test). Although English starts at grade 5 of primary school, an English test was only included in the exam in 2010. Despite the relevance of the new exam and the intended purpose of using it to influence high school teaching and learning, few studies have been carried out in order to gather evidence of its impact. Therefore, in an attempt to contribute to EFL teaching policies in Brazilian public schools, this ongoing study aims at investigating the washback of this EFL reading test on the perceptions and practices of two teachers from different high schools, as far as their approach, methodology and teaching sequence are concerned. It also provides information for further discussions of the social role and relevance of assessment in educational contexts. In each one of the schools two groups were observed throughout 2013. Aiming at data triangulation, information was gathered through interviews with the two teachers, school coordinators, classroom observations and field notes. Questionnaires were answered by students and analyses of textbooks and performance tests used by the teachers complement the data. Preliminary results suggest that apart from the relevance of ENEM, teachers and students from both private and public schools reacted differently to the exam. In the private school context, the influence of the traditional entrance exam (vestibular) is much more intense than the ENEM itself. In the public school context, on the other hand, although the teacher seems be familiar with the EFL exam, classroom observation shows that test wiseness was the main focus of the classes, suggesting a negative washback of the exam upon her teaching.

15. Understanding essay rating activity as a socially-mediated practice

Yi Mei, Queen’s University, yi.mei@queensu.ca

This study aims to understand raters’ essay rating activity (RA) as mediated by its sociocultural context of a high-stakes university entrance examination in China, the National Matriculation English Test (NMET). RA has long been considered a cognitive process of individual raters interacting with artifacts (e.g., essays, rating scales; e.g., Bejar, 2012; Crisp, 2010), but this decontextualized approach fails to consider potential interactions between individual raters, sociocultural contexts, and their influences on rating processes. RA is a socially motivated practice with social meanings and consequences, and situating it within its sociocultural contexts can make findings more meaningful (Barkaoui, 2008). A comprehensive understanding of RA has implications to rater training and fair assessment.
I apply cultural-historical activity theory (CHAT; Engeström, 1987, 2001) from sociocultural theory to reconceptualize RA as a 3-level hierarchical, socially mediated process with collective object-oriented activity (top), individual's goal-directed actions (middle), and individual's automatic operations (bottom). The distinction between the top and middle levels is critical: The central RA (top), in its networked relationships to other interconnected activities, is the primary unit of analysis, while individual's actions (middle) are the subordinate unit of analysis, which are more understandable when interpreted against the background of collective activity (Engeström, 2000). In an examination-driven society such as China’s, the consequences of NMET and the role of high school English teachers as NMET raters make them good candidates on understanding how their NMET RA is socially mediated. Two research questions (RQs) will guide this study: 1) What is the nature of raters’ NMET rating experiences? 2) How do NMET raters assess NMET essays to achieve their goals, against the background of their NMET rating experiences? The study will employ a three-phase qualitative case study design (Yin, 2009). In Phase 1, in one provincial NMET centre, I will interview the centre’s director, 4 trainers/team leaders, and 8 NMET raters with varying teaching and NMET rating experience, to investigate NMET raters’ central NMET RA. In Phase 2, I will interview the NMET raters, their principals, and 16 of their colleagues to investigate NMET rating-related activities at school. NMET rating-related documents will also be reviewed for additional contextual information. In Phase 3, the NMET raters will first conduct concurrent think-aloud protocols (Ericsson & Simon, 1993), followed by consecutive stimulated recalls (Gass & Mackey, 2000), to verbalize their thought processes while assessing 8 NMET sample essays. Lastly, individual interviews will be conducted on the raters’ specific goals of assessing NMET essays and perceptions of the NMET rating task. A constant comparative method (CCM; Corbin & Strauss, 2008) will be used to analyze Phase 1 and 2 data, to describe central NMET RA and its interconnected activity in school context (RQ1). Phase 3 data will also be analyzed using CCM but also conceptually informed by Cumming, Kantor, and Powers’ (2002) descriptive framework of rater decision-making behaviors, to describe raters’ specific goal-directed actions and situating them within collective activity experiences (RQ2). This study hopes to make an initial attempt at determining the potential value of a CHAT-based approach.

16. The development and application of cognitive diagnostic model for reading assessment

Junli Wei, University of Illinois at Urbana-Champaign, wei5@illinois.edu

A ssessing students’ achievement and providing constructive feedback is a critical responsibility of teachers. The Cognitive Diagnostic Assessment (CDA) provides valuable diagnostic profile information about students’ underlying knowledge and cognitive processing skills required for answering problems, thus the CDA has become popular in the testing field. However, most current empirical studies on the CDA rely on coding and analyzing existing items for cognitive attributes (Gorin, 2006), although the items were not originally intended for CDA use (Buck & Tatsuoka, 1998; Jang, 2005). Few efforts have been made to construct new tests considering the CDA, hence few guidelines are available for developing cognitive diagnostic language assessments (Lee & Sawaki, 2009). The present study employs cognitive diagnostic approaches to develop a diagnostic test model for fine-grained assessment, and classification of learners’ knowledge state and presence or absence of certain attributes. The study also develops a validation argument for the constructed CDA, and addresses the diagnostic report and corresponding remedial guidance. Questions addressed are: What attributes count in reading comprehension and how they are related hierarchically? How can we construct and validate the reading CDA model? What are the considerations in presenting the diagnostic reports? How can we guide the learners based
on their diagnostic reports? This study adapts the Evidence Centered Design paradigm (Mislevy & Riconscente, 2006) to describe a framework for the entire implementation process for diagnostic assessment. Participants are experienced test developers, content experts, teachers, and more than 2,000 students from senior high schools in China. The study consists of three steps: First, propose a preliminary diagnostic model through literature analysis and analyze a pilot study on it, including attribute identification, initial Q-matrix construction, Item Response Theory (IRT) and CDA analysis. Second, propose and validate the hypothetical model by comparing test developers’ judgment, examinees’ Think Aloud Protocol, and domain experts’ evaluation. The G-DINA model is psychometrically employed to analyze the test responses and the hypothetical model to find out the absolute and relative model-data fit statistics. The best-fit model will then be verified with new data and its goodness of fit will be confirmed. Third, diagnostic reports will be proposed, and followed by guidance for learners to cope with their diagnosed problems. Quantitative and qualitative information is collected to develop a validation argument for the composed CDA. The first two steps are completed. Preliminary findings identify six attributes for reading comprehension and suggest that the CDA method is scientific and feasible. It is expected to be verified by later analysis that the CDA diagnostic report can facilitate teaching and learning. Combining assessment of learning with assessment for learning has become a global trend in the education field. Complementing this trend, the study suggests that CDA is an effective way to achieve this goal by elaborating on practical aspects of cognitively based assessment design and application. In addition, the use of diagnostic reports for providing feedback would allow teachers to identify the learner’s specific deficiencies and to plan individualized instruction. It also facilitates the learners on their journey to autonomy.

17. Construction of rating scales for the assessment of peer-to-peer interactive speaking performance

Romulo Guedez-Fernandez, The University of the West Indies, romulo.guedez@sta.uwi.edu

In the Angophone Caribbean, Spanish plays an important role, both in high school education and as a desired communication tool/asset. This study forms part of on-going research aimed at developing rating scales for the assessment of interactive speaking performance for an advanced level Spanish course at the university level in an English speaking Caribbean country. Participants (n=40) are Spanish majors and minors who are in their final year of their degree. All of the participants have received at least nine (9) years of Spanish language instruction. In this academic context, Spanish language instruction and the respective curriculum follow the Common European Framework of Reference (CEFR) level C1. Classroom teaching activities include discussions and debates on complex topics such as politics, immigration, gender, etc., as these types of tasks demand more interactive ability of test takers at the level C1. Similar types of tasks are used in each of the two peer-to-peer oral assessments during the semester. These two tests are assessed by one rater who does not participate in the test, and one rater as interlocutor with minimal participation during the test. The set of descriptors provided by the CEFR’s rating scales for informal/formal discussion and conversation for the C1 level provides allowances for a wide range of discrepancies for similar test performances either when raters allocate marks during the rating process or when these performances are re-rated. For this specific context of teaching and assessment, the performance data-driven approach has been chosen in order to develop rating scales for the assessment of peer-to-peer interactive speaking performance. Feedback sessions on participants’ performance will be provided after the test using think-aloud protocols. Students’ performances, focus groups
with raters and teachers, feedback sessions as well as raters’ verbal reports while assessing students’ performances will be video-recorded, transcribed and further analysed. Not only will the data collected allow us to arrive at a more robust definition of the construct of interactional competence for this particular speaking test, but also enable the operationalization of this construct, provide more specific descriptors for the new rating scales, and facilitate the validation process. In the Work in Progress session data from this study will be presented and feedback about methods of analysis and interpretation will be sought.

18. Designing a post-entry language assessment for international students in a business administration co-operative education program

Christine Doe, Mount Saint Vincent University, christine.doe@msvu.ca

This work in progress outlines a research study that examined one business administration co-operative (co-op) education program at a Canadian university as an initial construct development of a Post-Entry Language Assessment (PELA) for international students. In other university contexts, a PELA based approach typically draws on diagnostic assessment to provide feedback and offer targeted support to English as Additional Language (EAL) students with the aim of enabling their successful completion of academic programs (See for example Fox & Haggerty [2014]; Read & von Randow [2013]). Co-op education programs offer university students the opportunity to gain valuable workplace experience. A challenge facing business administration co-op programs is preparing domestic students to obtain the cultural and situated workplace knowledge needed (Cord & Clements, 2010; Jones, 2007). International EAL students may face additional challenges related to culture and communication skills (Young & Schartner, 2014). Background. The intent of the study was to identify the strengths and weaknesses of EAL students in a business administration co-op education program to inform the construct development of a PELA. Previous research has identified course-based challenges international EAL students encounter in business-related programs (Foster & Stapleton, 2012) and the strengths the students bring to the programs (Macdonald, 2013), but such research has not been extended to workplace learning situations. In engineering, Fox & Haggerty (2014) conducted a needs analysis of EAL students by drawing on multiple data sources, including interviews with various stakeholders. Based on the needs analysis, an existing PELA was refined to inform targeted instruction for an English for Academic Purposes (EAP) course for engineering students (Fox & Haggerty). Methods. Taking a case study approach (Yin, 2009) this study was theoretically grounded in Davidge-Johnston’s (2007) framework for evaluating co-op education curricula as well as Read and von Randow’s (2013) approach to development of a PELA in a New Zealand university context. Interview data was collected from 5 stakeholder groups, including, 4 Business Administration faculty members, 5 potential employers, 1 co-op placement coordinator, 2 EAP course instructors, and 8 students (from Saudi Arabia and China). Drawing on theme and pattern analysis (Creswell, 2012) potential barriers (i.e., cross-cultural miscommunication) and strengths (i.e., intercultural competence, additional language) were identified. Based on preliminary results, interactive oral communication was identified as a fundamental skill for success in the co-op program. In addition, it was recommended that the students be exposed to case analyses earlier in the program and have an EAP instructor and business administration faculty member co-teach EAP courses. Presentation. In the work in progress presentation, I will discuss the current literature on PELA and co-op education for international EAL students, the methods used, and the preliminary findings and interpretations. This study is a pilot study for a larger-scale study that will include multiple institutions and a mixed methods design. I welcome comments and suggestions from the
19. Planning for washback in the development of a large-scale English language proficiency assessment

Kevin Joldersma, SLTI, kevin.joldersma@gmail.com
Karen Feagin, SLTI, kfeagin@2lti.com
Charis Walikonis, SLTI, cwalikonis@2lti.com
Catherine Pulupa, SLTI and University of Maryland, cpulupa@2lti.com
Jennie Norviel, SLTI, jnorviel@2lti.com
Mika Hama, SLTI, mhama@2lti.com
Toshi Watanabe, Benesse, t-watana@mail.benesse.co.jp

Washback effects are often believed to be negative. Buck (1988), commenting on the inability of many Japanese high school graduates to perform basic communicative functions after years of English language instruction, blames grammar- and translation-focused testing and the instruction practices that result. But according to Swain (1985), the washback effects of a well-designed test should be positive. Buck (1988) offers a solution to the problem in Japan: “If we want our students to learn to communicate in English, then we must give them tests which require them to process communicative English” (p. 18). In light of the need for tests which assess communicative English skills, a multinational group is currently developing a large-scale assessment at the university admissions level. Currently, the test is launching for administration in Japan to Japanese students seeking university admission, with plans for future expansion after continued development and research. It was designed within the framework of Task-Based Language Assessment (TBLA) and in alignment with the Common European Framework of Reference (CEFR). It tests four skills (reading, listening, speaking, and writing) and targets specific knowledge, skills, and abilities (KSAs) that students would use in real-life social and academic contexts. The KSAs were derived from syllabuses and aligned with the “Can do” statements from the CEFR. All items are designed to provide the test-taker with an authentic context in which s/he would interact with the item. This presentation will address multiple aspects of washback effects related to the development and implementation of the new large-scale assessment. There is already evidence of washback in Japan, thanks to end-users including universities, students, and the Japanese Ministry of Education. Assuming continued interest from the Ministry, there is the potential for shaping future education policy in Japan. This study will aim to assess washback effects that are already taking place, as well as make recommendations for curriculum development, instruction, and education policy within Japan. Specifically, we will describe the preliminary steps of our research into the washback effects of the assessment. We hope to use this research to plan for washback, continue to develop the test with washback in mind, develop teacher training resources aligned with our assessment framework, and plan research methodology for addressing washback effects after the test goes live. We hope to gain, from this work-in-progress session, input from the audience on our proposed methodologies for investigating washback of the assessment as well as suggestions for alternative approaches or sources of data collection.

Jian Xu, Chongqing University, 857119178@qq.com
Xiangdong Gu, Chongqing University/Cambridge English, xiangdonggu@263.net

One of the most common testing methods used in foreign language teaching is translation (Buck, 1992). Although it is a very common testing method, there has been very little research to indicate how effective a testing method translation really is (Buck, 1992; Ito, 2004). House (2009) ever put forward a system for analyzing and evaluating texts for translation purposes. As noted by Klein-Braley (1982), it is yet not clear what translation is supposed to measure. The National College English Test Band 4 (CET-4 hereafter) has adjusted its Chinese-English translation from sentence format (5% of its total score), to a short passage format (15% of its total score) since 2013. However, the test syllabus does not give a clear construct definition about this task, so it is difficult for teachers to teach and for students to prepare for passage translation in and out of class. Thus we think there is a need to conduct an empirical study to investigate what passage translation task actually measures in CET-4. Driven by Bachman and Palmer’s (1996) description of language ability, we tend to examine test-takers’ processes of handling the CET-4 Chinese-English passage translation task by adopting think-aloud protocol analysis and retrospective interviews. Six participants (3 males and 3 females) are involved in our exploratory study. The instrument used is a CET-4 passage translation task in December, 2013. The preliminary findings suggest that vocabulary and grammar are two main elements assessed by the task. As to vocabulary, the high group tends to choose the most appropriate English words or try to find synonyms as a compensation for memory loss while the low group may ignore the unfamiliar words. As to grammar, the two groups generally focus on single or plural forms, tenses and use of conjunctions with no significant difference. Most test-takers translate word by word and then synthesize them into a sentence. The most frequently used translation strategies are finding synonyms as substitute, reading, repeating, revising, reasoning, and monitoring. Significant differences of strategy use between the two groups are identified. Besides, test-wiseness strategies and construct-irrelevant factors are detected in our study. Based on the exploratory study above, we plan to design a questionnaire with items of translation competence and strategies, and ask 300 college students to tick the items they use when taking a CET-4 passage translation task in order to further reveal what the passage translation task measures exactly. The study intends to provide helpful suggestions for better classroom translation instruction and practice in and out of class. We hope the research findings may function as a bridge connecting translation assessment with translation instruction and practice in addition to bear implications for passage translation test development.
INDEX OF PRESENTERS

Acedo Bravo, Alberto Enrique 33, 63, 39, 83-84
Alderson, Charles 127-128
Arne Johannes, Aasen 34, 101-104
Bai, Ying 34, 62, 131-132
Baker, Beverly A. 34, 60-61
Ballard, Laura 34, 101-104
Bärenfänger, Olaf 130
Baten, Lutgarde 33, 63, 124-125
Battistini, Laura 40, 98-99
Bégg, Kaitlyn 34, 101-104
Berry, Vivien 33, 54-55
Bi, Nick Zhiwei 39, 89
Billington, Rosey 36, 76-77
Bitterman, Tanya 137-138
Bojarcki, Stefanie 34, 65
Boraie, Deena 40, 93-94
Brau, Maria 33, 55-56
Broeker, Roderick 147-148
Brookhart, Susan M. 8, 9, 32
Brooks, Rachel 33, 55-56
Cai, Hongwen 34, 60-61
Carlson, Cecile Hamnes 136-137
Caudwell, Gwenwydd 40, 93-94
Chapman, Mark 32, 50-51
Charest, Marie-Soleil 141
Chen, Michelle Yue 33, 48-49, 140
Cheng, Lijing 35, 37, 80-81
Cheong, Yun-Yee 148-149
Chi, Youngshin 30-31, 42-43
Cho, Ikkyu 30-31, 42-43
Chung, Sun Joo 35, 66
Colby-Kelly, Christian 35, 66
Cumming, Alister 11, 35
Dang, Trang 30-31, 35, 42-43, 142
Davidson, Fred 41, 119-122
Davis, Lawrence 41, 119-122
de Sousa Avelar, Flávia J. 39, 85
Deane, Paul 33, 63, 124-125
Deygers, Bart 142
Diaz Moreno, Yoennis 142
Do, Ha 158-159
Doe, Christine 33, 47-48
Donovan, Anne 153-154
Douglas, Scott Roy 33, 48-49, 140
Duke, Trina 34, 60-61
Dunlea, Jamie 32, 49-50
Dunlop, Maggie 32, 49-50
Elder, Catherine 128-129
Elwood, James 33, 52
Erickson, Gudrun K. 38, 90
Evanini, Keelan 40, 98-99
Evensen, Lars Sigfred 127-128
Fairbairn, Judith 40, 96-97
Fan, Jinsong 33, 51
Farhady, Hossein 154
Feaquin, Karen 40, 96-97
Fleckenstein, Johanna 34, 65
Fox, Janna 31, 34, 45, 101-104
Fraser, Catriona 39, 84-85
Gaillard, Stephanie 30-31, 42-43
Galaetz, Evelyn 33, 54-55
Gao, Manman 105, 149-150
Garanpayeh, Ardeshir 38, 41, 119-122
Gesicki, Adam 36, 67
Geva, Esther 37, 110-113
Ginther, April 119-122
Gomez, Pablo Garcia 128-129
Green, Anthony 3, 30-31, 37, 44, 107-109
Green, Rita 30, 43
Gu, Lin 32, 49-50
Gu, Xiangdong 32, 46, 160
Guedez-Fernandez, Romulo 157-158
Guetherman, Tim 32, 53-54
Haghi, Seyyedeh Saeedeh 132
Hama, Mika 159
Hamp-Lyons, Liz 30-31, 35, 37, 41, 44, 107-109
Han, Baochen 41, 116-119
Harsch, Claudia 34, 61
Hartig, Johannes 34, 61
He, Lianzhen 37, 77-78
Hill, Kathryn 37, 107-109
Hines, Susan 128-129
Hoffman, Christina 130
Hsieh, Ching-Ni 32, 49-50
Huang, Jing 35, 104-106
Huhtan, Ari 39, 83-84
<table>
<thead>
<tr>
<th>Presenter</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huynh, Tuan</td>
<td>142</td>
</tr>
<tr>
<td>Ikeda, Naoki</td>
<td>39, 92-93</td>
</tr>
<tr>
<td>Imao, Yasuhiro</td>
<td>134</td>
</tr>
<tr>
<td>In'nami, Yo</td>
<td>133</td>
</tr>
<tr>
<td>Inbar-Lourie, Ofra</td>
<td>40, 95-96</td>
</tr>
<tr>
<td>Inoue, Chihiro</td>
<td>33, 54-55</td>
</tr>
<tr>
<td>Isacs, Talia</td>
<td>38, 82-83</td>
</tr>
<tr>
<td>Isbell, Daniel</td>
<td>139-140</td>
</tr>
<tr>
<td>Iwashita, Noriko</td>
<td>33, 56-57</td>
</tr>
<tr>
<td>Jang, Eunice Eunhee</td>
<td>41, 113-115</td>
</tr>
<tr>
<td>Jarvis, Scott</td>
<td>123</td>
</tr>
<tr>
<td>Jia, Dingding</td>
<td>135-136</td>
</tr>
<tr>
<td>Jia, Guodong</td>
<td>91</td>
</tr>
<tr>
<td>Jin, Tan</td>
<td>39, 87</td>
</tr>
<tr>
<td>Joldersma, Kevin</td>
<td>141, 159</td>
</tr>
<tr>
<td>Kennedy, Lauren</td>
<td>141</td>
</tr>
<tr>
<td>Khalifa, Hanan</td>
<td>32, 53-54</td>
</tr>
<tr>
<td>Kim, Ahyoung Alice</td>
<td>123-125</td>
</tr>
<tr>
<td>Kim, Hyejeong</td>
<td>36, 76-77</td>
</tr>
<tr>
<td>Kim, Hyun Jung</td>
<td>125</td>
</tr>
<tr>
<td>Knoch, Ute</td>
<td>36, 74</td>
</tr>
<tr>
<td>Koizumi, Rie</td>
<td>32, 52-53, 133-134</td>
</tr>
<tr>
<td>Köller, Olaf</td>
<td>34, 65</td>
</tr>
<tr>
<td>Koyama, Yukie</td>
<td>134</td>
</tr>
<tr>
<td>Kramer, Brandon</td>
<td>34, 64-65</td>
</tr>
<tr>
<td>Kremmel, Benjamin</td>
<td>41, 113-115</td>
</tr>
<tr>
<td>Kunnan, Antony John</td>
<td>125-126</td>
</tr>
<tr>
<td>Kvistad, Anne Holten</td>
<td>127-128</td>
</tr>
<tr>
<td>Lammervo, Tiina</td>
<td>135</td>
</tr>
<tr>
<td>Laughlin, Veronika Timpe</td>
<td>130-131</td>
</tr>
<tr>
<td>Laurier, Michel</td>
<td>34, 101-104</td>
</tr>
<tr>
<td>Le Bel Beauchesne, Geneviève</td>
<td>143-144</td>
</tr>
<tr>
<td>Lee, Shinhye</td>
<td>33-35, 62, 71-72</td>
</tr>
<tr>
<td>Lee, Yong-Won</td>
<td>33</td>
</tr>
<tr>
<td>Leucht, Michael</td>
<td>34, 65</td>
</tr>
<tr>
<td>Levi, Tziona</td>
<td>40, 95-96</td>
</tr>
<tr>
<td>Lidster, Ryan</td>
<td>37, 78-79</td>
</tr>
<tr>
<td>Lira Gonzales, Maria-Lourdes</td>
<td>151</td>
</tr>
<tr>
<td>Liu, Hongyun</td>
<td>41, 116-119</td>
</tr>
<tr>
<td>Liu, Li</td>
<td>135-136</td>
</tr>
<tr>
<td>Liu, Sha</td>
<td>39, 91-92</td>
</tr>
<tr>
<td>Llosa, Lorena</td>
<td>33, 36, 47-48</td>
</tr>
<tr>
<td>Lucas, Ruth</td>
<td>147-148</td>
</tr>
<tr>
<td>Luce, Christine</td>
<td>40, 98-99</td>
</tr>
<tr>
<td>Ma, Jia</td>
<td>36, 67-68</td>
</tr>
<tr>
<td>Macqueen, Susy</td>
<td>36, 74</td>
</tr>
<tr>
<td>Maddox, Bryan</td>
<td>34, 101-103</td>
</tr>
<tr>
<td>Malone, Margaret</td>
<td>33, 47-48, 138-139</td>
</tr>
<tr>
<td>Margolis, Douglas</td>
<td>36, 72</td>
</tr>
<tr>
<td>May, Lyn</td>
<td>36, 74</td>
</tr>
<tr>
<td>McLean, Stuart</td>
<td>34, 64-65</td>
</tr>
<tr>
<td>McNamara, Tim</td>
<td>13, 38, 41</td>
</tr>
<tr>
<td>Mei, Yi</td>
<td>155-156</td>
</tr>
<tr>
<td>Mellow, Dean</td>
<td>34, 101-104</td>
</tr>
<tr>
<td>Mesquita, Alexandre</td>
<td>145</td>
</tr>
<tr>
<td>Min, Shangchao</td>
<td>37, 77-78</td>
</tr>
<tr>
<td>Montee, Megan</td>
<td>137-139</td>
</tr>
<tr>
<td>Mu, Yuanyuan</td>
<td>149-150</td>
</tr>
<tr>
<td>Musser, Samantha</td>
<td>138-139</td>
</tr>
<tr>
<td>Nakamura, Keita</td>
<td>133</td>
</tr>
<tr>
<td>Nakatsuhara, Fumiyo</td>
<td>33, 54-55</td>
</tr>
<tr>
<td>Natvig, Siri</td>
<td>127-128</td>
</tr>
<tr>
<td>Nguyen, Hoa</td>
<td>142</td>
</tr>
<tr>
<td>Nguyen, Huu</td>
<td>142</td>
</tr>
<tr>
<td>Nguyen, Lan</td>
<td>142</td>
</tr>
<tr>
<td>Nguyen, Minh</td>
<td>142</td>
</tr>
<tr>
<td>Nieminen, Lea</td>
<td>39, 83-84</td>
</tr>
<tr>
<td>Norviel, Jennie</td>
<td>159</td>
</tr>
<tr>
<td>O'Sullivan, Barry</td>
<td>37, 40-41, 93-94, 110-113</td>
</tr>
<tr>
<td>Ockey, Gary</td>
<td>37, 75</td>
</tr>
<tr>
<td>Pan, Yi-Ching</td>
<td>150</td>
</tr>
<tr>
<td>Papageorgiou, Spiros</td>
<td>37, 75</td>
</tr>
<tr>
<td>Park, Gina</td>
<td>33, 37, 48-49, 81-82</td>
</tr>
<tr>
<td>Petchko, Katerina</td>
<td>33, 52</td>
</tr>
<tr>
<td>Phakiti, Aek</td>
<td>34, 57</td>
</tr>
<tr>
<td>Pill, John</td>
<td>36, 74</td>
</tr>
<tr>
<td>Plakans, Lia</td>
<td>145-146</td>
</tr>
<tr>
<td>Plough, India</td>
<td>146-147</td>
</tr>
<tr>
<td>Pooler, Emilie</td>
<td>128-129</td>
</tr>
<tr>
<td>Potts, Joe</td>
<td>40, 100</td>
</tr>
<tr>
<td>Powers, Nancy</td>
<td>147-148</td>
</tr>
<tr>
<td>Pulupa, Catherine</td>
<td>159</td>
</tr>
<tr>
<td>Purpura, Jim</td>
<td>35, 37, 107-109</td>
</tr>
<tr>
<td>Qi, Luxia</td>
<td>41, 116-119</td>
</tr>
<tr>
<td>Qian, David</td>
<td>34, 58</td>
</tr>
<tr>
<td>R. Scaramucci, Matilde V.</td>
<td>155</td>
</tr>
<tr>
<td>Rahimi, Mohammad</td>
<td>37, 80-81</td>
</tr>
<tr>
<td>Raquel, Michelle</td>
<td>137</td>
</tr>
<tr>
<td>Rodríguez González, Roberquis</td>
<td>124-125</td>
</tr>
<tr>
<td>Roever, Carsten</td>
<td>39, 84-85</td>
</tr>
<tr>
<td>Ryu, Jin</td>
<td>39, 88</td>
</tr>
<tr>
<td>Saif, Shahrzad</td>
<td>37, 80-81</td>
</tr>
<tr>
<td>Saito, Hidetoshi</td>
<td>35, 68-69</td>
</tr>
<tr>
<td>Name</td>
<td>Pages</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Salamoura, Angeliki</td>
<td>32,</td>
</tr>
<tr>
<td>Sandvik, Todd</td>
<td>37,</td>
</tr>
<tr>
<td>Sawaki, Yasuyo</td>
<td>32,</td>
</tr>
<tr>
<td>Schmidgall, Jonathan</td>
<td>38</td>
</tr>
<tr>
<td>Schmitt, Norbert</td>
<td>41,</td>
</tr>
<tr>
<td>Schoonen, Rob</td>
<td>41,</td>
</tr>
<tr>
<td>Segalowitz, Norman</td>
<td>41,</td>
</tr>
<tr>
<td>Selonen, Marja</td>
<td>135</td>
</tr>
<tr>
<td>Shen, Mingbo</td>
<td>35,</td>
</tr>
<tr>
<td>Shin, Sun-Young</td>
<td>37,</td>
</tr>
<tr>
<td>Shiotsu, Toshihiko</td>
<td>141</td>
</tr>
<tr>
<td>Skar, Gustaf Bernhard Uno</td>
<td>127</td>
</tr>
<tr>
<td>Slomp, David</td>
<td>34,</td>
</tr>
<tr>
<td>Smemo, Jorun</td>
<td>127</td>
</tr>
<tr>
<td>So, Youngsoon</td>
<td>40,</td>
</tr>
<tr>
<td>Song, Xiaomei</td>
<td>33,</td>
</tr>
<tr>
<td>Stenner, A. Jackson</td>
<td>37,</td>
</tr>
<tr>
<td>Stevens, Lilly</td>
<td>33,</td>
</tr>
<tr>
<td>Stojanovska, Illina</td>
<td>142</td>
</tr>
<tr>
<td>Stone, Jake</td>
<td>33,</td>
</tr>
<tr>
<td>Storch, Neomy</td>
<td>36,</td>
</tr>
<tr>
<td>Sun, Hailiang</td>
<td>105</td>
</tr>
<tr>
<td>Suni, Minna</td>
<td>135</td>
</tr>
<tr>
<td>Suvorov, Ruslan</td>
<td>34,</td>
</tr>
<tr>
<td>Suzuki, Masanori</td>
<td>41,</td>
</tr>
<tr>
<td>Suzuki, Yulchi</td>
<td>142</td>
</tr>
<tr>
<td>Swartz, Carl</td>
<td>151</td>
</tr>
<tr>
<td>Tan, May</td>
<td>38,</td>
</tr>
<tr>
<td>Tao, Jidong</td>
<td>40,</td>
</tr>
<tr>
<td>Tavassoli, Kobra</td>
<td>154</td>
</tr>
<tr>
<td>Thirakunkovit, Suthathip</td>
<td>35,</td>
</tr>
<tr>
<td>Tian, Jie</td>
<td>32,</td>
</tr>
<tr>
<td>Tran, Phuong</td>
<td>142</td>
</tr>
<tr>
<td>Trottier, Sara</td>
<td>143</td>
</tr>
<tr>
<td>Tsagari, Dina</td>
<td>32,</td>
</tr>
<tr>
<td>Tschirner, Erwin</td>
<td>130</td>
</tr>
<tr>
<td>Turner, Carolyn E.</td>
<td>31,</td>
</tr>
<tr>
<td>Ullakonoja, Riikka</td>
<td>39,</td>
</tr>
<tr>
<td>Urmston, Alan</td>
<td>137</td>
</tr>
<tr>
<td>Vafaee, Payman</td>
<td>142</td>
</tr>
<tr>
<td>Vahabi, Khosro</td>
<td>132</td>
</tr>
<tr>
<td>van der Boom, Edith</td>
<td>37,</td>
</tr>
<tr>
<td>Van Gorp, Koen</td>
<td>38,</td>
</tr>
<tr>
<td>Van Maele, Jan</td>
<td>124</td>
</tr>
<tr>
<td>van Naerssen, Margaret</td>
<td>36,</td>
</tr>
<tr>
<td>van Splunder, Frank</td>
<td>124</td>
</tr>
<tr>
<td>Vasquez, Claudia</td>
<td>33,</td>
</tr>
<tr>
<td>Vaughter, Phallis</td>
<td>128</td>
</tr>
<tr>
<td>Vogt, Karin</td>
<td>32,</td>
</tr>
<tr>
<td>Volkov, Alex</td>
<td>36</td>
</tr>
<tr>
<td>Wagner, Elvis</td>
<td>125</td>
</tr>
<tr>
<td>Wagner, Maryam</td>
<td>33,</td>
</tr>
<tr>
<td>Walkkonis, Charis</td>
<td>159</td>
</tr>
<tr>
<td>Walter, Daniel</td>
<td>35,</td>
</tr>
<tr>
<td>Wang, Boran</td>
<td>39</td>
</tr>
<tr>
<td>Wang, Shujiao</td>
<td>152</td>
</tr>
<tr>
<td>Waring, Hansun</td>
<td>37,</td>
</tr>
<tr>
<td>Watanabe, Toshi</td>
<td>159</td>
</tr>
<tr>
<td>Wei, Jing</td>
<td>33,</td>
</tr>
<tr>
<td>Wei, Junli</td>
<td>156</td>
</tr>
<tr>
<td>Wei, Na</td>
<td>41</td>
</tr>
<tr>
<td>Weigle, Sara</td>
<td>41</td>
</tr>
<tr>
<td>Weyant, Katie</td>
<td>35</td>
</tr>
<tr>
<td>Wilmes, Carsten</td>
<td>123</td>
</tr>
<tr>
<td>Winke, Paula</td>
<td>35</td>
</tr>
<tr>
<td>Wolf, Mikyung Kim</td>
<td>40</td>
</tr>
<tr>
<td>Wong, Roxanne</td>
<td>151</td>
</tr>
<tr>
<td>Wu, Jessica Row-Whei</td>
<td>36</td>
</tr>
<tr>
<td>Wu, Rachel Yf-fen</td>
<td>36</td>
</tr>
<tr>
<td>Wu, Yong</td>
<td>35</td>
</tr>
<tr>
<td>Wu, Yongfie</td>
<td>126</td>
</tr>
<tr>
<td>Wu, Zunmin</td>
<td>37</td>
</tr>
<tr>
<td>Xu, Xiaoming</td>
<td>40</td>
</tr>
<tr>
<td>Xu, Jian</td>
<td>160</td>
</tr>
<tr>
<td>Xu, Yueting</td>
<td>36</td>
</tr>
<tr>
<td>Yan, Qiaozhen</td>
<td>32</td>
</tr>
<tr>
<td>Yan, Xun</td>
<td>35</td>
</tr>
<tr>
<td>Yang, Luna</td>
<td>37</td>
</tr>
<tr>
<td>Youn, Soo Jung</td>
<td>33</td>
</tr>
<tr>
<td>Yu, Guoxing</td>
<td>35</td>
</tr>
<tr>
<td>Zapata, Diego</td>
<td>40</td>
</tr>
<tr>
<td>Zhang, Wenxia</td>
<td>35</td>
</tr>
<tr>
<td>Zhang, Xiaoyi</td>
<td>39</td>
</tr>
<tr>
<td>Zhao, Wen</td>
<td>39</td>
</tr>
<tr>
<td>Zou, Shaoyan</td>
<td>39</td>
</tr>
<tr>
<td>Zumbo, Bruno D.</td>
<td>34</td>
</tr>
</tbody>
</table>
Language Testing in Asia (LTA) was founded in 2011 as a high quality peer-reviewed online academic journal. The purpose of LTA is to acknowledge and showcase scholarly findings in Asia in the field of language testing.

Submissions can be made online at http://www.languagetestingasia.com/

We also plan a special issue on Dynamic Language Testing (contact Miguel Fernadez at mferna20@csu.edu)
The British Council
Assessment Research Awards and Grants

If you’re involved or work in research into assessment, then the British Council Assessment Research Awards and Grants (ARAGs) might interest you. These awards recognise achievement and innovation within the field of language assessment and form part of the British Council’s extensive support of research activities across the world. The ARAGs have four categories:

- Assessment Research Awards
- Assessment Research Grants
- Innovation in Assessment Prize
- International Assessment Award

For more information visit www.britishcouncil.org/aptis/research or write to arag@britishcouncil.org

In addition to the global ARAGs, the British Council offers funding for language assessment research focused specifically on the East Asia region. For more information visit www.britishcouncil.cn/en/exams/aptis/researchers