LTEC 2014
THE 36TH LANGUAGE TESTING AND RESEARCH COLLOQUIUM

AMSTERDAM JUNE 2-6, 2014
TOWARDS A UNIVERSAL FRAMEWORK

PEARSON English
VU UNIVERSITY AMSTERDAM Faculty of Arts
Acknowledgments

The LTRC 2014 Organizing Committee

Chair
John de Jong

VU University Amsterdam
Gea Dreschler
Bert Weltens
Rebecca Present-Thomas
Helena van Veen
Mike Hannay

Pearson, London
Veronica Benigno
Alex Perry-Dean

We are grateful for generous support from our Sponsors:
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Getting Around Amsterdam</td>
<td>4</td>
</tr>
<tr>
<td>President's Message to LTRC 2014</td>
<td>6</td>
</tr>
<tr>
<td>Message from LTRC Chair</td>
<td>7</td>
</tr>
<tr>
<td>Awards 2014</td>
<td>8</td>
</tr>
<tr>
<td>Conference Program Schedule</td>
<td>9</td>
</tr>
<tr>
<td>Pre-Conference Workshops</td>
<td>10</td>
</tr>
<tr>
<td>Abstracts Wednesday June 4</td>
<td>15</td>
</tr>
<tr>
<td>Abstracts Thursday June 5</td>
<td>40</td>
</tr>
<tr>
<td>Abstracts Friday June 6</td>
<td>85</td>
</tr>
<tr>
<td>Conference Program Details Day by Day</td>
<td>118</td>
</tr>
<tr>
<td>LTRC 2014 Reviewers and Volunteers</td>
<td>121</td>
</tr>
<tr>
<td>Index</td>
<td>122</td>
</tr>
</tbody>
</table>
Getting Around Amsterdam

A comprehensive system of trains, buses, trams, and the metro ensure convenient public transportation in and around Amsterdam.

Travelling from Schiphol to Amsterdam

From the arrivals hall of Schiphol Airport, you can take an NS (the Dutch railway operator) train to the city center (station: Amsterdam Centraal) or VU University (station: Amsterdam Zuid). If you have a bank card with an appropriate chip and pin code, you can purchase your train tickets from the yellow NS machines near the airport exit. It is also possible to purchase tickets and ask questions at the NS desk, which is identified on the map below by the blue NS logo in front of Arrivals 2.

Trains from Schiphol leave several times per hour. Single, 2nd-class tickets cost approximately €2.60 to Amsterdam Zuid (~8 minutes travel time) or €4.00 to Amsterdam Centraal (~15 minutes).

Around Amsterdam with buses, trams, and metro

Once you're in Amsterdam, you'll want to switch to local transportation. Single-use tickets for trams and buses can be purchased on board for about €2.80 and are valid for one hour. However, we highly recommend purchasing a multi-day ticket for unlimited access to GVB buses, trams, and metro during your stay. You can purchase multi-day tickets in advance at the GVB ticket offices at Amsterdam Centraal and Amsterdam Zuid train stations. (Note that Connexxion buses and NS trains are not included in the multi-day tickets.)

Getting to/from VU University

The university is a short walk (~10 minutes) from train station Amsterdam Zuid. Additionally, the following local (GVB) lines, as shown on the campus map, will bring you directly to the VU Hoofdgebouw (Main Building):

- Tram lines 5, 16, and 24
- Metro line 51
- Buses 176, 241, 242, and 310
Getting Around Amsterdam

At VU University

All workshops, presentations and exhibitions will take place in the Hoofdgebouw (Main Building), which is located at de Boelelaan 1105.

The opening reception will take place at the Basket campus pub (next to the parking lots in the middle of the main campus block).

Conference dinner

Restaurant Café In de Waag is located on the Nieuwmarkt in the city center. From VU University, metro line 51 (direction: Amsterdam Centraal) stops directly at Nieuwmarkt, just across the square from the restaurant. Alternatively, tram lines 5, 16 and 24 will all take you to Dam Square, an 8-minute walk from the restaurant.
President's Message to LTRC 2014

Welcome to LTRC Amsterdam! - the 36th Language Testing Research Colloquium since our humble beginnings in a Boston hotel room in 1979. The conference organizers, chaired by John de Jong, and including Gea Dreschler, Bert Weltens, Rebecca Present-Thomas, Helena van Veen, Mike Hannay, Veronica Benigno, and Alex Perry-Dean, have put together a rich and varied program on the important theme of *Towards a Universal Framework*. The last time LTRC met in this lovely country was in August, 1993, when we experienced a two-nation colloquium, Part One in Cambridge and Part Two in Arnhem - the link between then and now being John de Jong. We're very fortunate this time around to enjoy the hospitality of VU University, and the facilities are not only first-rate but inviting and congenial, encouraging that most important aspect of any LTRC - socializing with old friends and discovering new ones.

Here in Amsterdam, we are being treated to a wealth of information and inspiration regarding language assessment, including four pre-conference workshops, 75 paper presentations, 4 symposia, 30 work in progress presentations, 19 posters, 7 plenary sessions, and 3 keynotes, featuring, the Samuel J. Messick Memorial Lecture, the Distinguished Achievement Award Lecture, and the John Trim Award Lecture. I’m sure we will all come away with intellectual riches that will guide and inspire language testing research for a long time.

We are also being indulged by wonderful social events, including the opening reception at *The Basket* and the conference banquet at *In de Waag*, surely the oldest banquet venue in LTRC history. Will there be any surprises during the banquet? It’s an LTRC tradition that remarkable, even magical things happen at the banquet. We’ll see...

ILTA, our professional organization, of which LTRC is the annual conference, continues to work toward its purpose of promoting the improvement of language testing throughout the world. We work with other professional organizations such as AAAL, TESOL, and the many regional and national language testing organizations worldwide to disseminate information about language assessment. We have an informal working group of experts in the assessment of aviation English working with the International Civil Aviation Organization.

Each year we publish an updated ILTA Language ILTA Bibliography and a Doctoral Dissertation Bibliography – we thank Tineke Brunfaut, of Lancaster University, for taking this project over this year from Annie Brown who has been conscientiously compiling the bibliographies for many years. Thank you Annie! We recognize excellence in language assessment research through the Cambridge/ILTA Distinguished Achievement Award and the ILTA Best Article Award. We promote education in language assessment with our annual Workshops and Meetings Awards. We encourage the professional development of MA and PhD students in language assessment by providing travel grants and a student presentation award. Jon Trace, University of Hawai‘i, and Maryam Wagner, University of Toronto, were the recipients of the Student Travel Award this year. We look forward to hearing the many student presentations at this LTRC and to seeing which will win the Lado Best Student Paper Award.

So, thank you very much to the organizing committee of this year’s LTRC and to all of you who are participating in the Colloquium this year. There is no doubt that it will be a memorable and merry meeting!

All best wishes,

Dan Douglas
ILTA President
On behalf of the Organizing Committee I am very happy to welcome you all to Amsterdam. It has been an interesting year preparing this meeting of the language testing minds. From the moment of putting out the call for papers correspondence has been proliferous. The call resulted in some negative and some positive reactions, but it did attract a large number of proposals: 224 in total at an interesting rate of submission as shown in the first graph.

The 224 proposals were each reviewed by three raters generating 672 ratings divided over 73 volunteer raters (see page121). The scoring rubrics were the same for the four types of proposals (paper, poster, WIP and symposium). Ratings resulted in 69% of the proposals being accepted (score of 4 or higher on scale of 1 to 6). Overall rater agreement did not meet high-stakes standards with 30% of exact agreement and a further 38% with a maximum difference of 1 score point between raters. The highest acceptance rate (87%) was for posters, followed by symposia (80%).

Acceptance rate for WIPs was 65% whereas 56% of the papers proposals were accepted. A total 18 proposals were withdrawn for various reasons.

The program opens each day with an award lecture, followed by one or two plenary sessions after which we continue in four different streams and come together again at the end of the day in a plenary session. For the streams we endeavoured to group papers by topic, but this was only partly successful as some proposals indicated up to five different themes of the 10 themes we suggested in the call for proposals and up to 9 keywords. Judging from the popularity of themes (shown in the second graph) it would seem that we are in general interested in serving the community of language learners and language teachers and hope to produce transparent information on learners’ progress towards proficiency.

There are many contributions that offer interesting perspectives in relation to the conference theme of seeking to define a universal framework and I am sure we will have interesting and heated debates in and around the campus. As of the time of this programme going to print we have 262 registered participants from 35 countries and from close to 200 different institutions. My wish for all of us is that we will be able to make some headway towards understanding what it means for learners to achieve progress in mastering a language.
Awards 2014

Award Committees:

ILTA Student Travel Awards
Chair: Tony Green (University of Bedfordshire, UK)
Elvis Wagner (Temple University, USA)
Jessica Wu (The Language Training & Testing Center (LTTC), Taiwan)

ILTA Best Article Award
Chair: Ute Knoch (University of Melbourne, Australia)

ILTA Workshops and Meetings Awards
Chair: Anthony Green (University of Bedfordshire, UK)
Margaret (Meg) Malone (Center for Applied Linguistics, USA)
Elvis Wagner (Temple University, USA)

Award Winners:

ILTA Student Travel Awards
Jon Trace, University of Hawai‘i, Paper: Building a better rubric: Towards a more robust description of academic writing proficiency (with Gerriet Janssen and Valerie Meier).
Maryam Wagner, Ontario Institute for Studies in Education/University of Toronto. Use of a diagnostic rubric for assessing writing: Students’ perceptions of cognitive diagnostic feedback

ILTA Best Article Award
Pending

ILTA Workshops and Meetings Awards
Classroom Language Assessment Workshops in Namibia and Tanzania: Lyle F. Bachman, Barbara Damböck, Liz Hamp-Lyons and Dan Douglas.
A workshop meeting to promote testing of Polish, Russian, Turkish and Arabic as heritage languages in Germany: Astrid Reich, Anna Soltyska and Anna Timukova

Jacqueline Ross TOEFL Dissertation Award
Jonathan Smidgall (University of California, Los Angeles) Modeling speaker proficiency, comprehensibility, and perceived competence in a language use domain
Supervisor: Dr. Lyle Bachman

Caroline Clapham IELTS Masters Award, 2013
Benjamin Kremmel (Lancaster University, UK). Explaining Variance in Reading Test Performance through Linguistic Knowledge: The Relative Significance of Vocabulary, Syntactic and Phraseological Knowledge in Predicting Second Language Reading Comprehension
Supervisor: Tineke Brunfaut.

2013 TIRF Doctoral Dissertation Awards in Language Assessment
Kimberley Woo (New York University)
Nick Zhiwei Bi (University of Sydney)
### Conference Program Schedule

**Monday, June 2**
- 8:15-17:00: Registration
- 9:00-17:00: Workshops

**Tuesday, June 3**
- 8:15-17:00: Registration
- 9:00-17:00: Workshops
- 19:00-21:00: Welcome Reception in The Basket

**Wednesday, June 4**
- 8:15-17:00: Registration
- 9:00-09:45: Opening of Conference by Prof. Michel ter Hark, Dean of the Faculty of Arts
- 9:05-09:45: Keynote Session 1
- 9:45-10:15: Plenary Session 1
- 10:20-10:50: Parallel Session 1
- 10:50-11:20: Break
- 11:25-11:55: Parallel Session 2
- 12:00-12:30: Parallel Session 3
- 12:35-13:15: Conference Photo
- 12:35-13:55: Lunch Break
- (Language Testing EB Meeting)
  - 14:00-15:40: Symposia Session 1
  - 14:00-14:30: Parallel Session 4
  - 14:35-15:05: Parallel Session 5
  - 15:10-15:40: Parallel Session 6
- 15:45-16:05: Break
- 16:10-16:40: Parallel Session 7
- 16:45-17:15: Parallel Session 8
- 17:20-17:50: Plenary Session 2
- 17:50-18:00: Poster Pitches 1-10

**Thursday, June 5**
- 9:00-09:45: Keynote Session 2
- 9:45-10:15: Plenary Session 3
- 10:20-11:05: Works-in-Progress
- 11:05-11:25: Break
- 11:30-12:00: Parallel Session 9
- 12:05-12:35: Parallel Session 10
- 12:35-13:55: Lunch Break
  (ILTA Annual Business Meeting)
- 14:00-14:30: Parallel Session 11
- 14:35-15:05: Parallel Session 12
- 15:10-15:40: Parallel Session 13
- 15:45-16:10: Break
- 16:10-16:40: Parallel Session 14
- 16:45-17:15: Plenary Session 4
- 17:20-17:50: Plenary Session 5
- 17:50-18:00: Poster Pitches 11-19
- 19:00: Conference Dinner in In de Waag

**Friday, June 6**
- 09:15-10:00: Keynote Session 3
- 10:00-10:30: Plenary Session 6
- 10:35-11:05: Parallel Session 15
- 11:05-11:30: Break
- 11:30-12:35: Poster Session
- 12:35-14:00: Lunch Break
  (Language Assessment Quarterly EB Meeting)
- 14:00-15:40: Symposia Session 2
- 14:00-14:30: Parallel Session 16
- 14:35-15:05: Parallel Session 17
- 15:10-15:40: Parallel Session 18
- 15:45-16:05: Break
- 16:10-16:40: Parallel Session 19
- 16:45-17:15: Parallel Session 20
- 17:20-17:50: Plenary Session 7
Pre-Conference Workshops June 2 & 3

Workshop 1: Using an Assessment Use Argument to Develop Language Tests and Justify Their Use

Monday, June 2 - Tuesday, June 3

Facilitators: Lyle Bachman, Daniel Dixon, & Adrian Palmer

Argument-based approaches to validation have been discussed in the educational measurement literature for over twenty years and in the language assessment literature for nearly a decade. One of the most highly developed exponents of this general approach is that of Bachman and Palmer (2010). They provide a detailed description of an “assessment use argument” (AUA) that can guide the development of language assessments, along with a set of procedures that test developers can follow in the development of language assessments and the justification of their use. Their concept of an AUA is beginning to inform language assessment development efforts around the world, and is also finding its way into the language assessment.

In this workshop, which will be based on Bachman & Palmer, 2010, Language Assessment in Practice, we will provide a detailed discussion of the process of assessment justification and the role of an AUA in this. We will also introduce an on-line tool for implementing Bachman and Palmer’s approach to language test development, and provide participants with hands-on practice in its use.

Lyle F. Bachman is Professor Emeritus of Applied Linguistics at the University of California, Los Angeles. He is a Past President of the American Association for Applied Linguistics and of the International Language Testing Association. He has received numerous awards for his research and service to the profession, including the TESOL/Newbury House Award for Outstanding Research, the Modern Language Association of America’s Kenneth Mildenberger Award for outstanding research publication (twice, with A. S. Palmer), the Sage/International Language Testing Association award for the best book published in language testing (with A. S. Palmer), the Lifetime Achievement Award from the International Language Testing Association, and the Distinguished Scholarship and Service Award from the American Association for Applied Linguistics. He has published numerous articles and books in language testing and other areas of Applied Linguistics. His current research interests include validation theory, classroom assessment, and epistemological issues in Applied Linguistics research.

Adrian Palmer teaches a variety of courses on language assessment, teacher training and professional development, and language teaching methodology, and he regularly offers workshops and presentations on these topics. Two of his recent books include Seeking the Heart of Teaching (University of Michigan Press), 2007, and Language Assessment in Practice (Oxford University Press), 2010. This latter book recently received the 2012 SAGE/ILTA triennial award for the best book in the area of language testing/assessment.

Daniel Dixon completed his Masters in Applied Linguistics at the University of Utah in December of 2013. In addition to being a graduate student and ESL instructor at the U of U, he worked with Dr. Adrian Palmer to develop the online Assessment Use Argument website. He was awarded a Fulbright Student Award and is currently an English Teaching Assistant at the Universidade Federal de Mato Grosso in Cuiabá, Brazil for the 2014 academic year. He also plans to use his ten years of experience working as a television news photojournalist to produce a documentary during his time in Brazil.
Workshop 2: Meta-analysis: Conceptual understanding and application

Monday, June 2 - Tuesday, June 3
Facilitators: Yo In’nami & Eric Surface

As with any academic domains characterized by the accumulation of knowledge over a large number of studies, research on language testing has seen increased development within recent decades, epitomized by high quality studies published in flagship journals such as *Language Testing* and *Language Assessment Quarterly*. The availability of a rich volume of studies heightens the need for more research syntheses. One way to address this issue is to use meta-analysis, which is a research method that summarizes a set of empirical data across studies. Meta-analysis, as a systematic review of the literature, is an important component of evidence-based practice and decision-making models (e.g., Briner, 1998; Briner & Rousseau, 2011; Rousseau & McCarthy, 2007) and of integrating research and informing future research directions. Although the concept of meta-analysis is a century old, Glass (1976) and Norris and Ortega (2000) initiated a widespread modern interest in this method in education and applied linguistics research.

Comparing meta-analysis with traditional, narrative reviews, Plonsky and Oswald (2013) argues that meta-analysis excels in terms of its systematic approach to data collection and analysis, its use of effect sizes as the unit of analysis, and its capability to conduct moderator variable analysis. Plonsky (2012) discusses the inherent relationship among meta-analysis, replication, and generalizability, and argues that the combination of meta-analysis and replication determines the generalizability of study findings across various settings. Since generalizability is of great concern to language testers, meta-analyses of similar or replicated studies in language testing hold great promise in terms of establishing a solid foundation for our current practice and propelling us in new directions.

**Yo In’nami** is an Associate Professor of English at Shibaura Institute of Technology, Japan. He is also a PhD candidate’s adviser and an external PhD examiner at the Temple University, Japan Campus. As an adjunct lecturer at the University of Tsukuba, Japan, he taught postgraduate courses on structural equation modeling and meta-analysis. He is a member of the AILA Research Network on Research Synthesis and Meta-Analysis (conveners John Norris, Lourdes Ortega, and Luke Plonsky). He is an organizing committee chair for the 18th Annual Meeting of Japan Language Testing Association held on September 2014 at Kyoto. His research interests include meta-analytic inquiry into the variability of effects and the longitudinal measurement of change in language proficiency. He has published widely in domestic and international journals, including Language Testing and Language Assessment Quarterly. He is currently co-editing a special issue on research synthesis and meta-analysis for the Taiwanese journal *English Teaching & Learning*. For further information, please visit [https://sites.google.com/site/yoinnami/](https://sites.google.com/site/yoinnami/).

**Eric A. Surface** is an industrial/organizational psychologist who co-founded SWA Consulting Inc. (SWA), an applied research and consulting firm in Raleigh, NC. SWA’s mission is to make a difference in learning and performance for individuals and organizations. Eric is currently the president of SWA and principal scientist and has been the principal investigator of numerous projects related to training, testing, performance and organizational effectiveness. Eric strives to facilitate the use of evidence (such as meta-analytic results) and data by clients to better analyze, learn, perform and succeed within their context and mission. His recent work involves the alignment of individual capability with mission or organizational requirements and tools to optimize learning outcomes and training effectiveness. Much of his work has specific focus on Language, Regional Expertise and Culture (LREC) in military and government organizations. Eric has presented numerous conference papers in the fields of I/O psychology, management and foreign language learning/testing, including meta-analytic studies at AAAL and LTRC meetings. His research appears in peer-reviewed journals, such as *Organizational Research Methods*, *Personnel Psychology*, *Journal of Management*, *Journal of Applied Psychology and Foreign Language Annuals*. His research interests include learning, training evaluation and effectiveness, needs assessment, work performance, context, validation and testing, and work-related foreign language acquisition and measurement. More information on publication and conference papers is available at [www.swa-consulting.com](http://www.swa-consulting.com).
Workshop 3: From ‘Languaging’ to ‘Translanguaging’: Assessing foreign language learners’ mediation competence
Monday, June 2

Facilitator: Maria Stathopoulou

Despite the fact that the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) has considered mediation as a significant aspect of language users’ proficiency (especially in today’s multilingual societies), it provides no mediation-specific can-do statements which would in turn facilitate the reliable assessment of the particular competence.

The workshop is mainly addressed to foreign language teachers who teach foreign languages either in the public or the private sector and want to enhance their knowledge of multilingual testing in general and mediation in particular. Additionally, the issues discussed in this workshop would be particularly useful for researchers in the field of applied linguistics and foreign language testing, language testing experts and teacher trainers. The workshop may also interest curriculum developers who want to link their curricula to multilingual testing and materials developers who wish to include mediation tasks in their materials. I do not expect participants to have any specific background or prior knowledge. Preparatory reading is not necessary. Finally, I do not expect participants to bring along any extra materials apart from the handouts distributed during the workshop.

Maria Stathopoulou is a PhD holder in Applied Linguistics, University of Athens. My scholarly interests range widely from multilingualism/interculturalism and the teaching of foreign languages to the assessment of written performance, foreign language curriculum design and language test development. My doctoral thesis focused on translanguaging and cross language mediation as basic elements of my wider concern with multilingual education and multiliteracies and my narrower concern with multilingual testing. Given that the CEFR does not provide any illustrative descriptors relevant to mediation, my study constitutes a systematic attempt to complement it by developing objective criteria so as to describe levelled mediation performance. The fact that my work has been published abroad and in Greece and presented in international conferences around Europe (i.e., the UK, France, Italy, Finland, Cyprus, Belgium etc) supports my strong background in foreign language testing and educational research. Also, in 2010, I offered a successful workshop (with Prof. Dendrinos) on mediation in the 1st International Conference on Language Testing and Assessment, University of Cyprus, Nicosia.

Additionally, I have been offering seminars to both undergraduate and post-graduate students at the University of Athens since 2010. Within the framework of the "Pre-service Teacher Training Programme" at the University of Athens, I am now developing an e-course about mediation for students of the Faculty of English Language and Literature who are unable to attend regular sessions. Also, I am currently a member of the RCeL team designing the online training material for assessors of the National Foreign Language Exams leading to state certification (KPG exams) and a member of the team investigating the quality of script evaluation, the ways that script raters use evaluation criteria and the characteristics of scripts which systematically cause problems in interrater reliability. The fact that, since 2008, I have been working as a research associate and task-developer for the KPG exams -which is actually the only exam battery in Europe which assesses mediation performance, seems to make me suitable to run the proposed workshop.

1 North (2007: 657) characteristically says: “What users complain about in relation to the CEFR descriptors are [...] The absence of descriptors for socio-cultural aspects and reading literature (both tried) and for mediation (not attempted). (North, 2007: 657).
Pre-Conference Workshops June 2 & 3

Workshop 4: Constructing test tasks that work: a hands-on introduction to item-writing and quality reviewing techniques

Tuesday, June 3

Facilitators: Jamie Dunlea, Gwen Caudwell, & Kevin Rutherford

Underlying all good test tasks is an explicit design template referred to as a specification, or spec. Test specifications define the test design at the macro level, and task specifications define the test at a micro level. This provides clear, detailed design templates for constructing test tasks and items to a consistent format. For the purposes of this workshop, an item is a discrete (or stand-alone) question where there is only one correct answer whereas a task is either a series of questions based on a single language input (e.g. a reading text which comes with a set of related comprehension questions) or an activity which is designed to elicit performance on a specific skill (e.g. speaking or writing). The task spec should not only make clear what the purpose of the task is, but also clearly identify criterial features of any input texts and item or items attached to it. In the past, adjusting tasks and items for difficulty was often a matter of expert judgment, something that is hard to quantify and even harder to replicate consistently. Item writers and test developers now have a range of software tools at their disposal, many of which are freely available online. These tools allow item writers to easily identify when texts differ from the criterial parameters in the specifications. This workshop will look at the process of identifying these important features, the use of automated analysis tools to examine texts in relation to the criterial features specified in a task spec, and finish by examining important considerations in the development of items and response alternatives for multiple-choice test items.

Working with test tasks is an increasingly important role for many teachers, either creating them for use within a class or institution, or using them to help students prepare for high-stakes assessments. Researchers already working in language assessment and test developers at all levels of test development also need to be able to construct items which perform well and measure what is intended. An understanding of the theoretical principals of language testing validation and statistical methods for test analysis may not be enough to guide a test item through the important stages between the definition of the construct and the realization of that definition in the form of a concrete test item. This workshop will help participants learn some of the important techniques professional item writers (should) employ to ensure test tasks work in the way they are intended to.

Jamie Dunlea is a researcher for the British Council specializing in English language assessment. Jamie is based in the UK and joined the British council in 2013 after a 20-year career in EFL education in Japan. From 2007-2013, Jamie was the Chief Researcher for the Eiken Foundation in Japan, leading validation research into a large scale, high stakes testing program. He has extensive experience in the application of validation theory to test development, and has presented widely in Japan and internationally. His research interests include applying international language testing validation frameworks to locally developed tests, investigating the relationship of tests to the Common European Framework of Reference, and applying automated textual analysis tools to investigate the lexical and linguistic features of reading and listening tests and the production of L2 learners.

Gwendydd Caudwell is Aptis Product Development Manager at the British Council, and is based in the United Arab Emirates. She started off as a teacher for the British Council in Bangkok in 2007 and continued with them doing her DELTA and becoming a Teacher Trainer and Manager and consequently managing a Teaching Centre in India. Her interest in assessment was sparked by attending the Roehampton University Certificate in the Theory and Practice of Language Testing in 2009. Following this introduction, Gwen undertook professional development training to become an item-writer, item-writer trainer and rater trainer. Gwen has extensive hands-on experience with item writing and the development of test specifications for an international testing system. She completed her MA in Language Testing at Lancaster University in 2012, and is particularly interested in researching the decision-making processes of raters and rater judgements.

Kevin Rutherford is Aptis Test Production Manager at the British Council, and is based in Poland. He is responsible for the commissioning and quality reviewing of items for item banking, and also oversees the
construction of operational test forms for an international language testing system. Kevin has extensive experience as an EFL teacher, and completed the Roehampton Certificate in the Theory of Language Testing and Assessment before undertaking professional development training in item writing. Kevin has experience working on the development of test specifications for an international language testing system from the design and implementation stage, and is an experienced item-writer trainer. He has just completed his dissertation for his MA in Language Testing at Lancaster University and is particularly interested in assessing writing.
The British Council John Trim Lecture Award

Keynote Session 1: Wednesday, June 4, 9:00

Presenter: Douglas Biber

Affiliation: Northern Arizona University

Title: Predictable patterns of linguistic variation across proficiency levels and textual task types

It is probably uncontroversial that language production in the spoken versus written mode is a crucially important component of task variation. Recognizing this fact, major standardized English language exams (e.g., TOEFL-iBT, IELTS, PTE-Academic) have been designed to assess performance in both speech and writing, including multiple task types in each mode (e.g., ‘independent’ vs ‘integrated’ task types in the case of TOEFL-iBT and PTE-Academic). Despite this awareness, there has been relatively little research carried out to date to explore the linguistic differences among the task types included on standardized exams (see Brown et al. 2005; Cumming et al. 2005; Banerjee et al. 2007). In addition to being few in number, previous linguistic descriptions have been limited in three major respects. First of all, these studies have been restricted mostly to vocabulary measures and a few grammatical measures that are difficult to interpret linguistically (e.g., the number of dependent clauses per T-unit). Second, no previous study (to my knowledge) has directly compared the linguistic characteristics of spoken versus written texts that are produced for standardized language exams. And finally, researchers have simply set out to establish the existence of linguistic differences (as evidence for the validity of task types), with no specific predictions concerning the expected patterns of linguistic variation among task types. The present study offers an approach that attempts to address these limitations. It first begins by surveying previous linguistic research on register variation, arguing that this framework enables specific predictions concerning the expected patterns of grammatical variation among spoken and written task types (and across proficiency levels) on standardized exams. Those predictions are then evaluated through linguistic analysis of a large corpus of TOEFL iBT responses, exploring the linguistic patterns of variation across spoken and written modes, independent and integrated task types, and proficiency levels. Three linguistic studies are presented, each building on the previous one. The first is based on consideration of 23 individual grammatical complexity features emerging from previous register research; the second based on the co-occurrence patterns among linguistic features, using Multi-Dimensional (MD) Analysis; and the third uses Cluster Analysis to explore the possibility of linguistic developmental differences that are not captured by proficiency ratings. The results of these investigations show that the linguistic differences among task types are predictable based on their situational/register characteristics, and that developmental progressions are similarly predictable, as students make the transition from the grammatical styles associated with speech to the grammatical styles associated with advanced academic writing. The Multi-Dimensional study further shows that these linguistic predictions are much more reliable and informative when they are based on sets of co-occurring lexico-grammatical features. And finally, interpreting the results of the cluster analysis indicates the existence of systematic development in the use of complexity features that is not captured by holistic exam ratings.

Douglas Biber is Regents’ Professor of English (Applied Linguistics) at Northern Arizona University. His research efforts have focused on corpus linguistics, English grammar, register variation (in English and cross-linguistic; synchronic and diachronic), and how corpus-based research findings can be applied in English Language Teaching. He has published over 190 research articles and 21 books and monographs. For example, he has written research monographs on the linguistic characteristics of American university registers (Benjamins 2006) and methodological approaches to corpus-based discourse analysis (Benjamins 2007). He is widely known for his work on the corpus-based Longman Grammar of Spoken and Written English (1999) and for the development of ‘Multi-Dimensional Analysis’ (a research approach for the study of register variation), described in earlier books published by Cambridge University Press (1988, 1995, 1998). More recent textbooks of immediate relevance to ELT professionals include an introduction to Register, Genre, and Style (Cambridge, 2009), the college-level Longman Student Grammar of Spoken and Written English (2002), and the corpus-based ESL grammar textbook Real Grammar (Longman 2009).
Abstracts June 4

**Plenary Session 1: Wednesday, June 4, 9:45**

**Presenter:** John Read and Norbert Schmitt

**Affiliation:** University of Auckland and University of Nottingham

**Title:** The Potential of Vocabulary Measures in Informing Proficiency Frameworks

It is well established that vocabulary knowledge is a strong predictor of reading comprehension ability, and indeed vocabulary tests are commonly used as broad indicators of language proficiency for placement purposes and in applied linguistic research studies. However, vocabulary measures do not figure prominently in the major proficiency frameworks, except in the form of descriptors which make general statements about the range and appropriate use of lexical expressions in candidate production. One of the criticisms of the CEFR has been the lack of specification of the linguistic features of learner ability at various levels of the framework (see, e.g., Hulstijn, 2007, 2013). There are essentially two approaches to addressing this situation. One, using word frequency as a metric, is to measure the vocabulary size of learners at different levels of proficiency, as for instance Milton (2009) has done. This is also the role of the vocabulary size measure incorporated in DIALANG (Alderson, 2005). The other approach is to describe the lexical features of candidate production in proficiency tests. Read and Nation’s (2006) study of the IELTS Speaking Module is a case in point, and the approach has been implemented much more comprehensively in the English Profile Project (www.englishprofile.org). This paper will present a critical review of the empirical evidence from both approaches for the contribution that lexical measures can make to defining proficiency levels in operational assessments. We will argue that vocabulary testing can add a valuable diagnostic dimension to proficiency assessment. The paper will go on to outline a research agenda to investigate the extent to which vocabulary tests can provide a valid basis for inferences about the corresponding functional language skills. For example, can we draw any conclusions about learners’ ability to rapidly recognize the meanings of words in a reading passage from their performance on a conventional vocabulary test which presents the target words in a relatively discrete and context-independent manner? We need to ask similar questions about so-called “productive” vocabulary tests which implicitly claim to assess whether learners are able to use the target items in their own written production. When we move to oral-aural skills, the research gap is much greater. Very few published measures of L2 listening vocabulary even exist, and little is known about how knowledge of the spoken form of words might contribute to fluent listening comprehension. The same applies to oral vocabulary knowledge and learners’ spoken production. Thus, there is much still to be learned about how appropriate vocabulary measures can inform the assessment of learner proficiency across the skill areas.

**Parallel Session 1.1: Wednesday, June 4, 10:20**

**Presenter:** Tim McNamara

**Affiliation:** University of Melbourne

**Title:** Using a general proficiency scale for a specific occupational context: How valid are the ICAO scale descriptors for radiotelephony communication?

Concerns within the aviation industry about levels of English proficiency among non-native English speaking aviation personnel led to the establishment of common English standards or Language Proficiency Requirements (LPRs) by the International Civil Aviation Organization (ICAO), a special agency of the UN. The general proficiency rating scales and descriptors associated with these LPRs have since been used as the basis for English tests developed to meet these requirements. These scales, however, have been criticised for their failure to fully reflect the communicative demands of radiotelephony communication (Kim, 2013; Kim and
Elder, 2009) and in particular for their insensitivity to the strategies widely recognized as being necessary for successful interaction in English as a lingua franca contexts involving users from diverse L1 backgrounds (e.g., Canagarajah, 2007; Firth, 1996; House, 2003; McNamara, 2012; Meierkord, 2004). This paper reports on a study which set out to investigate the validity of such criticisms by investigating the construct of radiotelephony communication from the perspective of stakeholders from a particular aviation community, namely Korea. The study employed multiple methods including an analysis of focus group feedback on six recorded samples of emergency and urgency radiotelephony discourse collected from the Incheon air traffic control towers, supplemented with interview and questionnaire responses from a sample of 300 Korean pilots and 100 air traffic controllers. A thematically based coding scheme was developed with a view to identifying the features of radiotelephony discourse deemed by occupational experts to be critical for effective communication in the aviation context. The features which emerged from this analysis were then compared with the six assessment criteria (Pronunciation, Structure, Vocabulary, Fluency, Comprehension, and Interactions) and associated linguistic proficiency descriptors which comprise the ICAO rating scale. Results revealed that the strategic features most valued by aviation personnel include dealing with various L1 influenced pronunciations and accents, using simpler structure, avoiding choice of words whose meanings are unlikely to be shared by the majority of the community, delivering messages succinctly, and being sensitive to an interlocutor’s level of English proficiency and willing to make an effort to achieve successful communication. However, these valued strategic features were not well reflected in (and in some cases even at odds with) the content of the ICAO scale. Accordingly, it is concluded that the traditional view of language proficiency underlying the LPRs’ scales is too narrow and underrepresents the radiotelephony communication construct. The paper goes on to consider the broader implications of using general proficiency scales for specific occupational contexts and argues for a grounded empirical approach to scale development.

Parallel Session 1.2: Wednesday, June 4, 10:20

Presenter: Cathie Elder and Ute Knoch

Affiliation: University of Melbourne

Title: Tensions between general language frameworks and language-specific achievements: the case of Asian languages in Australia

The paper briefly reviews ongoing debates surrounding the development and use of frameworks characterizing student achievements in language learning. These frameworks are part of the broader context of outcomes-based assessment for accountability purposes operating in many educational settings around the world. In Australia, discussion has thus far focused mainly on English, as the language of mainstream education, and has highlighted the dangers of generalization and standardization inherent in such frameworks (McKay, 2000) and their potentially negative impact on particular learner groups, such as students of English as a Second Language (both immigrant and indigenous). The dangers of standardisation when applied to languages other than English are arguably even greater in the Australian context, because of a) the different linguistic challenges and complex local ecologies of the particular languages taught in schools and b) the diversity of the learner population, with new learners from English and other L1 backgrounds often in the same class as heritage learners who may speak non-standard varieties and vary in levels of exposure to or prior instruction in the taught language (Elder, 2000; Elder, Kim and Knoch, 2011). The generic national and state frameworks devised for teachers of languages neutralize this contextual diversity to the point of rendering them meaningless to their users (Scarino, 2011). The paper proceeds to outline a recent project designed to address some of the problems with current outcomes-based frameworks, with specific reference to four Asian languages (Chinese, Japanese, Indonesian and Korean) commonly taught in Australian schools. The project was designed with the aim of developing context-sensitive, evidence-based descriptions of achievement which would capture
important differences both between the languages and within the language learner population for each
language. Tests of reading, writing and oral interaction at 3 levels of schooling (end of primary, mid secondary
and senior secondary) were devised to the same broad specifications in each of the four languages,
administered to school age learners across four states of Australia and rated by trained language assessors
against a common set of agreed criteria. Scores derived from this process were cross-referenced to language
background information elicited via questionnaire from the same learners. Based on this information learners
were assigned to statistically distinct language background subgroups based on a range of variables including
parents’ first language, extent of home exposure and/or prior instruction, proximity of the L1 to the taught
standard). Test scores were then used to group learner scripts within each subgroup into ‘high’ and ‘average’
achievement bands. These scripts were analysed qualitatively by expert language teacher panels working
collaboratively with researchers to produce detailed profiles of achievement and annotated exemplars. The
different characteristics of these profiles across languages point to the complexity inherent in any universal
scaling exercise and to the inevitable tension between the institutional demand for uniform goals and scalable
outcomes and the local need for context-sensitive frameworks which highlight learners’ differing learning
trajectories and achievements and associated learning needs.

Session 3: Wednesday, June 4, 10:20

Parallel Session 1.3: Wednesday, June 4, 10:20

Presenter: Yi Mei

Affiliation: Queen's University

Title: Scoring Fairness in Large-scale High-stakes English Language Testing: Examining the National
Matriculation English Test (NMET)

Fairness is “subject to different definitions and interpretations in different social and political circumstances
(AERA, APA, & NCME, 1999, p. 80).” It is therefore essential to investigate test fairness within various
 educational contexts. Empirical research exploring test fairness in scoring written performance has been mostly
conducted in the North American context where group-level fairness concern dominates. There has been little
research conducted in countries such as China where individual difference is the dominant fairness concern,
especially in large-scale high-stakes tests. This study employed Kunnan’s (2004) fairness framework and
collected evidence of construct validity from the rater’s perspective. The research context is the high-stakes
National Matriculation English Test (NMET) in China whose results are used for university admission decisions
involving over 9 million test-takers every year. This study examined scoring fairness by investigating what and
how raters’ scoring decisions were affected by writing elements intended (or unintended) to be measured in
the NMET. This study also further explored whether there was any difference in the writing elements, the
intended (or unintended) test constructs, that novice and experienced NMET raters attended to when rating.
This study was conducted in June 2012 after the NMET was administered, and employed a mixed methods
design. First, six NMET raters each scored five NMET essays while thinking aloud (Ericsson & Simon, 1993), then
attended a follow-up interview about how they made judgment on the five scripts so as to explore, in more
details, the effect of various writing elements on their scoring decisions. Later, a questionnaire was issued to
162 NMET raters to explore their NMET scoring experience. In addition, the NMET-related documents were
also collected for analysis of the NMET writing constructs. Qualitative analysis was conducted thematically, and
descriptive statistics and t-test analysis were used for quantitative analysis. It is found that the NMET raters
generally made scoring decisions based on the intended NMET writing constructs. However, their over-
emphasis on handwriting and superficial understanding of the NMET writing constructs (i.e., following the
superficial meaning of the rating scale without fully understanding the communicativeness of the test
construct) may pose threats to NMET scoring fairness. These fairness-threatening behaviors are closely related
to raters’ prior teaching and NMET rating experience, and the intense nature of the NMET rating sessions. The
findings also indicate that the NMET rating may also be influenced by the raters’ teaching community. Among all the NMET raters, the novice teachers with prior NMET rating experience seemed to attach most importance to the intended constructs. The findings highlight that it is the adequacy with which the raters’ rating foci reflect the intended (unintended) constructs that has enriched our understanding of fairness. Scoring fairness is not the presence or absence of considering the writing constructs within this Chinese large-scale high-stakes testing context. This study has implications not only for the choice and training of NMET raters to ensure scoring fairness, but also for further systematic investigation into scoring fairness in other educational contexts where individual-level fairness is a primary concern.

Parallel Session 1.4: Wednesday, June 4, 10:20

Presenter: Gerriet Janssen and Valerie Meier

Affiliation: University of Hawai‘i at Manoa; Universidad de los Andes–Colombia

Title: Building a better rubric: Towards a more robust description of academic writing proficiency

Because it is common for rubrics to form the basis of raters’ scoring inferences in performance assessments, principled rubric use requires both that a rubric’s initial construction be conceptually and empirically sound and that its validity and reliability be systematically reevaluated as it is adopted into different contexts. Such reevaluations can contribute not only to local test development efforts but also to a more complete understanding of functional language proficiency. The Jacobs et. al. (1981) writing rubric has, in its original and modified forms, enjoyed widespread use in performance tests of second language writing, as well as high visibility in the field (e.g., Weigle, 2002). Although this rubric has been used in studies of rater effects (Kondo-Brown, 2002), instrument development (Brown and Bailey, 1984), and formative task-based assessment (Ke, 2006), to our knowledge, the rubric itself has not been the subject of research. Given its popularity, we have sought to evaluate its reliability and validity in our specific context—a high-stakes English placement exam for incoming PhD students at one Colombian university—and to explore how it can be made a more robust instrument for describing and measuring academic writing development. Here we present the results of two phases of research: a mixed-methods investigation of the original rubric using many-faceted Rasch measurement (MFRM), profile analysis, multiple regression, and rater interviews, and a methodologically parallel investigation of pilot data from our revised version of the rubric. Data for the initial analyses were collected from three years of operational test administrations (n = 524). MFRM was used to estimate rubric category severity and fit, and to model the rating scale structure of each rubric category in order to investigate the frequency and consistency of score use and determine the optimal number of steps for each category scale. Next, profile analysis of the rubric categories by test taker placement level was used to determine the extent to which the rubric categories were differentiated across different proficiency profiles. Multiple regression was used to identify the contribution of each category to predictions about placement levels and to explore how to make the category selections more efficient. Finally, interviews were conducted with all five raters of the writing assessment in order to understand their perspective on the rubrics’ utility. Based on these multiple sources of information, we eliminated one of the five original rubric categories and modified the scale structure of the remaining four. We then piloted our revised rubric using a random sample of 150 essays that were drawn from the original data set and re-scored by two raters. MFRM was used to cross validate these new scores and check for invariance. The data was then analyzed using both profile analysis and multiple regression; follow-up interviews with the raters were also conducted. Our results indicate revising the Jacobs et. al. rubric did improve its reliability and validity in our context and suggest possibilities for a more robust and efficient operationalization of academic writing proficiency.
**Abstract: No descriptor available: plugging gaps in the CEFR**

The Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) has gained widespread recognition as a tool for elaborating and comparing language curricula and assessments. However, serious gaps in the illustrative scales threaten to limit its value for such purposes and restrict its claims to universality. One issue is that relatively few descriptors were included in the framework to cover functional language abilities at the higher levels of proficiency: the C1 and C2 levels of the CEFR. In many cases, the cells representing the highest levels in the illustrative scales state that there is simply ‘No descriptor available’. Building on the work of North (1996) on scaling descriptors that contributed to the development of the CEFR, this paper will report on a project to address the shortage of C level descriptors. The first phase of the project involved identifying language functions or speech acts that appear for the first time at the C levels in materials for English language education. A database of materials designed for use with high proficiency learners was compiled. This included materials addressing the productive, receptive and interactive language skills covered by the CEFR. The materials comprised internationally popular ELT course book series for high proficiency learners, national and local curricula/syllabuses from a range of countries, English language test specifications and proficiency scales. Over 2,000 ‘Can-Do’ descriptors extracted from these sources provided an initial indication of the kinds of functional uses of language that have come to be identified with higher levels of proficiency over at least the past 30 years of educational practice. This is a greater number of descriptors than was originally compiled for the CEFR to cover all levels. An iterative process involving keywords analysis and synthesis of the collected descriptors was then used to generate a list of language functions that might distinguish higher proficiency language learners and to form the basis for a set of sixty new draft Can-Do statements. In a global survey, 750 English language educators then ranked these statements and Rasch scaling with anchoring was used to integrate the results into the existing framework. The results showed high levels of consistency in scaling the statements and the process has helped to fill in many of the gaps in the illustrative scales, adding value to the CEFR. On the other hand, it has also raised questions about the distinctions between the C1 and C2 levels and the interpretation of descriptors in diverse local contexts. References: Council of Europe (2001) The Common European Framework of Reference for Languages. Strasbourg: Council of Europe. North, B. (1996) The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. Unpublished PhD.
such a framework in a country where the number of English learners exceeds that of native speakers around the world conveys no doubt a political message. The experience of the Common European Framework of Reference (CEFR) has shown that a language framework has important impact on society, institutions as well as individuals. A key issue under debate at the outset is whether we should follow the practice of the Eiken Foundation in Japan (for CEFR-J, see www.cefr-j.org/english/index-e.html) by adapting the CEFR to the Chinese context, or whether we should develop a framework specific to the Chinese context. In this paper, this issue will be explored from the perspective of language education at both macro and micropolitical levels (Alderson 2009). The educational system in China is highly segmented with different governmental departments taking charge of education at each key stage: elementary, secondary, vocational, undergraduate and postgraduate. National curricula and assessments are developed and implemented for each key stage under the auspices of these different departments. Coordinating resources and reconciling and negotiating the different views of professionals and policymakers at different levels are therefore particularly challenging due to an inherent lack of communication among administrators, researchers and practitioners working with different departments. Decisions on varieties of English to be taught and the age to begin learning English are also political at the macro level. In terms of micropolitics, the project will meet considerable institutional and individual resistance: publishers may not be willing to risk losing their current market share; exam boards may be reluctant to align their tests to the new standards; and teachers will be pressed to take on a new set of requirements and adjust to new textbooks. We argue that the aim of the project is not to search for or borrow a yardstick against which learners’ proficiency can be measured, but to develop, validate and continuously improve the yardstick. This ongoing process intends to improve the coherence and transparency of English language education at different levels. More importantly, it will increase the acceptability of the CCFR because stakeholders, including policymakers, will be involved in or informed of the development and validation of the framework. Learners’ performance on local and international English language tests as well as Dialang-China, a modified version of Dialang (www.lancaster.ac.uk/researchenterprise/dialang/about) will be analyzed to unpack the cognitive processes of Chinese EFL learners and supplement the English Profile Program with the criterial features of Chinese learners. In summary, we argue that the development of a language framework is not simply a professional endeavor but a complex political initiative, at both macro and micro levels.

Parallel Session 2.3: Wednesday, June 4, 11:25

Presenter: Wei Yan

Affiliation: Queen's University

Title: Determinants of Chinese students' academic success in Korean universities

There are more than 80,000 Chinese students studying in South Korean universities (Shin, 2011). In order to gain admissions, these students are required to submit their scores on the Test of Proficiency in Korean (TOPIK). Once admitted, students will study in Korean medium of instruction. Despite the wide acceptance of TOPIK scores as admission criteria among Korean universities, little research has been conducted to examine whether TOPIK measures the Korean proficiency needed in academic settings. Further, English also plays a powerful role in academic studies of Korean universities, and high English proficiency is beneficial to students studying in Korean universities (Jambor, 2011). Therefore, both Korean and English proficiency could contribute to the academic success of Chinese students along with educational and psychological factors such as prior achievement and academic self-efficacy (Richardson et al., 2012; Robbins et al., 2004). The results from previous studies on international students’ language proficiency and academic success are, however, mixed and highly inclusive (Cho & Bridgeman, 2012). This study addressed the research gap by investigating the key determinants of Chinese students’ academic success indicated by their first semester GPA and credit hours earned in Korean universities. The determinants include demographics, prior achievement, academic self-
efficacy, and Korean and English language proficiency. Questionnaires were issued and collected from 138 undergraduate and 63 graduate Chinese students studying in 27 different Korean universities. The questionnaire consists of four sections: demographic information, academic background, language proficiency and psychological factors. Multiple regression and correlation analyses were conducted to explore the predictors of Chinese students’ academic success in Korean universities, especially the contribution of language proficiency. The results indicate that both undergraduate and graduate participants’ first semester GPAs are significantly correlated with the TOPIK score along with other Korean proficiency indicators. Both of them are correlated with English proficiency indicators, too. The hierarchical regression models accounted for 20% and 38% of the variance in undergraduate and graduate participants’ first semester GPA, respectively. Korean proficiency added an extra 6% to the contribution of demographic, educational and psychological factors in the undergraduate model. For graduate participants, in the presence of academic self-efficacy, Korean and English proficiency added 22% to the contribution of demographic and educational factors. Korean proficiency indicators including the TOPIK score accounted for 22% of the variance in graduate participants’ first semester credit hours. These findings suggest that TOPIK has a significant small to medium predictive effect on academic success and serves well as an admission requirement, especially for graduate students. The findings also highlight that both Korean and English proficiency influence on academic success in this particular context. The relationship of language proficiency and academic success is complicated, and this study contributes to the field by exploring this relationship in a less researched test, TOPIK and by investigating two languages at the same time. This study also contributes to the overall understanding of international university students’ academic success, in particular the success of Chinese students.

**Parallel Session 2.4: Wednesday, June 4, 11:25**

**Presenter:** Yuanyuan Guan  
**Affiliation:** James Cook University  
**Title:** Using MCQs and stimulated recall protocols to diagnose listening comprehension processes

Listening comprehension, as an invisible complex process, involves activation and application of a set of cognitive functions and skills. This makes the diagnosis of listening comprehension a very challenging task. Amongst various techniques to measure listening comprehension, multiple-choice questions (MCQs) have been extensively used tests given its convenient scoring and low cost. However, it is also challenged by researchers that it provides little information on what test takers actually perceive and how they process the spoken text. Admittedly, psychometric analysis provides valuable insights into listening sub-skills, it is more important that these findings be justified by qualitative data. As Ross (1997) asserted, the immediate retrospection method “provides a useful tool for investigating the psycholinguistic validity of item response patterns and can offer detailed qualitative data to supplement traditional and probabilistic approaches to test analysis” (p.219). This paper aims to investigate the effectiveness of using multiple-choice questions and stimulated recall protocols to diagnose test takers’ listening comprehension performance on the listening component of the Diagnostic English Language Assessment (DELTa). The DELTA test was administered to approximately 2000 first-year and second-year university students in three Hong Kong universities. The item bank of the listening component consisted of 33 spoken texts, with five to eight multiple-choice questions for each text. Each MCQ is intended to measure a particular listening sub-skill which is assumed to be essential in understanding general or academic spoken texts. These listening sub-skills include: identifying specific information, understanding main ideas and supporting ideas, understanding information and making an inference, etc.. By adopting the Rasch measurement, all the MCQs were calibrated and analysed to determine their difficulty levels in the form of item difficulty measures. The Item Characteristic Curves (ICCs) generated from the WinSteps software indicated the expected and observed responses of each item. High and low ability test takers were invited to re-take the
Abstracts June 4

test a month later in a simulated test situation, where their test-taking process was video-recorded. In the subsequent Stimulated Recall Process (SRP) they were asked to recall and verbalise what they were thinking and how they processed the four options of MCQs and worked out the answer to each question. The SRP data were analysed by using NVivo. The presentation will first report on item and person performance of the DELTA listening component. It will show how different groups of test-takers perform on easy or difficult items relative to their listening proficiencies. Their cognitive processings of items of various item intents will be compared. The diagnostic information of the MCQ options will be discussed with implications of the diagnosis effectiveness of MCQs.

Parallel Session 3.1: Wednesday, June 4, 12:00

Presenter: Junghee Byun and Jungyeon Koo

Affiliation: Seoul National University

Title: The L1-L2 Comparability of Measuring the Writing Construct: Are We Measuring the Same Construct?

Motivated by the interconnectedness hypothesis of L1 discourse competence and L2 writing literacy (Cummins, 1981), this study attempts to explore the comparability of assessing the discourse competence in Korean (L1) and English (L2) writing, focusing on the non-linguistic features. For this, 31 college students wrote a pair of argumentative essays - one in Korean and the other in English on the same topic. And 6 raters with half for Korean and the other half for English scored the writing samples. The analytic writing scale with common features being extracted from two sources of scoring rubrics in each language was developed into 4 criteria - task fulfillment, organization, elaboration of idea, and coherence. The obtained data were analyzed by the multi-facet Rasch model, two-way mixed-repeated measures ANOVA, and multiple regressions. The results indicate the potential comparability of Korean and English discourse competence, which will add further evidence to the aforementioned hypothesis and serve to advance L2 writing development and L2 writing pedagogy.

Parallel Session 3.2: Wednesday, June 4, 12:00

Presenter: Lieve De Wachter and Jordi Heeren

Affiliation: KU Leuven

Title: The construction and validation of an academic literacy test: providing first-year university students with an early (warning) signal related to academic achievement.

Academic bachelor programs at Belgian universities, and more specifically at the University of Leuven are confronted with a low academic success rate of first year students: only 38.76% of the students passed all their exams in 2013. De Wachter & Cuppens (2010) make it clear that not all students are equally well prepared for a university education. Several retention studies focus on cognitive variables such as high school GPA as typical predictors of study success (Tyson 2011; Veenstra et al. 2008; Zhang et al. 2004), others on motivational characteristics (Guay et al. 2008; Marsh & Craven 2006). This article will bring up yet another powerful predictor of study success: academic literacy (Brown & Hudson 2002; De Wachter & Heeren 2013; Van Dyk 2010; Holder et al. 1999; De Wachter, Heeren, Marx & Huyghe 2013; Peters, Van Houtven & Morabit 2010). This presentation will discuss the validity of the results of an academic literacy test developed at the Leuven Language Institute (KU Leuven). Together with the Teaching and Learning Department of the university of Leuven, the external validity of that test has been evaluated. The language test results of 2660 KU Leuven
students were correlated with the results of their January exams (De Wachter, Heeren, Marx & Huyghe 2013). There was a positive and very significant correlation between the language test scores and the exam results ($r=0.35$, $p<0.001$). On average, if a student scored lower on the language test, he/she scored lower on the exams and vice versa, although there were a lot of exceptions to this general tendency. Nevertheless, 72% of the students who failed the academic literacy test did not pass their January exams. Research thus confirms the claim that the test can be an early warning signal for students regardless of their field of study. At the same time the group of students that pass the language test have to be aware of the fact that passing the language test is no guarantee for academic success: it is a necessary, but not a sufficient condition for study success. This also implies that a very high correlation would be impossible, as academic achievement depends on much more than language skills alone. The results of the correlation study show that the academic literacy test developed at the Leuven Language Institute may not be a strict predictor of study success but it does select an at-risk target audience. The observation that academic language skills are an important indicator of students’ future academic achievement, regardless of their field of study, suggests that a general academic language skills test could be used to inform both science and non-science students about their chances on early academic achievement. Even though the test is rather low-stakes and the remedial courses and tools are not obligatory, it is necessary to keep validating the test. Test development has to be seen as a continuous process of designing, testing and revising in order to meet the intended purpose as accurately as possible.

Parallel Session 3.3. Wednesday, June 4, 12:00

Presenter: Beverly Baker

Affiliation: McGill University

Title: Investigating Language Assessment Literacy in Canadian University Admissions

There are increasing numbers of non-native English speaking applicants to English Canadian universities (AUCC 2008a, 2010). One result of this trend is that university admissions officers, as gatekeepers, are faced with a growing and potentially confusing array of language test scores when making their decisions. These admissions decision makers need a certain amount of language assessment literacy (LAL) to enable them to make use of these language test scores effectively and ethically (O’Loughlin 2011, 2013). This research project examines a collaboration between language assessment professionals and Canadian university admissions decision makers in addressing this challenge. Our research questions were the following: 1) What is the LAL needed for users of language test scores in admissions decision-making at post secondary institutions in Canada? 2) Once the LAL needs of this group are established, what is the nature of the materials needed to develop this LAL? The first phase of the project consisted of a survey with Canadian university admissions decision makers, informing Phase 2—the development of a series of workshops held at 8 universities representing every region of the country. The final phase of the project was the creation of resource materials for all Canadian post-secondary institutions. Research data were collected in both Phase 1 and Phase 2. In the survey, admissions decision makers were asked about their knowledge, beliefs, and levels of confidence in making use of language test scores in their decision-making. During the workshops, participants completed a questionnaire to evaluate the workshop content. In addition, participants were recorded during the workshops as they discussed issues of language assessment related to their work. All data were analysed with content analyses as well as techniques associated with grounded theory (Strauss & Corbin, 1998). Results suggest that these stakeholders are generally confident in their use of test scores in decision-making—and this confidence increases when they inquire into the subsequent academic performance of admitted students. However, there is no evidence of awareness that the various tests may be addressing different constructs. Despite this general feeling of confidence, participants voiced concerns about issues of security and testing fraud. While some evidence of misunderstanding was identified, respondents demonstrated often keen awareness of key concepts related to
validity in language assessment, albeit without making use of the conventionalized language of the field. It was determined that the most beneficial materials for these stakeholders would not include a formal treatment of validity and validation, as might be found in LAL materials for language teachers or other stakeholders. Instead, the concept of validity was most usefully framed for this audience in terms of confidence, i.e. confidence in interpretations made from test scores, and what evidence would be necessary to increase this confidence while minimizing negative impacts. In this presentation, we discuss the contribution of this study to the new research area of LAL, as well as the benefit of framing LAL as competency development rather than the building of a knowledge base.

Parallel Session 3.4: Wednesday, June 4, 12:00

Presenter: Maryam Wagner

Affiliation: Ontario Institute for Studies in Education, University of Toronto

Title: Use of a Diagnostic Rubric for Assessing Writing: Students’ Perceptions of Cognitive Diagnostic Feedback

Feedback has been widely researched in second language (L2) education, especially in L2 writing (e.g., Ferris, 2003; Lee, 2003; Lyster & Ranta, 1997) with much of the research focusing on the effectiveness of the types and delivery modes of feedback, particularly among university-aged learners. Despite the numerous studies there is a lack of consensus and many unanswered questions (Hyland & Hyland, 2006). Much of this literature has focused on learning outcomes rather than on cognitive processes and strategies (Author). Furthermore, investigations of feedback have been place primarily focused on whether: 1) errors and/or concerns have been identified in the writing; 2) the writing has included comments; and 3) the feedback has provided corrections and or descriptions of errors (Biber, Nekrasova, & Horn, 2011). The focus has not been on evaluating the degree to which feedback helps students to identify cognitive gaps. This type of feedback may be referred to as cognitive feedback (Balzer et al., 1989; Butler & Winne, 1995). This research study developed and validated a Diagnostic Rubric for Assessing Writing (DRAW) in order to investigate: 1) the extent to which the DRAW provided diagnostic writing information for adolescent English language learners (ELLs), and its sub-skills and descriptors allowed for discrimination between adolescents with differing writing proficiencies; 2) how students’ goal orientations interacted with students’ perceptions of feedback from the DRAW. The research engaged 75 ELLs and their teachers from three 10th Grade classes in different schools in a Canadian city. The basis of the DRAW was a language assessment framework developed by the local government. The framework was revised and validated through an iterative process with the teachers resulting in the DRAW which included diagnostic assessment descriptors distributed among six writing sub-skills: Ideas, Organization, Vocabulary, Sentences, Mechanics, and Grammar. The DRAW was used to evaluate two genres of students’ writing at the start of the investigation, at the mid-point, and at its conclusion. One writing sample (mid-point) was used to generate individualized writing profiles for students containing diagnostic feedback about their cognitive writing skills. Using multiple data sources (e.g., interviews, surveys), students’ perceptions of feedback were used to develop profiles of students according to their preferred: type of feedback (cognitive/evaluative); source of feedback (external/internal); focus of feedback (e.g., content/organization); and identified use of feedback (goals/teacher expectations). Students also responded to a goal orientation survey. Three feedback profiles emerged: 1) Maximizers: students who value cognitive feedback generated both internally and externally, who focus on the ideas as well as self-identified weaknesses, and who use feedback to achieve goals; 2) Minimizers: students who prefer evaluative feedback from external agents, who focus primarily on grammar and use feedback only for directed purposes; 3) Seekers: students who exhibit a varying combination of feedback qualities at neither of the Maximizers/Minimizers extremes. This paper presents exemplar case studies of each of these profiles using multiple explanatory variables including: students’ writing achievement,
Abstracts June 4

teachers’ classroom goal structure and assessment literacies, student’ goal orientations, and the school context to elucidate the relationship between feedback and writing development.

**Parallel Session 4.1: Wednesday, June 4, 14:00**

**Presenter:** Kinnie Kin Yee Chan

**Affiliation:** The Open University of Hong Kong

**Title:** Computer vs. Human Scoring in the Assessment of Hong Kong Students’ English Essays

Maintaining fairness and consistency in essay scoring both in the classroom and high-stakes assessment is crucial as the results of essay scoring will impact the development of students either directly or indirectly. Thus, teachers, students and parents are under considerable pressure in the educational system in Hong Kong. Automated Essay Scoring (AES) employs computer technology to evaluate and score written prose in the place of the usual human grading. It has become more widely accepted as an educational technology for both assessment and classroom instruction, but little research has been done on validating the performance of the AES models in an applied educational context, and only some foundational work has been performed in their calibration against human ratings. This study investigated the effectiveness of the Lexile Analyzer, an Automated Essay Scoring (AES) system, in measuring English language essay writing ability of Hong Kong school students. 137 Grade 6 Hong Kong students responded to three essay writing prompts from the US National Assessment of Educational Progress (NAEP) and essays were scored by their four experienced Hong Kong English teachers. They used the NAEP holistic essay marking rubrics for three genres: narrative, informative and persuasive. The Lexile Analyzer and Hong Kong human rater scales were co-calibrated by applying the Many-facets Rasch Measurement (MFRM) model. The MFRM was implemented to adjust for prompt difficulty and to detect rater severity in judging essay writing, because Rasch measurement models provide a framework for obtaining objective and fair measures of writing ability which remain invariant over raters, writing tasks and other aspects of the writing assessment process. The MFRM analyses revealed that 4 Hong Kong teachers of English scored essays consistently. Moreover, the Lexile Analyzer scored consistently with the trained and professional human raters in this research. For the essay writing performance of students, most of the misfitting cases were students who performed erratically, produced off-topic essays or wrote essays that were scored by the most severe or lenient rater(s) in this research. For example, the students misinterpreted the instructions of the prompt; as a result, the ratings for that prompt were far worse than the students’ average performances on the other prompts. This research was conducted to calibrate the reliability of the AES system, the Lexile Analyzer, so teachers, students and parents are provided insight into the nature of the essay scoring process. Most importantly, the research probes issues about the fairness and consistency of essay scoring when the Lexile Analyzer is used as an assessment tool and highlights the time- and stress-saving potential of this tool for Hong Kong language teachers by overcoming problems of validity and reliability and by encouraging students’ independence during their English language learning.

**Parallel Session 4.2: Wednesday, June 4, 14:00**

**Presenter:** Kirby Grabowski

**Affiliation:** Teachers College, Columbia University

**Title:** Describing pragmatic performance at a range of ability levels on a role-play speaking test
Research in applied linguistics has demonstrated that language used in the real world is often indirect, and misunderstandings in interaction often stem from a failure by interlocutors to appropriately convey and/or interpret implied, or pragmatic, meanings. Given their relative importance in everyday communicative contexts, pragmatic aspects of language have been integrated into current influential models of language ability. However, even with these models as a guideline for test development, pragmatic aspects have remained, in large part, noticeably absent not only in both large- and small-scale assessments, but also in the descriptive scales meant to inform the underlying constructs associated with those tests. As a result, it has been difficult to make inferences about a test taker’s overall language ability as it pertains to real world contexts. For the current study, a test was developed to measure test takers’ grammatical and pragmatic knowledge in the context of speaking. Test takers from four different levels of ability participated in four role-play tasks, representing real world situations, in which each test taker was paired with the same native speaker partner. The conversations were recorded and then scored by two raters. Although the data were analyzed statistically using both multivariate generalizability theory and many-facet Rasch measurement in a separate analysis, the purpose of the current study is to show how raters were able to isolate and score the appropriateness of a range of pragmatic meanings in the responses. To accomplish this, a discourse analysis of pragmatic meanings was performed for one task across four ability levels, highlighting contextualization cues to uncover degrees of pragmatic success and failure in the performance samples. Findings from the statistical analysis will also be briefly discussed in order to support the qualitative analyses.

Symposium 1: Wednesday, June 4, 14:00

Presenters: Gary Ockey Educational Testing Service, Paula Winke Michigan State University, Paul Gruba University of Melbourne, and Gary Buck Lidget Green

Title: Critical Issues in Assessing L2 Listening Ability: Construct and Consequential Validity

This symposium brings together empirical and theoretical papers that investigate current, critical issues related to the assessment of L2 listening ability. Each of the four papers adds an argument to the same common theme: how the construct of L2 listening ability is defined and operationalized impacts the construct validity of L2 listening assessments. This common emphasis addresses the conference theme “Towards a Universal Framework” by furthering understanding that listening is a multi-faceted construct, and by suggesting ways in which such a listening construct can best be assessed. The papers investigate L2 listening ability in relation to real world functional skills by acknowledging the complex realities of current understandings of communicative language ability. The first paper uses quasi-experimental design to compare two groups’ performance on L2 listening tests that have scripted spoken texts (used on most traditional tests of L2 listening) versus unscripted spoken texts (representative of most real world spoken discourse). The author argues that using only scripted texts in L2 listening tests results in an overly-narrow definition of the construct. The second paper reports on a large-scale study (n = 21,726) that examines how test-taker performance varies when using spoken texts that have a variety of accents on L2 listening tests. Test-takers listened to texts with different accents and answered comprehension items. ANCOVA was then used to examine how familiarity with a particular accent affected performance. The author argues that using only scripted texts in L2 listening tests results in an overly-narrow definition of the construct. The second paper reports on a large-scale study (n = 21,726) that examines how test-taker performance varies when using spoken texts that have a variety of accents on L2 listening tests. Test-takers listened to texts with different accents and answered comprehension items. ANCOVA was then used to examine how familiarity with a particular accent affected performance. The author argues that the inclusion of multiple accents on assessments, while theoretically justifiable, can present threats to test fairness. The third paper uses quasi-experimental design to examine test preparation and testwiseness in relation to multiple-choice and fill-in-the-blank test task response formats. Test-takers were placed into a control group, an explicit test preparation group, and an implicit test preparation group. In addition to performance data, eye tracking technology was used to examine test-takers’ eye movements while taking the exams. The author argues that test-takers’ difficulties in responding to listening test items may stem from the item format, and explores how the use of different task response formats affects the measurement of the construct. The fourth paper explores the key issues for L2 listening assessment that arise from the use of new media as the mode of presentation, including issues such as task
design, text selection, and listener characteristics, concluding that L2 listening is best understood as a multifaceted construct that must be acknowledged by test developers. In this vein, the paper includes a multi-media presentation on the issue of multi-media and listening assessment. The symposium concludes with a discussant summarizing and critiquing the four papers, and critiquing and problematizing the notion of assessing a multifaceted construct of listening in the quest for a “universal framework”. Each paper is 15 minutes long, and the discussant presentation is 10 minutes. The ultimate goal of the symposium is not only to investigate how these issues affect the construct validity of L2 listening assessments, but also to explore the consequences that their use have on test-takers and test users, as well as on L2 learning and teaching.

**Symposium 2:** Wednesday, June 4, 14:00

**Presenters:** Tobias Haug University of Applied Sciences of Special Needs Education Zurich, Peter Hauser Rochester Institute of Technology, Wolfgang Mann City University London, and Krister Schönström Stockholm University

**Title:** Issues and challenges related to signed language assessment

Signed languages are natural languages that make use of visual communication channels, using hands, arms, face, and the upper body to express something and the eyes for understanding. Users of signed languages include deaf children with deaf/hearing parents, hearing children of deaf parents, and hearing and deaf adults, who learn sign language as a second language (L2). As with spoken languages, it is necessary to have instruments that can assess the sign language development of these learners. The aim of this symposium is to present complementary perspectives on current (and future) issues and challenges related to sign language assessment and the efforts to document learners’ strength and weaknesses. The development of sign language tests and its subsequent discussion on research, theory and practice is fairly new, dating back only to the late 1990’s. This is in part because research on signed languages, such as American Sign Language (ASL) or British Sign Language (BSL) is a rather young field compared to spoken languages. Consequently, the number of existing sign language assessments is limited with a focus on ASL and BSL. The purposes of these tests vary and may include assessing signed language development in L1 or L2 learners or supporting research. Also, there are several challenges to test development, including regional sign variations and the small population size of native signers, who serve as norm group. In our symposium, an international team of deaf and hearing presenters will discuss different aspects related to the use of sign language assessments to support learners by providing evidence of their strength and weaknesses. The first talk addresses issues related to the adaptation of existing sign language tests to other signed languages, based on an example from ASL and Swedish Sign Language. The second talk discusses the use of dynamic assessment procedures to measure signers’ response to mediated learning experiences with focus on vocabulary development in ASL. The third talk discusses item development and psychometric analyses needed for testing different L1 and L2 populations. The final talk focuses on the use of new technologies in signed language assessment and the potential advantages of such technologies, including web-based test formats, automatic sign recognition, and signing avatar have for sign language learners. The importance of gathering and discussing issues and challenges in the area of sign language assessment cannot be stressed enough. This symposium will contribute towards a better knowledge basis for developing future tests to effectively document strength and weaknesses by different groups of sign language learners. 10’: Christian Rathmann, Hamburg University, Germany (moderator) 50’ (presenters): Presenter 1: Krister Schönström, Stockholm University, Sweden: Adaptation of sign language tests Presenter 2: Wolfgang Mann, City University London, UK: Dynamic assessment with deaf populations Presenter 3: Peter Hauser, Rochester Institute of Technology, USA: Sign language tests for different populations Presenter 4: Tobias Haug, University of Applied Sciences of Special Needs Education Zurich, Switzerland: Use and implementation of new technologies in sign language testing 10’: Beppie van den Bogaerde, Hogskole Utrecht, Netherlands (discussant): 10’: Rebuttal from presenters 10’: Discussion
**Title: Using the CEFR as the foundation for an academic rating scale.**

When used as a rating scale for assessing general language proficiency, the CEFR constitutes a flawed instrument that shows vagueness, inconsistencies and overlaps. When used for LAP assessment, the same problems apply but additional issues arise. Firstly, the CEFR does not cover all the metalinguistic and cognitive skills required for authentic academic discourse. Secondly, those skills that the CEFR does cover, are not dealt with systematically or consistently. It follows logically then, that the CEFR does not allow for an immediate operationalization as an LAP rating tool. For that reason, the Certificate of Dutch as a Foreign Language (CNaVT) has not simply adopted the CEFR descriptors in its LAP rating scale. Instead, an iterative data-driven rating scale construction process has served to overcome some inherent weaknesses of the CEFR. Subject specialists were consulted to establish the main rating criteria and raters offered input concerning the wording and interpretation of the descriptors. This presentation retraces the steps taken to refine and rework the CEFR into a valid and reliable rating scale for a test of Dutch for academic purposes and wishes to answer two research questions: 1. Can an empirical and iterative data-driven construction process eliminate the inherent weaknesses of the CEFR as an LAP rating scale? 2. Does co-constructing a rating scale with novice raters help to stimulate a shared interpretation of its criteria? In order to answer these research questions, the presenters will summarize the main points of the iterative three-year rating scale construction and validation process. This overview includes qualitative data from focus groups with subject specialists and raters as well as reliability analyses of three pilot studies undertaken in 2011 and 2012. It will show how the current four-band scale came about as well as why and how the current descriptors differ from the original wording. The focal point of the presentation however is the final leg of the rating scale validation, undertaken in October-November 2013. For this, six trained novice raters judged 200 written and spoken performances, using a missing data design whereby each performance was rated three times. After the rating process, all six raters took part in a two-hour focus group. The quantitative data was analysed using both Classical Test Theory statistics and IRT. The focus group data was transcribed and double coded using NVivo 10. The analyses show that even after four pilot runs and after an intensive iterative data-driven rewriting process in conjunction with the raters, some original CEFR flaws persist, both on the level of usability as on the level of reliability. These problems can be grouped into four categories: broadness of a criterion, non-exhaustiveness of a descriptor, overlap among criteria and vagueness of a descriptor. The data also shows however that by following an iterative-data driven construction process the major shortcomings of most CEFR scales can be alleviated, leading to a valid and reliable scale that discriminates well for most criteria and allows for consistent rating.

**Title: Impact of Response Length on Reliability of Automated Scoring for Short Spoken Responses**

Language assessments designed to measure spoken English proficiency typically use constructed -response items of varying length and type to provide an adequate construct coverage, and to assess test-takers with a wide range of spoken English proficiency levels. Depending on the item type and the scale, automated scoring using speech scoring engine is a viable option for some items. One such item type is a sentence-repeat item
where test-takers provide verbatim repetition of sentences they listen to as part of the item. Each item in this case is typically one sentence, and the examinee’s spoken response is scored using a scale designed to evaluate the accuracy, pronunciation, and clarity of the spoken response. This research investigates the impact of response length on the reliability of automated scoring models for items that elicit short and highly predictable responses such as for sentence-repeat items. Memory constraints favor the use of rather short items, typically 5 to 15 words of length, but concatenation of multiple responses to achieve a greater response length can allow more robust extraction of features and improve the reliability of the scores. Data were gathered for 16 sentence-repeat items designed as part of an assessment to evaluate spoken English proficiency of non-native speakers for employment opportunities. The items were presented in sets of four and scored on a scale of 2-6. Data from roughly 2,500 test-takers were used for the study. Automated scoring models were developed to estimate the proficiency scores on the basis of features related to fluency, pronunciation, and repetition accuracy. Features were calculated using a fully automated process based on the output of an automatic speech recognition system. The scoring models were constructed in two different ways: (i) Individual item scoring model that calculated a score for each individual item; (ii) Group item scoring model for which sets of four thematically-linked responses were concatenated to calculate one automated score. The group item scoring model outperformed the individual item scoring model on evaluation criteria of agreement (weighted kappa, Pearson correlation) between human and automated scores. These results were based on the evaluation of individual item scores or group item scores (for four items). We will further evaluate aggregated automated scores without concatenation of responses, and also investigate whether the high performance of group item-based approach is still maintained at the test-taker level. The additional analyses will be completed by the end of this year and will be therefore ready in time for presentation at the conference. This research supports increased reliable scoring of items that elicit short responses from speakers of varying proficiency levels, with implications for more robust identification of candidates’ strengths and weaknesses on the target construct.

Parallel Session 6.1: Wednesday, June 4, 15:10

Presenter: Lorena Llosa

Affiliation: NYU

Title: The Role of Linguistic Complexity in Content Assessments

English language learners’ (ELLs’) performance on content area standardized tests is consistently lower than that of their non-ELL peers. One explanation for ELLs’ lower performance is the linguistic challenge of the tests themselves. Studies have reported correlations between the linguistic complexity of content area test items and the items’ level of differential item functioning (DIF) against English Language Learners. These correlations have been interpreted to mean that linguistic complexity causes construct-irrelevant variance against ELLs on content area tests. As a result, linguistic modification has been used as an accommodation for ELLs on content area tests in order to eliminate potential test bias against ELLs. This interpretation of DIF evidence and resulting linguistic modification accommodation depend on the assumption that complex language forms are always irrelevant to content area achievement. In contrast, this paper argues that linguistic complexity in test items is not a source of construct irrelevant variance simply by virtue of its complexity. We address the issue of language-related construct-irrelevant variance on content area tests from the perspective of systemic functional linguistics (SFL), which posits that language is integral to content area achievement. An SFL perspective suggests that specific language forms may function to facilitate scientific discourse and may therefore be relevant to science achievement. This perspective suggests the need for a mechanism by which to judge the relevance of a language form to a content area, independent of the form’s complexity. We propose a ‘matching’ mechanism for determining the construct relevance of lexical, syntactic, and text level language
forms based on Bachman and Palmer's (1996) concept of a TLU domain. We propose that relevance, and thus related claims of construct-irrelevant variance and bias, should be judged according to the degree of correspondence between language use in the assessment and language use in the educational contexts in which the content is learned and used. In order to demonstrate this approach, we first present findings from literature that documents the language forms, lexical, syntactic, and textual, of school science discourse. Next, we compare these forms to language forms assumed to be irrelevant to science achievement in previous studies of content area test bias and accommodations. The comparison reveals that many of the language forms previously assumed to cause construct-irrelevant variance on content area tests are forms which have been well-documented in school science discourse. This finding weakens the assumption that these forms are irrelevant to the science achievement construct by virtue of their complexity and thus complicates previous DIF-based arguments for test item bias. In addition, results suggest that participation in school science discourse includes use of text-level language forms that are not typically tested on standardized science tests. Therefore, we suggest that future construct definitions for content area tests should explicitly include grade-appropriate levels of proficiency with the language of the content area at the lexical, syntactic, and text levels.

Parallel Session 6.2: Wednesday, June 4, 15:10

Presenter: Megan Montee and Margaret Malone

Affiliation: Center for Applied Linguistics

Title: Computerized oral proficiency assessment and young language learners: Comparing performances in two contexts

Are performances on a computer-delivered oral proficiency tests consistent with examinee ability? When developing tasks to assess academic speaking proficiency, it is vital to authentically elicit the oral language students use in real-world academic contexts. At the same time, there is limited research about the nature of the development of academic oral language of young English language learners in United States (U.S.) schools. By comparing examinee performances in test and non-test situations, this paper presents research that supports the validity argument of a computerized oral language test for young learners. While previous research has compared student performances in different testing situations, such as face-to-face and semi-direct (Shohamy, 1994; Kenyon, Malabonga and Carpenter, 2005) methods, little research has addressed similarities and differences between young learner performances in test and non-test situations. The focus for this study is the speaking section of an academic English language proficiency test designed for English language learners in grades 1-12 to be used by 30 states in the U.S. The test, which will be made operational in 2015, will be administered annually for accountability and program purposes. The speaking section will include computer-administered speaking tasks that will be centrally scored by trained raters. In this qualitative study, we compare student test performance with performance in a face-to-face research interview conducted as part of cognitive laboratories about test items. Research questions include: 1) Are young learners’ ratings on computerized tasks consistent with ratings of language produced in a non-test situation? 2) How do the linguistic features of language produced in each context vary within and across language functions? The research questions are addressed through analyzing test responses and cognitive interview data from English language learners (N=30) with differing levels of English proficiency in grades 4-5 in U.S. public schools. Each learner completed a series of test tasks followed by a stimulated recall interview (Gass and Mackey, 2009). For each student, responses from two test tasks with different functions, or communicative purposes, were rated, transcribed and coded. Portions of each student’s stimulated recall interview were also rated using the same procedures. Interview data was also transcribed and analyzed for vocabulary use, fluency, and linguistic complexity. Results include a review of correlations between computer task and interview ratings, as well as similarities and differences between the features of examinee language within each context. The results of this
study have implications for the design of computer-based oral proficiency tasks in general and for young learners in particular, and provide evidence about the extent to which performances on these tasks reflect language ability.

Parallel Session 7.1: Wednesday, June 4, 16:10

Presenter: Xiangdong Gu and Qiaozhen Yan

Affiliation: Chongqing University/Cambridge English

Title: A Pilot Study into Aligning Chinese English Majors’ Language Proficiency Levels to the CEFR

The Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR), as a widely used common framework of reference based on six broad reference levels and an ‘action-oriented’ approach to language teaching and learning, has become highly influential in language and education policy within Europe and the wider world. It is being adopted by many countries to help define language proficiency levels with resulting implications for local pedagogy and assessment. In keeping with the globalization more generally around the world and in keeping with the expansion of the Chinese economy and its international profile, English has become widely accepted in China as a utilitarian tool for international mobility and study purposes, and English language education in China has also become internationalized. In this context, there is a strong need and a significant meaning to explore Chinese students’ English proficiency levels with reference to the CEFR. The present paper is a pilot study aligning Chinese English majors’ language proficiency levels to the CEFR. We adapted the self-assessment checklists consisting of five common reference levels (from A2 to C2) from the European Language Portfolio (ELP) to a five-point Likert scale (from 1: I can’t do it at all to 5: I can absolutely do it) and employed them as research instruments. The research subjects were 400 English majors of four grades (freshmen, sophomores, juniors and seniors) from a top university in southwest China. For each grade of the students, two checklists with adjacent levels, i.e. freshmen A2 and B1, sophomores B1 and B2, juniors B2 and C1, and seniors C1 and C2, were distributed to them separately within a time interval of two weeks. Data were collected and analyzed to explore which proficiency level each grade of the Chinese English majors have reached with regard to the CEFR. The preliminary findings of the study are discussed and implications for English language teaching, learning and assessment in the Chinese context and beyond are considered. It is hoped that the study would set as a trial to align Chinese English learners’ English proficiency levels to the CEFR in a nationwide context in China. The findings and implications are thought to have relevance to language and education policy, pedagogy and assessment. Therefore, it is also hoped that our study will encourage more researchers to discuss the relevance of the findings from this study to the context and assessment they are familiar with.

Parallel Session 7.2: Wednesday, June 4, 16:10

Presenter: Jonathan Schmidgall

Affiliation: Educational Testing Service

Title: The impact of 'real-world' construct under-representation on decision errors for a test of oral proficiency

This study investigates the potential impact of ‘real-world’ construct underrepresentation on decision errors for a test of oral proficiency (TOP). Schmidgall (2013) examined factors that influenced naive listener perceptions of a speaker’s comprehensibility in a TLU domain (a ‘real-world’ construct), and found that both
speaker- and listener-related factors influenced whether the speaker was judged to be functionally competent. While some of these factors were relevant to the construct of oral proficiency as defined by the TOP (e.g., speaker pronunciation), other factors were not relevant to the TOP’s construct definition (e.g., speaker teaching skills, listener attitude homophily, listener interest in the topic). Conversely, the TOP may underrepresent the ‘real-world’ construct as defined by naïve listener perceptions. A consequence of the misalignment between the TOP’s and naïve listeners’ evaluations of oral proficiency might be different views of what constitutes classification or decision error. Decision error – or the rate of false positives and false negatives for each decision category – is typically viewed as a function of the consistency of test scores and the location of the cut scores that define the decision categories. Another way to quantify decision error is to compare decisions based on test scores to those based on a ‘gold standard’ (Kupermintz, 1999). In the case of the TOP, decision error could be investigated by comparing decisions based on TOP scores (Pass, Provisional Pass, or Fail) to those based on a gold standard, the ‘real-world’ construct. Ultimately, ‘real-world’ construct under-representation may only concern TOP administrators to the extent that it impacts decision errors. In Bachman and Palmer’s (2010) assessment use argument (AUA), an argument-based approach to validity, an important claim regarding score interpretations is that these interpretations provide information that is sufficient to support decision-making. If the construct articulated by a test’s design statement under-represent the ‘real-world’ construct, test score interpretations may provide insufficient information for decision-making. This aspect of validity has been comparatively under-researched in language testing publications (Choi & Schmidgall, 2010). Several analyses were conducted using TOP decisions as predictors of naïve listener evaluations. Ordinal regression models were used to predict naïve listener evaluations of oral proficiency based on TOP decisions and other factors identified based on previous research, including speaker teaching skills (organization/clarity, respect/rapport), listener attitude homophily, and listener interest in the topic. Classification consistency for the models was examined using contingency tables (Xi, 2007). The results of this study can be used to evaluate and propose modifications to several claims in the TOP’s AUA, including (1) the claim that cut scores are set to minimize false positive decision errors, and (2) the claim that TOP score interpretations are sufficient to make decisions about a candidate’s functional oral proficiency. There are several implications for language testers. This study highlights a method for evaluating a comparatively under-researched aspect of validity (the sufficiency of test score interpretations for making decisions), and the results show how ‘real-world’ construct under-representation may undermine claims about the interpretation of scores and subsequently alter the overall argument for test use.

Parallel Session 7.3: Wednesday, June 4 16:10

Presenter: Maria Stathopoulou

Affiliation: National and Kapodistrian University of Athens

Title: Shuttling between languages in a testing context: defining interlinguistic mediation as translanguaging practice in the light of research results

In the new multilingual contexts of cultural diversity, people are very likely to assume the role of mediator, to translanguage or use two or more languages in a parallel fashion. Although the influential Common European Framework of Reference for Languages (CEFR) has considered mediation as a key communicative activity (along with production, reception and interaction), it does not provide any can-do statements relevant to mediation which would in turn facilitate its reliable assessment. This paper is the result of a longitudinal project aiming at exploring the complex mechanisms of interlingual mediation, a communicative activity which involves relaying in one language messages purposefully extracted from a source text in another language, so as to restore communication gaps between interlocutors (who do not share the same language). Drawing data from the Greek national standardized foreign language exams (known as KPG), the only examination system in Europe
which assesses test-takers’ mediation ability, the research investigated which strategies lead to successful mediation performance and ultimately explored what it means to mediate. Considering performance as being inextricably linked to the tasks which trigger it, it involved the linguistic analysis of (32) KPG written mediation tasks and (653) scripts produced by mediators of different proficiency levels over a period of 6 years. Tasks and scripts have been analysed by following different methodological procedures ranging from top-down to bottom-up. Specifically, while KPG tasks have been linguistically described on a basis of genre-oriented model with predefined categories adopting the Hallidayian view of language as a social semiotic, the analysis of mediation scripts has been based on an inductively developed framework of analysis, the so-called Inventory of Written Mediation Strategies (IWMS). Both quantitative and qualitative methods have been used in order to discover to what extent task parameters affect written mediation strategy use and which mediation strategies differentiate successful from less successful mediation scripts and scripts of differing proficiency levels. Overall, the outcomes of this research (a levelled mediation task typology and of the IWMS), may facilitate the development of mediation levelled tasks for testing (and teaching) purposes, while they may also contribute to the creation of standardized measures and clear benchmarks for reliable assessment of mediation competence, thus providing a critical supplement to the CEFR and to the curricula of language courses interested in developing learners’ interlinguistic competence. Based on the results derived from the aforementioned research, the present paper provides an empirically based definition of interlinguistic mediation and discusses this rather neglected issue in a critical perspective. The question it addresses is: what does mediation competence entail and on the basis of what criteria can it be assessed? The paper concludes by arguing that defining mediation on the basis of empirical evidence is a crucial step towards the construction of mediation-specific illustrative descriptors and thus towards the development of a Universal Framework which will consider mediation as an important aspect of communication.

Parallel Session 7.4: Wednesday, June 4, 16:10

Presenter: Hye Won Shin

Affiliation: Teachers College, Columbia University

Title: Measuring L2 vocabulary depth: A comparative analysis of word association and vocabulary knowledge scale tasks

This study reports on two independent L2 vocabulary test formats with the purpose of measuring the same target words in a given lesson. It is the outgrowth of research examining how different instructional techniques affect deep lexical knowledge in elementary school-aged EFL learners over time. The focus of this study is to investigate the two vocabulary assessments often employed to assess L2 vocabulary depth: the word association test (WAT) and the vocabulary knowledge scale test (VKST). WAT is generally thought to have reliability and validity (e.g., Greidanus & Nienhuis, 2001; Qian, 1999; Read, 1993; Schoonen & Verhallen, 2008) than VKST (e.g., Wesche & Paribakht, 1996). Interestingly, however, VKST is used more frequently by researchers, despite having receiving criticisms (e.g., Wolter, 2005). Modified versions of the VKST format have been developed, but no studies have assessed their reliability. Furthermore, while applied linguists encourage the use of multiple lexical measures to give a more complete assessment of the learning that has occurred (Milton, 2009; Webb, 2005), SLA researchers rarely employ more than one assessment. In fact, a synthesis of SLA studies on the effectiveness of receptive and productive-based vocabulary instruction shows VKST as the primary outcome measure used to determine effect sizes (Shin, manuscript under review). Indeed, Malvern, Richards, Chipere, and Duran (2004) call our attention to the pressing need for a more rigorous method of validating vocabulary measures. The goal of the study was to examine the validity and reliability of two independent tests, a WAT and a VKST, and their respective subtests (i.e., word knowledge and comprehension belong to WAT, and receptive knowledge and productive knowledge belong to VKST). The study included 153
adolescent Korean students, a group that has traditionally been under-represented in L2 vocabulary research. They are in their last year of Korea’s six-year elementary school system (the equivalent of the sixth grade, or the first year of middle school, in much of the US). The WAT and VKST tests both assessed the ability of the students to identify 21 target words that they were supposed to have learned in the classroom. The use of target words chosen from a school curriculum reflects conventional practices in an FL secondary classroom. A battery of statistical analysis was employed to determine the (a) internal consistency (Cronbach’s alpha), (b) reliability (test-retest), (c) reliability of subtests (correlations), (d) receptive-productive distinction (confirmatory factor analysis), (e) concurrent validity (Gates-MacGinitie), and (f) cutoff score accuracy of each test (receiver operating characteristic curve analysis). The results showed that WAT and VKST are equally effective measures of L2 vocabulary depth, but that WAT may potentially be a better test format. In short, WAT should play a more prominent role in evaluating L2 lexical knowledge. Implications for the test formats in FL classrooms are also discussed.

Parallel Session 8.1: Wednesday, June 4, 16:45

Presenter: Colin Finnerty and Anthony Green

Affiliation: Oxford University Press

Title: Balancing statistical evidence with expert judgement when aligning performance to a description of language ability.

Linking language test results to a common scale – such as the Common European Framework of Reference (CEFR) – allows test takers and users of results (such as receiving institutions) to interpret test outcomes in terms of meaningful descriptions of practical performance in the target language. In evaluating the value of a test as an indicator of a level of ability defined in CEFR terms, we need to investigate the quality of the alignment achieved as well as the fairness of the scoring decisions. The Council of Europe manual (Council of Europe 2009) provides a range of methods for aligning language tests to the CEFR, but is not prescriptive concerning which methods carry greater weight in determining cut scores. This paper outlines the activities that underpinned the alignment of the Oxford Test of English B (OTE-B) to the CEFR and explores the process of triangulating objective test data with subjective expert judgement. Expert judgement is an inexact science, but it allows us to make the essential connections between test content and framework descriptors. Statistics provide objective information on the relative difficulty of each test item, but require interpretation. The two rarely match perfectly, but must be reconciled in the interests of quality and fairness in score reporting. OTE-B is a general proficiency test targeting CEFR levels B1 and B2 that is taken entirely online. The test was written with the CEFR levels in mind and each item was intended to target a CEFR level. A series of benchmarking activities have been undertaken to ensure that test scores appropriately align to CEFR bands. Following the development of test specifications and the trialling of tasks, materials were commissioned from groups of trained item writers. Reading and Listening tasks were pretested on sample populations from a variety of L1 groups. In addition to classical analysis, Rasch analysis was used to develop an item difficulty scale. To assist with the initial development of the scale, pretests contained anchor items drawn from the Oxford Online Placement Test (OOPT), which had previously been aligned to the CEFR. Speaking and Writing materials were also trialled on sample populations and analytical criteria for raters were developed following iterative review processes. A benchmarking activity was then carried out to map materials to levels using human judgement. This activity involved 14 trained benchmarkers rating a sample of Reading and Listening tasks using the basket method to place items in CEFR bands. Analysis was carried out and this scale was compared to the Rasch-derived scale. The test was then piloted on a population of approximately 300 test takers from a range of L1 backgrounds. Teacher ratings of student performance were also collected for the pilot participants. The results from the above exercises were then combined to map Rasch estimates against expert ratings of materials and
teacher ratings of students. This presentation will explore the findings of these different alignment methods and the approach adopted to reconciling them. Reference: Council of Europe (2009). Relating language examinations to the CEFR: A manual. Strasbourg: Author.

Parallel Session 8.2: Wednesday, June 4, 16:45

Presenter: Rachel Brooks and Maria Brau

Affiliation: Federal Bureau of Investigation

Title: Diagnosing Proficiency and Performance Abilities

Before FBI linguists can work on improving their language abilities, they need to first gain a better understanding of the strengths and weaknesses in the language skills pertinent to their positions, particularly speaking and translation. Consequentially, the FBI’s Language Testing and Assessment Unit developed and implemented diagnostic assessments of language proficiency (speaking) and performance (translation) abilities according to the framework of the Interagency Language Roundtable Skill Level Descriptions. The diagnostic assessments give stakeholders a comprehensive understanding of their abilities rather than simple, numerical ratings. The diagnostic reports provide detailed written commentary to the examinee and list multiple examples from the assessment that support the evaluator’s conclusions. The resulting multi-page report gives linguists the directed information to focus their efforts on their areas of weakness leading to progress in their language skills. Five years into the diagnostic assessment program, over 150 assessments have been administered and the washback has been significant for both test takers and administrators. A qualitative and quantitative study of the first five years of diagnostic assessments was conducted to determine a) the effectiveness of receiving a diagnostic evaluation prior to retesting, b) a comprehensive profile of the ILR Level 2 ranges for the speaking and translation skills, where most of these assessments occur, and c) information on how to improve the effectiveness of the diagnostic assessment for future use. To examine the effectiveness of the diagnostics, a statistical comparison was made of the change in test/retest ratings between the group of examinees who took the diagnostic assessment prior to retesting and the group of examinees who did not receive explicit feedback prior to retesting. Additionally, the written diagnostic reports were coded and analyzed to determine the construct of the ILR Level 2 from a bottom up approach and contrast those concrete manifestations of the Skill Level Descriptions with the theory of the ILR framework. Lastly, the format of the feedback from the diagnostic was revisited in light of the previous research. The diagnostic report for the speaking evaluation evolved from being a written summary accompanied by written support for the claims made to a hybrid written/oral format, including audio excerpts from the diagnostic followed by rater commentary. Research from surveys and a comparison of change in pre-diagnostic and post-diagnostic speaking proficiency ratings determined the effectiveness of the new format. Preliminary results indicate that the targeted and comprehensive feedback in the first version of the diagnostics was somewhat effective in conveying a person’s strengths and weaknesses in speaking proficiency or translation performance; however, the report itself seemed overwhelming, which may have hindered its utility. Outlining the linguistic patterns that emerged in the diagnostic reports led to the development of a user-friendly diagnostic checklist for speaking proficiency and audio excerpts to support the evaluation. The participants in the diagnostic evolutions have been receptive to the new feedback and it is expected to make a significant improvement and examinees’ abilities to understand their own speaking proficiencies.
Parallel Session 8.3: Wednesday, June 4, 16:45

Presenter: Maria Tagarelli De Monte

Affiliation: Istituto Statale per Sordi di Roma


When dealing with language teaching and assessment, there is often low or no consideration of what is there in teaching and learning sign language (SL); a different modality, same communicative objectives, a higher need of competence (when considering the rehabilitating function of it). The lack of a shared framework of reference for SL and of shared methodologies in its teaching and assessment, leads to striking differences in the preparation and competences by sign language learners. This situation have drawbacks not only from a linguistic standpoint in the description and profiling of sign language, but also in the professional preparation of communication assistants and interpreters. This paper will present the attempt to build and design a complete educational methodology for sign language starting from the analysis and application of the outlines contained in the Common European Framework of reference for languages (CEFR) to at least 3 different sign systems: Italian sign language (LIS), Austrian sign language (ÖGS) and Catalan sign language (LSC). The outcomes of this work, started within the SignLEF project (signlef.aau.at - financed by the European Commission), includes the design of a preliminary Common European Framework of Reference for Sign Language (CEFR4SL) and of a set of materials (manual and DVD) to support sign language teachers in their job. However, the work done within this project by a team of professional researchers, deaf LIS teachers and interpreters has also opened the way to a number of issues and questions that still needs to be properly addressed, and that we will discuss and analyzed within this paper. Starting from the consideration of an international language framework, how should this address the description of non-verbal abilities required from a competent signer? Some of the fundamental grammatical and syntactic rules governing SL are proven to rely heavily on facial expressions and body movement. In the development of a framework for LIS, we’ve decided to add new descriptors for non-verbal skills, opening a totally new line of development from A1 to C2 level. From a linguistic standpoint, the definition of variant vs. dialect in the case of sign language is still a matter of discussion, as the definition of what is a C level of competence for a “native” signer vs. a learner. In a context where most of the signers are learners and only a very small percentage of deaf signers can be called “native”, the definition of a C level needs to be grounded in sociolinguistic analysis of sign language and its daily use. For this project, we have proceeded to the discussion and selection of a number of videos produced by the deaf community without any input from the hearing community, looking for videos showing sign language “as it is”, not adapted or simplified for the audience. Other issues such as the written form of sign language and the learning path of adult hearing learners will be addressed and discussed to show what is being done in the context of defining an international framework for sign language.

Parallel Session 8.4: Wednesday, June 4, 16:45

Presenter: Michelle Raquel

Affiliation: The Hong Kong Polytechnic University

Title: The Excel@English Scheme: Supporting language learning after diagnostic testing

Maintaining and developing Hong Kong tertiary students’ English proficiency throughout their studies continues to be a challenge. In Hong Kong universities, despite the fact that English is the medium of instruction, anecdotal evidence suggests that if English is not a formal subject of study, proficiency levels could
Abstracts June 4

decline or even hinder academic success. Tsui (2012) claims that this decline could be attributed to the fact that English is not the lingua franca in Hong Kong universities and this limits opportunities for English language use outside the classroom. To address this problem, the Hong Kong PolyU English Language Centre (ELC) launched the Excel@English Scheme (EES), a platform through which students take a diagnostic language test (i.e., the Diagnostic English Language Tracking Assessment (DELTA)) and then obtain support on areas of weakness as identified by their DELTA diagnostic report. Support to improve weaknesses are in the form of a mentoring programme, non-formal language learning activities, and a points collection system wherein students can keep track of their learning activities outside the classroom. At the end of the Scheme, students take the DELTA again to determine their English proficiency development gains. The initiative aims to provide first-year undergraduates and higher diploma students with a structure for self-directed English language learning by keeping track of their efforts and progress outside the classroom. This presentation discusses the principles, opportunities, and challenges involved in launching the EES and its role in promoting language development together with the DELTA diagnostic report. It will outline the conceptual framework of the Scheme and how elements of the Scheme (i.e., the DELTA diagnostic report, required learning plans, mentoring system, e-learning platform, and language learning activities) work together to promote language development. This is followed by a discussion on the issues involved in monitoring students’ progress particularly in tracking development of specific language skills. The presentation concludes with an evaluation of the Scheme as a means of validating diagnostic English language testing.

Plenary Session 2: Wednesday, June 4, 17:20

Presenter: Carolyn Turner and Beverly Baker

Affiliation: McGill University

Title: Process and performance in an online formative assessment tool for French-speaking nurses

In the unique context of Quebec, the only Canadian province where French is the official language, French-speaking nurses often need to be functional in English to provide services effectively. Quebec recognizes that essential services often need to be made available in English for its minority English population as well as for its increasingly diverse immigrant population. The present study was borne from the healthcare community’s concern that low L2 proficiency remains a major barrier to healthcare access for linguistic minorities and adversely affects health outcomes. To date this program of research has supported the development and validation of an online ESP self-assessment tool to support nurses’ language training. Earlier steps have been reported (XXXX, 2010, 2011) and include drawing on input from nurses concerning the traits/constructs underlying nurse-patient interactive tasks; calibrating those tasks onto a well-known, previously validated rating instrument commonly used for workplace purposes, the Canadian Language Benchmarks; and, finally developing a nursing-specific oral interaction tool for formative assessment purposes. The final result can be defined as an ESP assessment tool as described by Douglas (2000). This paper reports on the a priori validation phase of the assessment tool (a major phase to approximate situational authenticity) which leads to the development of the a posteriori validation phase using a small subsample. This is in line with an evidence-based approach to validation (Weir, 2005) and focuses on the growing interest in domain use construct definition and assessment usefulness (Bachman & Palmer, 2010). Our intent was to more accurately describe the construct through a priori validation, thus contributing to the usefulness of our next phase, a posteriori validation procedures which would answer our research questions: How do test-takers navigate in an online formative assessment environment? Which tasks work better in the context of formative assessment for ESP? How do the indices of linguistic and pragmatic items compare? When provided speech models, what is the nature of change in nurse speech across several attempts? The target population of the self-assessment online tool is pre-service and practicing nurses across the province of Quebec. The first section of the tool
Abstracts June 4

consists of six brief online scenarios which represent typical and engaging clinical nurse-patient interactions. Participants respond to several incidents within each one of these scenarios which tap into linguistic and pragmatic features (objective items). In the second section, they are requested to provide voice recordings of what they would say; then they listen to a selection of models, pursue further attempts and receive feedback. For the a priori validation phase of the assessment tool, a mixed methods approach was used: authentic scenarios were drafted and vetted by expert nurses; then objective prompts based on the scenarios were drafted; next, nurses responded to the items to generate language needed to create authentic distractors and keys; and finally, a rating scale was empirically derived through a content analysis (Tesch, 1990) from the nurse responses. The results of this phase will be discussed in light of the development of the a posteriori validation approach using a small subsample.
A critical aspect of professionalism in language testing is an awareness of historical antecedents in our field. This talk considers the past in language testing and discusses three important lessons we might learn from it. The first relates to Cicero’s dictum: “Not to know what happened before one was born is always to be a child.” Seminal work has been carried out in language testing in the ages before us and clearly we would benefit from an appreciation of it. The second concerns the necessity for clearly and explicitly defining a priori what we are trying to measure through our language tests, rather than attempting to establish this after the test event through sophisticated statistical procedures. The constructs language testers have sought to measure over time has changed in line with contemporaneous pedagogical priorities and a consideration of the diachronic relationship between pedagogy and testing provides valuable insights into this change. The third locates our field in a wider societal context and examines some of the social, economic and institutional forces that have influenced the directions language testing has taken in the last hundred years.

Cyril Weir is the Powdrill Professor in English Language Acquisition and the Director of the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire. The Centre offers scholars, examining boards and government organisations both domestic and foreign a centre in Britain that can provide and assist in quality research and development in these areas.

Cyril Weir has a PhD in language testing. He is the author of a wide range of educational publications, including Communicative Language Testing, Understanding and Developing Language Tests and Language Testing and Validation: An evidence-based approach. He is also the co-author of Examining Writing, Examining Reading, Evaluation in ELT, and Reading in a Second Language, Measured Constructs, and the co-editor of six Studies in Language Testing (SILT) volumes.

Cyril Weir is a member of the Editorial Board of Language Testing. He is co-editor for the Studies in Language Testing series published by Cambridge University Press.
Plenary Session 3: Thursday, June 5, 9:45

Presenter: John H.A.L. de Jong

Affiliation: Pearson / VU University Amsterdam

Title: Standards & Scaling

Governments, corporations, educational institutions, parents and indeed the learners themselves, all are demanding to understand what we are actually teaching in language classes and what the results of assessments mean in real life; what the relation is between those results and how well bearers of the diplomas and certificates can be expected to deal with the demands of everyday language use.

So, though professional language testers talk a lot about validity, the relative importance of validity and reliability, the different types of validity and how extremely complicated these matters are, in fact the lay public is asking us: please provide information on the validity and reliability of your stuff, of the things you claim to be doing for us.

In this contribution I will discuss the major approaches to designing descriptive scales from a measurement point of view and review some forms of reporting test results and how they relate to validity and reliability.
Abstracts June 5

Work in Progress 1: Thursday, June 5, 10:20

Presenter: Troy Cox

Affiliation: Brigham Young University

Title: A new look at selecting a language for assessments in L2 reading comprehension

A new look at choosing a language for L2 reading assessment When measuring foreign language reading, test developers need to decide whether the language of questions should be in the first (L1) or second (L2) language. Heretofore, there has been a limited amount of research to guide the making of such decisions, particularly among more advanced learners. The available research indicates that, with the exception of fluent and very highly proficient L2 readers, questions in the L2 result in lower test scores, and that the effect is greater with lower proficiency readers (Godev, 2002; Gordon, 1995; Lee, 1986; Shohamy, 1984). With those studies, however, the reasons for the lower scores have not been deeply explored, nor have the affective characteristics of participants including confidence, anxiety and attitude been measured. The forthcoming study will explore these issues among advanced adult L2 learners by investigating the effect of language of assessment on test performance and the relationship of the language of assessment with affective characteristics. A computer-administered reading comprehension test with 20 short passages in Russian will be given to over 60 advanced learners of Russian. For each passage, a single multiple-choice question will be administered in both English and Russian and examinees will see the same question for each passage in both languages one after the other. The language in which the participants first see the questions will alternate such that Group A will see questions 1-10 in English first and 11-20 in Russian first. Incorporating a counter-balanced design to account for language order, Group B will see questions 1-10 in English first and 11-20 in Russian first. When examinees change their answers, they will be asked to explain why they changed their answer upon seeing the question in the other language. After each question, learner affect will be assessed with survey items that measure confidence and anxiety. A post-assessment survey will also be administered to measure overall attitudes toward the language of assessment in terms of difficulty, anxiety, preference, and understanding the question. The results and analysis of the reading comprehension exams will add to the small body of similar research, and the findings from the survey items will better inform educators and test-developers of the implications of choosing a language for assessment.

Work in Progress 2: Thursday, June 5, 10:20

Presenter: Meg Malone

Affiliation: Center for Applied Linguistics

Title: United States tertiary education institution administrators’ perceptions of validity of the Cambridge English: Advanced exam

The Cambridge English: Advanced assessment (CAE) is widely used as a certification of advanced proficiency (CEFR C1) in English. Among the 3,000 organizations that accept the CAE are over 100 institutions of tertiary education in the United States. These institutions make admissions decisions and additional language support decisions for international students based partially on test scores, making the CAE a high-stakes exam. Evidence-based test validation (Weir, 2005) suggests that the consequential validity of an exam be evaluated. Kane (2006) emphasizes the importance of exploring the consequences of test use, including whether the test is used for the purposes intended by the developer. Given the CAE’s potential for use as a gate-keeping mechanism in the admissions processes of tertiary education institutions, support for the validity of the CAE should include evidence that its use in such contexts is in line with the purposes intended by the test developers and fair to all stakeholders. This would include evidence that the CAE is being used in an appropriate way by institutions to provide evidence of students’ advanced English proficiency during the
Abstracts June 5

admissions process and to make decisions about the provision of any additional English language services after admission. Investigation of stakeholder perceptions of an exam is one way of gathering such evidence. Results illuminate stakeholders’ understanding and perceptions of the test, which are likely to impact its use (Chapelle, Enright, & Jamieson, 2008; Hawkey, 2006). While considerable research has been conducted on test taker and instructor perceptions, administrator perceptions are less studied, despite the fact that administrators often play a gate-keeping role by interpreting test results. Even less studied are the perceptions of administrators towards the CAE in the context of the United States. The current study aims to fill this gap by conducting preliminary focus groups and a large-scale online survey to examine the knowledge and perceptions of U.S. administrators in tertiary institutions regarding the CAE. The outcomes of this study will shed light on the ways in which the CAE is being used by this group, which should be considered when evaluating evidence of the validity of the instrument. This information will further the language testing field’s understanding of the extent to which the meanings of test results are understood beyond the audience of test developers. This Work in Progress session will present preliminary results of the focus groups and the surveys to seek input from the language testing field regarding interpretations of findings and directions for analysis. References: Chapelle, C., Enright, M., & Jamieson, J. (Eds.) (2008). Building a validity argument for the Test of English as a Foreign Language. New York: Routledge. Hawkey, Roger (2006). Impact Theory and Practice: Studies of the IELTS test and Progetto Lingue 2000. Cambridge, England: University Press. Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education. Weir, C. (2005). Language testing and validation: An evidence-based approach. Hampshire, UK: Palgrave Macmillan.

Work in Progress 3: Thursday, June 5, 10:20

Presenter: Simon Dawson

Affiliation: Nottingham University, Ningbo, China

Title: Providing EAP test candidates with fuller feedback

The University of Nottingham, Ningbo, China runs a preliminary year EAP program averaging 1,500 students annually. A team of assessment specialists work throughout the year on the development and administration of EAP assessment for the various courses run within the program. A recurring issue faced by the team is that of feedback from assessment. While a great deal of work goes into each assessment (from the part of the assessment team, tutors, students and assessors) candidates receive just a single score or set of scores following an assessment. More feedback could be of considerable use to stakeholders in identifying student strengths and weaknesses that can inform future teaching and learning. However, giving fuller feedback to test candidates is viewed with reservation as it is generally assumed it will involve a great deal more work on everyone’s part. This study investigates whether there is a way of providing fuller feedback to candidates without significantly adding to the work of assessors, test developers, or administrators. The study uses the assessment of Academic Oral Presentations (AOP) as a focus, but the outcomes should be applicable to other forms of performance assessment. The researcher intends firstly to use focus groups to investigate how assessment feedback is interpreted and used by students and tutors. This will help to identify areas where fuller feedback could be of use. Next, the researcher intends to trial an alternative method of assessment of AOPs which involves the assessor making notes while watching presentations instead of only allocating scores. Candidates would receive the written up notes with their set of scores. The researcher plans to review the potential of this alternative method through focus groups with different stakeholder groups. The paper will be of interest to those who are involved in the development of language assessment, especially performance assessment, and to those who make use of the outcomes of language assessment.
**Work in Progress 4:** Thursday, June 5, 10:20

**Presenter:** Rie Koizumi and Yo In’nami

**Affiliation:** Juntendo University

**Title:** Effectiveness of role play versus discussion: Comparing paired oral tasks in classroom assessment

Recent oral assessment research shows that paired tasks based on interaction between learners have benefits as an assessment method in comparison to interviews between interviewers and learners, in that paired tasks elicit a wider range of speech functions and co-constructed discourses (e.g., Galaczi & ffrench, 2011; Taylor & Wigglesworth, 2009). In addition, paired orals would be able to relate teaching to testing more firmly when similar tasks are used for instruction and assessment; thus, this approach can probably produce more positive washback on teaching. Previous studies have examined the effects of various factors related to paired orals on peer–peer performance, rating processes, and resultant test scores (e.g., Ducasse, 2010; May, 2011). Most such studies used unguided discussion tasks, but some (Csépes, 2009; Kormos, 1999; Teng, 2013) used guided role play tasks. It is expected that discussion tasks demand more interactive ability and thus are appropriate for more proficient learners than role plays; role plays would be useful for less proficient learners if conversation structures are provided and would be able to cater to learners at various proficiency levels by using either simple or complicated roles and situations presented on role play cards. However, to the present authors’ knowledge, no studies have further compared role play and discussion tasks in terms of respective advantages and disadvantages and common features. This study aims to investigate these areas by considering test discourses and scores and students’ and teachers’ reactions toward the two tasks in the context of classroom assessment of Japanese learners of English. This research would clarify the features of paired orals, facilitate their use in speaking tests, and help teachers select effective tasks for in-class student evaluation. This project will first develop role play cards and discussion topics, ranging from easy to relatively difficult and then execute a pilot study to check their appropriateness and task difficulty. Rating scales and feedback sheets will be based on Ducasse (2010), May (2011), and Taylor (2011). The participants will be Japanese university students who have studied English as a foreign language for at least six years and are novice or intermediate level speakers. They will engage in pair-work during class, perform both role play and discussion tasks in their midterm and final exams, conduct self-assessment, and reflect on their performance using teacher-assigned scores. Their performance will be video-recorded, evaluated by two raters using the rating scales, and transcribed for further analysis. We will employ multifaceted Rasch analysis to examine test-taker ability, task difficulty and consistency, rater severity, and bias patterns, and conduct an analysis using generalizability theory to clarify the number of task administrations necessary to obtain consistent scores. Finally, we will conduct student and teacher questionnaire surveys to examine their feelings about the two paired oral tasks and the score reports and the manner in which they use them for learning and teaching. In the presentation, we will present prototype tasks and rating scales alongside the pilot study’s results.

**Work in Progress 5:** Thursday, June 5, 10:20

**Presenter:** Ute Knoch, John Pill and Lyn May

**Affiliation:** University of Melbourne

**Title:** Transitioning from university to the workplace: Stakeholder perceptions of academic and professional writing demands

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
Abstracts June 5

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. The transition point from university to the workforce remains under-researched; nevertheless, some employers have suggested that graduates are entering the workforce with insufficient language skills. This study aims to gain a detailed insight into the changing writing demands from the last year of university study to the first year in the workforce for two groups of professionals, accountants and engineers, and relate these to the demands of the writing component of IELTS, which is increasingly used for exit testing, although not expressly designed for this purpose. The study will collect data in several phases, using a multi-method design. Interviews and questionnaire data will be collected from final year students as well as their lecturers. These interviews will focus 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. Students will also be asked to supply samples of their written work. Lecturers of final year core courses will also be asked about common problems in students’ writing and about typical writing tasks in university courses. Similarly, interviews will be conducted with new graduates and employers of new graduates. In this phase, we will likewise attempt to collect written samples and tasks required from new graduates in employment. For the final stage of this study, we will hold a series of focus group discussions with employers of new graduates and lecturers of final year students to investigate the linguistic features graduates are perceived to need in order to participate successfully in the workplace and whether these can be identified in the various task types and work samples collected in the earlier stages of the study (including IELTS tasks, university tasks and workplace tasks and writing samples). The study will provide important insights into the similarities and differences of writing demands between IELTS writing tasks and those of accountancy and engineering university courses and workplaces. It will provide important information about the perceptions of students, graduates, lecturers and employers on the development of English language proficiency through university courses and in the workplace as well as the suitability of the use of academic proficiency tests (such as IELTS) as university exit tests and for entry into the professions.

Work in Progress 6: Thursday, June 5, 10:20

Presenter: Troy L. Cox

Affiliation: Brigham Young University

Title: Improving the reliability of ASR rating in elicited imitation assessments

Automatic speech recognition (ASR) is being explored by many as a way to reduce costs in rating speaking exams. The accuracy of ASR in processing speech, however, is highly dependent on the quality of the recorded speech. Graham et. al (2011) had a correlation of .91 (n=103) between human-rated elicited imitation (EI) and oral proficiency interviews, yet when the EI files were processed with ASR, the correlation dropped to .68. While some testing companies attempt to improve ASR accuracy by automatically sorting out anomalous sound files (Higgins et al, 2011) for human inspection, little has been reported on open-source technical procedures of automatically processing the files for improved accuracy. To investigate the effects of different audio post-processing, a two-step approach was adopted. First, a comparison between an ASR-scored EI test and a human-rated speaking proficiency test (SPT) was conducted. As part of an institutional test battery, 295 students took a 45-item EI test and a 10-item SPT comprised of a hierarchy of level specific prompts. The SPT was double-rated by trained human raters and the results were analysed with a many-facet Rasch measurement to obtain a speaking rating. The EI sentences were scored via the CMU Sphinx-4 speech recognition system using a Kleene Star Grammar, and the correlation between the SPT scores and the EI ASR score was .75 with a number of EI
files being misrated due to distortions (e.g. background noise, breathing on microphones, fluctuations of intensity, etc). In the second step, files were human-rated to obtain a standard with which to compare the ASR findings. The files were then post-processed with a variety of filters (e.g. frequency, intensity, etc) to determine which combination most closely aligned with the standard. Correlations between the ASR-filtered EI, human-rated EI, and PST will be reported.

Work in Progress 7: Thursday, June 5, 10:20

Presenter: Simona Sunara and Jeffrey Steele

Affiliation: University of Toronto

Title: Challenges in the development of a web-based French global proficiency test

This presentation discusses the on-going creation of a web-based French proficiency test designed to evaluate Canadian University-level second language learners’ vocabulary, grammar, reading, writing, speaking, and listening abilities. The test’s development is guided by the following general objectives: (1) program administration: place students in the appropriate French-as-a-second-language (FSL) course within a five-level language program; (2) learning & pedagogy assessment: provide qualitative feedback on specific linguistic features and skills that require particular learning/instructional attention, and quantitative evaluation of the extent to which curricular objectives are being met; (3) acquisition research: obtain empirical data in order to establish developmental learner profiles (e.g., Ågren, Granfeldt & Schlyter, 2012) that will contribute to our understanding of patterns of second language development including the extent to which linguistic knowledge and skills in different areas of competence develop in tandem. Such profiles will also serve to evaluate the content and current sequencing of the learning objectives of our FSL series. Along with these main goals, the test (4) must be self-correcting due to the level of assessment training of individuals involved in our French program, and (5) exist in multiple versions of equivalent difficulty, given that some learners take the test multiple times over the course of their studies. We discuss challenges and solutions in meeting these objectives with reference to the development and pilot testing of the vocabulary and grammar modules. In order to test vocabulary breadth and depth, French versions of the Vocabulary Levels Test (Nation, 1990) and the Word Associates Test (Read 1993, 1998) were developed. Grammatical competence, operationalized as learners’ passive and productive use of a variety of (morpho)syntactic phenomena (e.g., noun phrase features of determiners including grammatical gender, number, and definiteness; verb phrase features including tense, mood, aspect, person and number; sentential features including syntactic order and complexity) were assessed with multiple-choice, and sentence-completion and scrambling tasks. Challenges include (i) creating a test for a highly heterogeneous learner population in terms of both first language (50% of students at our English-language institution have a mother tongue other than English) and French proficiency (learners range from absolute beginners to near-native speakers); (ii) establishing proficiency levels given the absence of a recognized standardized French test; (iii) having a test that is sufficiently long to provide a valid and reliable assessment of learners’ proficiency in the six areas yet not so long as to be discouraging and/or fatiguing; and (iv) determining which variety/varieties of French should be targeted, given that test takers will have been exposed to both Canadian and European varieties in their previous learning experience. All of these issues, including solutions adopted, will be discussed with reference to the test’s development and evaluation including qualitative data from post-test interviews and quantitative results from piloting.
Abstracts June 5

Work in Progress 8: Thursday, June 5, 10:20

Presenter: Yuichiro Yokouchi

Affiliation: Graduate School, University of Tsukuba

Title: Investigating Effective Information Presentation in Retelling Tasks and Its Characteristics of Utterances

Retelling tasks are widely used in language testing situation and classroom activities. Hirai and Koizumi (2009) developed a practical speaking test framework, which is called as Story Retelling Speaking Test (SRST), and Koizumi and Hirai (2012) reported that SRST can discriminate speaking performance between beginner and intermediate level. Accordingly, SRST or retelling tasks is effective to assess EFL learners speaking performance; however, there is no guideline for presentation of information. The length, difficulty, and input mode (e.g., listening input or reading input) should be considered when teachers apply retelling tasks to their classroom activities or tests. To use retelling tasks effectively, investigating appropriate text length and difficulty of input in retelling tasks is necessary. The effect of presentation mode in retelling tasks also should be considered. Consequently, the objective of this study is investigating the effect of input length, input difficulty, and presentation mode in retelling tasks. To see the effects of those factors, utterances will be analyzed from many directions. The features of utterances will be explained by the indices of complexity, accuracy, and fluency (hereafter, CAF). CAF are used as variables to assess language learners’ performances of production. Some researchers use CAF indices to see the learning effects (e.g., Gilabert, Baron, & Levkina, 2011; Kuiken & Vedder, 2011), and those indices may be able to apply to assessing speaking abilities. The term of CAF had started use in Housen and Kuiken (2009); in actually, some indices such as type-token-ratio as for the lexical complexity, numbers of errors per AS-unit, or words per minute as for the fluency are traditionally used for performance measurement. However, there are disagreements in the use of CAF indices (Koizumi, 2005; Sakuragi, 2011). Therefore, considering the choice of CAF indices may be an important point for this study. Part of experiment has been executed. Eighty-nine Japanese college students participated to this research. Five data were excluded because of the failure of recording, and finally 84 data were used to analysis. Participants took two reading-based story retelling tests and two listening-based story retelling test. Two graduate students who are majoring English education transcribed all utterances. The transcribed data are pruned to measure the performance by the perspectives of CAF. Furthermore, all performances were scored by the empirically-derived, binary-choice, boundary-definition (EBB) scales which is published by Koizumi and Hirai (2012). The variables of CAF and EBB scores are examined quantitatively, and the transcribed utterances examined qualitatively.

Work in Progress 9: Thursday, June 5, 10:20

Presenter: Claudia Harsch

Affiliation: University of Warwick

Title: Could a “universal framework” improve TOEFL iBT® score interpretation for students and their tutors?

With the increased intake of international students in UK universities, English language tests have gained importance in their ‘gate-keeping’ role. One of the tests used in the UK is the TOEFL iBT®, the predictive validity of which is the focus of our project (funded by the ETS TOEFL ® COE Research Programme; RFP 2012-21; duration 2013 – 2016). One known factor which influences academic success is linguistic preparedness. It is hence imperative for universities to be provided with diagnostic information about incoming students’ linguistic preparedness. Since this angle is often neglected in predictive-validity studies, our study enriches the traditional approaches by addressing the issue of how students and their tutors interpret and use the test
Abstracts June 5

reports. We set out to answer the following questions: What use do test users make of the descriptions provided by the TOEFL iBT report, along with relevant CEFR-descriptors? Do test users perceive the test reports as helpful in identifying strengths and weaknesses so that incoming students can be adequately supported by EAP tutors? Do users perceive a need for more specific or more universal descriptors in test reports? The outcomes will inform decisions about language support offered to incoming students based on their TOEFL iBT score profiles, and can inform possible revisions (by ETS) to the reporting system to meet users’ perceived needs. Our study targets postgraduate students entering with a TOEFL iBT score in one major UK university, as well as lecturers from the admitting departments and all EAP tutors. We will employ an interdisciplinary mixed-methods approach in order to enrich traditionally quantitatively-oriented predictive-validity studies with qualitative data from questionnaires and interviews. The work-in-progress presented here uses data from these questionnaires and interviews. The questionnaires are currently being administered to 227 postgraduate students entering the university in October 2013, to all EAP tutors of the university, and to all academic tutors of the relevant departments. Follow-up interviews will take place in early 2014 with volunteering students and nominated tutors. The questionnaires and interviews cover the following concepts: 1) Students: perceived usefulness of test reports for interpreting their strengths and weaknesses; exploitation of test results with a view to seeking language support; perceived usefulness of self-assessment using CEFR-descriptors versus descriptors used in test reports; their expectations of what a ‘good’ test report should convey. 2) Tutors: perceptions of interpretability and usefulness of test reports; reported usage of test reports for EAP support referral, placement and teaching; expectations of a ‘helpful’ test report in terms of specificity. We will present first results of the perceived usefulness and diagnostic information provided by TOEFL iBT test reports for students and tutors. Our findings should yield insights into the perceived utility of different sets of descriptors (TOEFL iBT test descriptors versus CEFR descriptors), and into the perceived preferences for either more (discipline-) specific or more universal descriptors, thus contributing to enhancing diagnostic decisions about language support offered to incoming international students.

Work in Progress 10: Thursday, June 5, 10:20

Presenter: Yehbeen Yang

Affiliation: Ontario Institute of Studies in Education University of Toronto

Title: Assessing Multilingual Competence Conditional upon Interaction between Linguistic Knowledge and Cognitive Processing Skills

In our globalized society, multilingualism has become a highly desired asset. In various cultures learning a second or third language has become a common practice at the start of a formal education. In the earlier research, bilingualism was seen as a handicap for young children (Craik, Bialystok, Gollan & Green, 2009). It was thought that bilinguals’ dominant language would impede the development of their learning language. However, current research shows contrary results. It has been found that multilinguals have greater cognitive processing skills by the demand of juggle between multiple language systems (Craik et al, 2009). Such phenomenon – having two or more language systems in one mind –has been termed as multicompetence (Cook, 1992). In order to assess multicompetence, two parts need to be considered: linguistic knowledge and cognitive processing skills. However, linguistic proficiency of bilingual minds has been judged by the summation of their L1 and L2 linguistic knowledge (Jessner, 2008). Such method of assessment disregards the enhanced cognitive processing skills that multilingual minds have compared to monolinguals. Multilingual minds process their multiple languages holistically – their L1 and L2 share the same mental lexicon and processes in roughly the same areas of the brain (Carlson & Meltzoff, 2008). Hence, language systems are not in isolation but are processed in a way that the skills in one language are transferred in assisting the development of the second language (Bialystok, 2007). The practice of assessing multilinguals’ linguistic competence against monolingual
native-speaker standard cannot capture their multicompetence. Linguistic knowledge is nevertheless important since it is the resource being consciously processed in gathering information that is necessary in producing what is communicative. This is the works of cognitive processing skills. The present study examines the interactional relationship between linguistic knowledge and cognitive skills that operate in multilinguals’ minds. It builds on the threshold hypothesis that a certain level of linguistic knowledge is required in both L1 and L2 in order to benefit from higher cognitive abilities without the deterioration in either language (Cummins, 1979), and seeks to answer the following question: How does the knowledge and cognitive processing skills in both L1 and L2 coordinate themselves in developing multicompetence? In an attempt to answer the question, I would like to measure 100 bilingual English-French grade 2 (age 7-8) students enrolled in French immersion program in Toronto District School Board (TDSB) of their French and English knowledge by obtaining their provincial English and French reading and writing scores. Participants’ verbal fluency will be measured by the Delis-Kaplan Executive Function Battery (DKEFS; Delis, Kaplan & Kramer, 2001), which assesses both semantic and phonological fluency. Participants are free to use both languages in the tasks. It is hypothesized that participants would perform just as well as comparable monolinguals when they are able to use both of their languages, instead of having to inhibit one language system. Hence, the benefit from the use of their entire language system represents their multicompetence.

Work in Progress 11: Thursday, June 5, 10:20

Presenter: Payman Vafaee

Affiliation: University of Maryland

Title: Persian Linguistic Correlates of Proficiency

Recently there has been increasing demand for advanced learners of less commonly taught languages (LCTLs) in the United States. To determine which learners are qualified to fill positions requiring ability in a LCTL, the U.S. government uses proficiency tests which are based on the Interagency Language Roundtable (ILR) proficiency scale and considers ILR 3 (the ILR scale is 0–5) as the minimum acceptable level of proficiency (Brecht & Rivers, 2005). However, the majority of LCTL learners do not go beyond ILR 2, even after many years of instruction and learning (Long, Gor & Jackson, 2012). Although proficiency scales like the ILR are attractive to end users and test administrators, from a language-learning or language-teaching perspective, these scales and ratings based on them are less helpful. Numerical ratings, like ILR 1+ or 2, may mean something to testing specialists, but they may mean little to learners or their teachers. So far, few data exist on the appropriate linguistic content for the different ILR levels, or on psycholinguistically defensible learning sequences within the levels—especially at the advanced levels for LCTLs. It would therefore be valuable for teachers, textbook writers, testers, and the learners themselves to know precisely what is learnable, and therefore teachable, at a given stage in proficiency development (Long et al., 2012). In the current paper we report on preliminary results of the Persian Linguistic Correlates of Proficiency (PLCP) project. The goal of the PLCP project is to provide empirically based checklists of Persian linguistic correlates of ILR proficiency, and thereby to increase the transparency and utility of ILR ratings and, ultimately, to improve the teaching, learning, and testing of Persian. The specific purpose of the current study is to identify linguistic features of Persian that are implicated in progress on the ILR scale—specifically, from ILR 2 to 2+, and from ILR 2+ to 3. To achieve this goal we developed a battery of ten different receptive-based test tasks (e.g., lexical decision, grammaticality judgment, naturalness, collocation, and vocabulary-translation) to measure test takers’ control of Persian phonology, morphology, syntax, lexis, and collocations. The language structures and features that were the focus of the tasks were chosen based on interviews with experienced teachers and advanced learners of Persian (e.g., Persian subjunctive, sub-verb agreement, light verbs, collocations, vowels and liquids pronunciation). 60 advanced learners of Persian (who have completed at least two years of language training) and 30 Persian
native speakers (NSs) are being recruited. First, learners will take the official Persian Oral Proficiency Interview (OPI), which reports scores based on the ILR scale, and then both NSs and learners will complete the PLCP battery. Subsequently, NSs' responses will be used as a baseline and to detect native-like response patterns, and the learners' scores on PLCP tasks and the OPI will be statistically compared, and the relationship between them will be investigated. The results of this project will enable researchers to identify the psycholinguistically measurable Persian linguistic components that are correlated with scores of ILR 2, 2+ and 3.

Work in Progress 12: Thursday, June 5, 10:20

Presenter: William Pellowe and J. Lake

Affiliation: Kinki University Fukuoka

Title: Many-faceted Rasch analysis of peer-assessment as a diagnostic tool

Language proficiency frameworks can guide curriculum planners and classroom teachers, exemplified by Hadley’s (2001) accessible introduction to the ACTFL framework as the basis for instructional planning. The use of rubrics in instruction requires learners to both understand the rubric and to be able to assess the strengths and shortcomings of their own work relative to the rubric though, presupposing that learners have sufficient metalinguistic knowledge to understand the rubric, an assumption that Tokunaga (2010) challenges. Previous research has supported the formative use of peer assessment but found peer assessors to be inconsistent in their interpretation of rating rubrics (Cheng & Warren, 2005; Farrokhi, Esfandiari, & Schaefer, 2012; Mok, 2011; Saito, 2008). However Toppings (1998) notion of "learning by assessing" holds that interaction with the rubric during peer assessment can drive learning suggesting that instructional explanations based on proficiency frameworks may be less effective than using peer assessment as a mechanism to improve students understanding of the rubric. Additionally, fit analysis of peer assessors’ interpretation of the rubric can guide remedial instruction by identifying rubric items that students struggle to interpret. Peer assessment was piloted in academic writing classes at a Japanese women’s university (n = 24). Students assessed each others’ essays using a 9-item rubric and entered their ratings into an on-line database. Many-faceted Rasch analysis found general agreement in rank ordering between students' ratings and teachers' ratings, but that students tended to rate holistically, did not use the full range of the rating scale, and were much more lenient than teachers. The rating patterns provide evidence that students were unable to interpret the rubric clearly rather than simply interpreting it differently than teachers, meaning that as well as being unable to provide diagnostic feedback to each other, students were unlikely to understand feedback from teachers. By anchoring the difficulty of the rubric items against teacher ratings using the Facets software package, the most misfitting items when rated by peer assessors could be identified, these being "Introduction", "Thesis statement", and "Conclusion". In their second essay, all rubric items improved substantively, but "Thesis statement" was the only item showing substantively and statistically significant improvement greater than overall improvement. These pilot results indicated points of weakness in the instructional materials and also suggested that the observed gains in proficiency were more likely due to practice and learning by assessing than to instruction and feedback. Revised instructional materials were produced to address the problematic rubric items and are being operationally piloted during the second semester of 2013 (n = 105). Results due in February 2014 will confirm or disconfirm whether the revised instruction improved student understanding of the rubric and how this affected gains between the first and second essays produced by students.
Abstracts June 5

Work in Progress 13: Thursday, June 5, 10:20

Presenter: Heidi Han-Ting Liu and Bill Bonk

Affiliation: Pearson Knowledge Technologies

Title: Assessing Pragmatics: Developing a Two-Turn Integrated Spoken English Item Type for a Computerized Workplace English Test

The testing of L2 pragmatic knowledge has received increasing attention in recent years. Various types of instruments have been developed to measure L2 learners’ pragmatic knowledge, such as DCTs, role plays, and self-assessments (e.g., Hudson, Detmer, & Brown, 1992, 1995). Role plays are arguably one of the common elicitation methods adopted for assessment purposes, particularly in oral proficiency testing. However, research shows that many of the current approaches to testing L2 pragmatic knowledge share the challenges of construct underrepresentation and lack of interactiveness (e.g., Grabowski, 2008; Roever, 2011). In addition, role plays are currently typically scored by human raters. The present study has two goals. The first goal is to investigate the feasibility of developing an item type to more holistically represent the construct of L2 pragmatic knowledge by taking into consideration the roles of language users, contexts, and interaction, and by incorporating the concepts of implicature, routines, speech acts, and conversation analysis. The other goal is to construct two-turn interactions in such a way that items elicit performances analyzable automatically. The new items are specifically designed to elicit L2 learners’ pragmatic knowledge regarding workplace appropriateness, as the item type is to be implemented in a workplace English test. In the item design, each item contains two parts: the situation, and two pairs of turn-taking between the interlocutor (a recorded voice) and test-takers. The interlocutor turns are designed to fit the situations; the style, politeness, and register are written to be appropriate for the relative power, distance, and imposition between the interlocutor and test-takers as described in the situations. In the first turn-taking, test-takers either respond to the interlocutor who initiated the first pair-part with an appropriate second pair-part, or initiate a first pair-part that’s appropriate for the situation. In order to elicit further responses from test-takers in the second pair of turn-taking, the interlocutor is designed to say a dispreferred response, which requires test-takers to provide further accounts to maintain the appropriateness of the conversation. Whether test-takers can successfully respond appropriately in the second pair of turn-taking depends on their ability to decode the situation-specific implied meanings. In the first pilot testing, item format variations were investigated. Based on the results, a total of 30 items were developed and implemented in a computer-delivered test system. The second pilot testing involved 42 test-takers, with a mixture of native English speakers, ESL learners, and EFL learners. The results showed that, in general, the items elicited the expected responses. The test-takers consensually agreed that the second pair of turn-taking in all situations were adequate. Test-takers with lower proficiency level (i.e., CEFR level below A2) were found to have difficulties in comprehending both the descriptions of the situations and the implied meanings of the interlocutor utterances. The research study is currently developing scoring rubrics. Potential criteria considered to be included in the rubrics are content (i.e., amount of information and topic relevance) and appropriateness (i.e., prosody and structure). However, additional elements for scoring are still under investigation.

Work in Progress 14: Thursday, June 5, 10:20

Presenter: Toshihiko Shiotsu

Affiliation: Kurume University

Title: Evaluating the Effects of False Alarm Rates and Non-words on a Yes/No Vocabulary Test
Since being introduced to second language testing by Meara and Buxton (1987), the Yes/No vocabulary test format has received a considerable amount of attention as an efficient means of estimating the vocabulary size of learners for purposes such as placement and diagnosis. This computer-based format is essentially a self-assessment task in which test-takers indicate whether or not they know the meaning of a series of words presented on screen, with a certain percentage of non-words included as a control for guessing. We have conducted a project involving 345 EFL learners at universities in western Japan where we explored the effects of three extensions of the basic Y/N format: oral vs. written presentation of the target words; the addition of sentence contexts to the presentation of the words; and the measurement of reaction times as well as response accuracy. The main findings of the project have been reported elsewhere, but recent studies by Pellicer-Sánchez & Schmitt (2012) and by Stubbe (2012) have prompted us to undertake a re-analysis of our data, with a focus on two aspects in particular. The first is the effect of the false alarm (FA) rate, where test-takers claim knowledge of non-words. Other studies with Japanese learners, including Stubbe’s, have found low FA rates overall, indicating that learners in Japan tend to be conservative in their claims of word knowledge. We have found FA rates that extend much higher than that. Thus, while looking at variables in our design that might have contributed to increased false alarms, we have been re-analysing our data by excluding test-takers with FA rates of 20% or more. With the remaining sample of 246, we have found among other things that our Yes/No test results correlate more strongly with an external index of the participants’ knowledge of the tested words. The second aspect of the re-analysis is related to the first, but focuses on the difficulties of the non-words. The role of the non-words in a Y/N test is a matter of ongoing debate. Like earlier researchers (Mochida & Harrington, 2006; Shillaw, 1996), both Pellicer-Sánchez & Schmitt (2012) and Stubbe (2012) argue on various grounds that it may not actually be necessary to include non-words in the test in order to obtain a valid estimate of the learners’ vocabulary size. We are searching our data for evidence that might further elucidate the issue of whether the non-words make a worthwhile contribution to the measurement of knowledge of the target words or whether they constitute in effect a form of construct-irrelevant variance.

Work in Progress 15: Thursday, June 5, 10:20

Presenter: Yumiko Moore

Affiliation: University of Bedfordshire

Title: Investigating valid constructs for writing tasks in EAP tests for use in Japanese university entrance examinations

There has been an increasing debate about English tests used in Japanese university entrance examinations. One of biggest criticisms of the current English tests is the appropriateness of tasks in the tests. Politicians and business leaders in Japan is advocating that the Test of English as a Foreign Language (TOEFL) should replace current English tests, whilst educators feel that TOEFL (which is primarily intended for North American university admissions) is inappropriate for Japanese university entrance purposes and have proposed a new English test, a localised Test of English for Academic Purposes (TEAP), specifically intended for Japanese University admissions. Despite the extent of the argument over the tests, little research has theoretically examined the validity of the tests. Drawing on the socio-cognitive framework for validating writing tests (Shaw and Weir, 2007), this study aims to establish writing tasks in English for Academic Purposes (EAP) tests which are valid for Japanese university entrance purposes. The findings will contribute to further developing EAP tests for use in university admissions in Japan. Data will be gathered from multiple sources in Japanese universities at various points in time during the 2013, 2014 and 2015 academic years. The mixed methods approach has been chosen for this study because a combination of qualitative and quantitative approaches provides a better understanding of research matters than either approach alone; this approach will enable the study to reveal complementary strengths and non-overlapping weaknesses (Johnson and Onwuegbuzie, 2004). First, the initial sample for a survey will consist of 200 to 300 Japanese teachers from a number of prestigious Japanese universities, with a follow-up semi-structured interview of five or more teachers chosen from among the survey
participants. Second, the investigation will include an analysis of (a) course syllabi and course assignment samples, and (b) English test papers used in the entrance examinations during the 2013 and 2014 academic years. Finally, approximately 300 Japanese university students will be recruited. The students will be divided into two groups; one group will take the writing tests of a TEAP sample test and the other will take those of a TOEFL iBT sample test accompanied by a cognitive processing questionnaire, which will then be followed by a post-test questionnaire. Furthermore, a semi-structured interview or focus group discussion with the students and their teachers will provide an in-depth understanding of their cognitive processing during the tests.

**Work in Progress 16**

**Presenter:** Bart Deygers  
**Affiliation:** Centre for Language & Education, KULeuven  
**Title:** Determining the predictive validity of high stakes tests of Dutch for academic purposes

This paper presents an ongoing research project that critically examines the two main language tests that grant access to Flemish higher education. Both tests aim to identify language users that have reached the B2 level of the CEFR and - since attaining that level is a prerequisite for academic enrolment - function as gatekeepers to higher education. The two tests that are currently in use for this purpose are ITNA and PTHO, which share the same goal and target population but operate under a different construct. ITNA is a primarily computer-based analytic test, whereas PTHO is a task-based and integrated test. The first part of the presentation focuses on the pilot study; a concurrent validity study that was conducted in 2011-2012. The first leg of this study is quantitative: in order to determine whether both tests judge the same candidate equally, 78 candidates took ITNA and PTHO. The analysis shows that even though both tests correlate moderately (.60), the correlations for the written (.77) and the oral part (.36) differ substantially. The qualitative part of the pilot study helps to explain this difference, which stems from the importance that PTHO attributes to meaning, a criterion that ITNA does not consider. After discussing the pilot study, this paper discusses the larger research project that consists of two research questions. The first research question concerns the diagnostic potential and predictive validity of PTHO and ITNA. In order to determine the extent to which these entry tests of Dutch for academic purposes predict future linguistic academic proficiency, 300 test takers will sit the ITNA and the PTHO tests at the sites of the three main Flemish associations of higher education. This study will show whether these tests judge the same student in a similar way. It will also allow for the selection of 99 L2 students who will be tracked for one year. One third of these students will have passed both PTHO and ITNA, the others will have passed one test, but not the other. Since all respondents will be allowed to enroll in higher education, it will be possible to map the L2 progress of each test taker and to map this data onto the original test score. The progress and real-life performance of these candidates will offer information concerning the predictive validity of PTHO and ITNA. The second RQ concerns PTHO’s and ITNA’s generalizability to the broader field of academic language. Subject specialist focus groups will serve to determine the generalizability of ITNA and PTHO. These subject specialists will be academic staff holding a teaching position in the first year of higher education. The focus group respondents will be asked to comment on the authenticity, representativeness and generalizability of the test tasks in ITNA and PTHO. By discussing previous research and by introducing a future large-scale predictive validity study the researchers wish to stimulate discussion about the predictive potential of language tests. They would also welcome feedback on the ongoing research they propose.
Work in Progress 17: Thursday, June 5, 10:20

Presenter: Kris Buyse and Emma De Rijk

Affiliation: KU Leuven

Title: To CEFR or not? Towards a universal template for the assessment of oral language skills

Since several decades, the three major standards for a high quality assessment are validity, reliability and transparency. In the last two decades, with communicative and task based approaches as mainstream language teaching methodologies, emphasis is not only laid on transparency and reliability —promoting the use of more transparent templates and scales in a formative assessment process—, but also on validity, aiming at assessments in authentic communicative contexts in which not only knowledge, but also skills and attitudes are required (Keeves 1994; Parrondo Rodríguez 2004). Based on the Common Framework of European Reference for Languages (CEFR(L)), a growing number of templates are created for the communicative assessment of language competence templates. In order to check the effectiveness and adequacy of current evaluation templates, we compiled between 2009 and 2012 a corpus of 100 templates coming from three different education types and languages, all CEFR(L) levels, and both from general and specific purposes language courses. On the basis of a state of the art of current language assessment, we formulated the following research questions: (i) are the evaluations based on those templates sufficiently valid? (ii) are they reliable? (iii) are they transparent? (iv) do they take into account the progress made by the learner? are they formative and/or summative? (v) are they easy to use? At the same time, we asked 74 colleagues to participate in a survey in order to discover their views regarding the research questions. Nevertheless, we did not find any templates which, besides being sufficiently valid, reliable and transparent, (1) assess at the same time the communicative output and the linguistic accuracy (i.e. a concern that we share with the vast majority of the colleagues we interviewed), (2) combining the advantages of synthetic and analytical assessments, (3) being sufficiently flexible for use in any type of education and their corresponding objectives, (4) allowing to combine summative and formative assessment of knowledge, skills and attitudes, (5) turning into a sufficiently detailed and reliable instrument to give feedback to the student. The templates analysed are based on and referring to the CEFR criteria and descriptors, but their flexibility is minimal: every change in the objectives, level or education type requires the development of another template: CITO/SLO (2011: 23) points out that, as CEFR distinguishes between oral expression, interaction and understanding, each skill also requires its own evaluation. Therefore, after presenting briefly the state of the art on performance assessment (Meyer 1992; Gipps 1994) in general and the objectives and demands of current language teaching (such as the need to integrate triggers and rewards, see Gijselaers 2007), we will evaluate our corpus of assessment templates. On the basis of the data of our analysis, we will finally propose flexible criteria and templates for a communicative assessment of oral language skills, based on the CEFR, but going beyond it allowing us to adapt the assessment to the objectives and demands of each language course without losing sight of transparency, validity, reliability and user-friendliness.

Work in Progress 18: Thursday, June 5, 10:20

Presenter: Ying Zheng and David Booth

Affiliation: Southampton University

Title: Exploring the effectiveness of online feedback at different CEFR levels

Over the years, the role played by corrective feedback in language acquisition has become a highly controversial issue (Price, Handley, Millar, & O’Donovan, 2010). The majority of studies in Second Language
Acquisition (SLA) have focused on feedback given in classrooms or on homework/assignments. A strong hypothesis is that providing effective feedback is crucial in facilitating students’ learning as they are usually limited in what they are able to notice in the mistakes/errors they make. Nowadays, with the ‘digital natives’ becoming the dominant group in the language learning population worldwide, it is of vital importance to look into how language teaching and language learning can be effectively intertwined with good assessment practices by making effective use of online feedback mechanism. This study will examine the nature of corrective feedback and the extent of impact different modes of feedback can have on language learners’ interlanguage or intake in the context of online assessment modules. Language learners of different CEFR proficiency levels in a UK university will be invited to participate in this study. After they complete an online assessment module at their respective proficiency levels, they will be asked to reflect on the online feedback they receive in terms of the formats of feedback, the types of feedback, i.e., analytic/linguistic feedback (e.g., grammar, semantics), and holistic feedback (e.g., English writing convention, genre), and the effects of feedback on their perceived improvement. One particular research interest of this study is to find out whether online feedback should be tailored to different CEFR proficiency levels, if so, in what ways and to what extent. Currently, the project is in the process of developing this online assessment system based on the analysis of common errors and the next step is to validate the efficacy of different types of feedback in terms of their temporal dimension and relational dimension. A questionnaire will be issued to the invited participants to gather their opinions of the online feedback provision. Retrospective interviews will be carried out with a number of learners and their teachers to probe into their perceptions of the effectiveness of the feedback they receive in their online assessment modules and how they think that feedback provided influences their follow-up learning and teaching as well as their possible repeated attempt(s) on future assessment modules. Theoretically, the results of this project can contribute to the theoretical development and integration of SLA on feedback and diagnostic online language testing. In the SLA field, research has demonstrated the significance of certain types of feedback (but not all) for L2 development, but how this knowledge can be transplanted to the language testing area for diagnostic purposes is a promising area with uncertain answers. Pedagogically, the results of this project can also shed light on effective language learning and teaching that could potentially intertwine with the good practices in diagnostic assessment and feedback.

Work in Progress 19: Thursday, June 5, 10:20

Presenter: Ying Zheng

Affiliation: University of Southampton

Title: Linking a test to CEFR: An application of extended descriptors

This work in progress presents the methodology adopted by Pearson in using an extended version of CEFR descriptor set in the development of test items for a new Pearson test. The test was designed to be aligned to both CEFR and GSE scale. The GSE is a numerical scale of language proficiency aligned to the CEFR, originally used to report scores on Pearson Test of English Academic. In order to align the test to both CEFR and GSE scale, the work commenced at core; Experienced items writers with a detailed knowledge of CEFR and descriptors were selected and trained on the GSE scale and the extended CEFR descriptors. These descriptors were the outcome of a research study conducted involving 216 teachers and 89 trained Pearson staff members in over fifty countries worldwide. Extended descriptors include an additional 89 can-do-statements to the original CEFR descriptors. Each CEFR band was divided into three sub categories of high, medium and low (expect for A1 and C1 which were divided into two categories of high and low). These bands were linked to a range on the GSE scale. Item writers were asked to select and tag each item they wrote with a suitable can-do statement and an appropriate GSE value. Once the item development was completed, field testing was conducted in several regions; South America, Asia, and Europe. The US, Canada, Australia and the UK were also
targeted in order to gather a wide variety of other L1s from international students studying there. Candidates from over 30 countries sat the test and the collected data is being analysed. The field testing is drawing to a close at this stage, and this work in progress will report whether item difficulty parameters generated from IRT analysis would confirm item writers’ estimation of the CEFR levels as well as the consistency of their initial assignment of can-do statements to the test items, as compared to the empirically validated item difficulty as a whole. The aim of this study is to be able to expand the CEFR can-do statements and provide an extended list of more granular statements with empirically consolidated evidence. In addition, from the previously completed studies, the extended can-do statements have also been linked to the GSE Syllabus; therefore the test has a clear and identifiable link to the Syllabus. This could potentially help teachers identify students’ abilities and recommend suitable stretch or remedial activities from the learning content.

Work in Progress 20: Thursday, June 5, 10:20

Presenter: Timothy Farnsworth

Affiliation: CUNY Hunter College

Title: Characterizing beginning teachers’ formative language assessments

The area of language assessment literacy has recently been the subject of much attention, notably with the publication of the special issue of Language Testing in July 2013. Experts have identified trends in language assessment education, both in textbooks and coursework (Davies, 2008; Brown & Bailey, 2008), and others have used survey methodologies to examine language teacher and language tester knowledge (Malone, 2013; Fulcher, 2012). Other work in this area has focused on assessment literacy from a pedagogical perspective, examining issues such as the use of Microsoft Excel to teach and conduct classroom-level statistical analyses (Carr, 2008). However, fewer researchers have examined the formative assessments designed by beginning language educators to determine their usefulness and to find ways to improve the quality of these teacher-developed instruments through improved assessment education. In the current study, which is a work in progress, language educators with zero to five years’ experience in a large urban public school setting (primary and secondary education) in the United States will be asked to contribute assessment materials they have developed for use in their own classrooms. The contributions will include the assessments themselves as well as a narrative describing the intended purposes of the assessment and the curriculum within which it functions. This work will be qualitatively categorized using an instrument developed for the purpose of characterizing difficulties and strengths of the submitted work. Results will have significance for teacher education programs, and a subsequent publication or presentation will provide concrete recommendations for language teacher educators preparing future educators. In the work in progress session at the Language Testing Research Colloquium, a draft of the evaluation instrument and numerous sample assessments will be provided to attendees for commentary and suggestions. The primary goal of the session will be to solicit ideas for cataloging and characterizing the strengths and weaknesses of the submitted materials. References Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? Language Testing, 25(3), 349-383. Carr, N. T. (2008). Using Microsoft Excel to calculate descriptive statistics and create graphs. Language Assessment Quarterly, 5(1), 43-62. Davies, A. (2008). Textbook trends in teaching language testing. Language Testing, 25(3): 327–347. Fulcher, G. (2012). Assessment literacy for the language classroom. Language Assessment Quarterly, 9(2), 113-132.
Work in Progress 21: Thursday, June 5, 10:20

Presenter: Romulo Guedez-Fernandez

Affiliation: The University of the West Indies

Title: Development and Validation of a Framework for the Assessment of Interactive Speaking Performance

This investigation was motivated by the need for an adequate instrument to assess Interactive Speaking for a Spanish Programme at the university level in the English-speaking Caribbean. Participants will be Spanish Majors/Minors who have already completed 7 years of high school Spanish Language instruction. This will be the first time in which they would sit an oral examination together with another classmate (peer-to-peer interaction) while being assessed by two examiners. Therefore, relying on existing models (Bachman, 1990; Bachman and Palmer, 1996; McNamara, 1996; Skehan, 1998; Fulcher, 2003) and based on the descriptions of the Common European Framework of Reference for Languages (CEFR), the purpose of this research is to develop and validate a framework for the assessment of Foreign Language (FL) interactive speaking performance. This new framework will provide the foundation for the development of appropriate scales for describing FL interactive speaking performance. The size of the sample is estimated to be about 150 year one, two and three majors/minors. A multi-method approach will underpin this investigation. The data collection instruments will include tests scores, questionnaires, semi structured interviews as well as focus groups in which instructor, raters and other stakeholders will be involved. The analysis of both quantitative and qualitative data will help to develop and validate the new framework and to determine in what ways the existing framework is inadequate for the assessment of interactive speaking. This study also aims to determine to what extent the newly developed framework can promote development of listening comprehension and interactive speaking of year one, two and three university student’s. The findings of this study will have implications for the construction of appropriate interactive speaking assessment scales and for improving our Spanish programme. In the Work in Progress session data from a pilot study will be presented and feedback about data methods of analysis and interpretation will be sought.

Work in Progress 22: Thursday, June 5, 10:20

Presenter: Maggie Dunlop

Affiliation: Ontario Institute for Studies in Education, University of Toronto

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

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 at a medium-size bilingual Canadian undergraduate institution which
developed a computer-based diagnostic placement test for French as a Second Language for Academic Purposes (FSLAP). The assessment was designed according to cognitive language development principles, and describes current learner abilities in key cognitive language skills. Students participating in field testing of FSLAP test items (n=900) were invited to participate in this study. Study participants (n=260) are randomly assigned to one of two treatment groups. One treatment group is receiving ‘general’ feedback (n=130) and one group is receiving ‘specific’ feedback (n=130). Four data collection activities are taking place, each activity maintaining a 50:50 ratio of treatment groups. The four activities are designed to collect multiple sources of data to inform the following research questions: How do learners engage with and process computer-based feedback differently? What features of descriptive, substantive computer-based feedback facilitate deeper cognitive processing among language learners? Two hundred participants will complete either an immediate or delayed survey, each designed to draw out cognitive aspects of attention, processing, interaction, and application. A further 30 participants will receive computer-based feedback while using an eye tracking device then complete a stimulated recall interview to elicit information on participants’ cognitive attention, processing and interaction. A final 30 participants will receive feedback and subsequently complete a delayed recall interview to elicit information on ongoing processing, interaction and application. Data collection will be completed December 2013. The results of this study, which will be available by June 2014, 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.

Work in Progress 23: Thursday, June 5, 10:20

Presenter: Daniel Reed and Aaron Ohlrogge

Affiliation: Michigan State University

Title: The issue of age-appropriateness in establishing fair and valid test tasks, practices, and policies in a universal framework

The issues and challenges associated with the need to create age-appropriate tests that align with a universal scale such as the Common European Framework of Reference are considerable. The challenges and complexities of English language testing in Greece, which has one of the world’s youngest English language testing populations, have been documented by a number of scholars (cf. Gabrielatos, 1993; Gass and Reed, 2011; Kenny, 1995; Prodromou, 1993; Tsagari, 2009). Gass and Reed (2011) emphasize the importance of taking into account intercultural sensitivities and age-appropriateness in developing an English test for Greece based on a case study involving an American university collaborating with two educational institutions in Greece. However, whereas their discussion focuses primarily on the intercultural communication among the collaborators and how a comprehensive understanding of the cultural context is essential to the success of the collaborative process, the authors of the current presentation will emphasize the importance of what tasks, practices, and policies are age-appropriate. In particular, they will present an account of the details of this project that exemplify how fair and high-quality testing practices can be established through an intercultural collaboration. The presentation will highlight details of multiple facets of the project including how age-appropriate content was established through interviews and classroom observation, how design features of the test were determined, and what measures were taken by the collaborating institutions for planning and administering the tests to the young population. Steps that have been implemented in addition to interviewing students, teachers, and school owners included a thorough cultural sensitivity review process focused on the age groups typical of the testing populations, holding regular tripartite, joint coordination meetings in Greece to discuss issues proactively and as they arise, insistence on standardized test
administration practices at all test sites (even with regard to details like the kinds of pencils used for the writing), and processes for obtaining and responding to feedback from examinees, examiners, teachers, school owners, and other stakeholders. These operational features will be illustrated and discussed as part of the ongoing improvements of the testing program.

Work in Progress 24: Thursday, June 5, 10:20

Presenter: Robin Stevens

Affiliation: Rosetta Stone, Inc.

Title: Self-assessment as an efficient screener for placement into CEFR-aligned English curricula

This paper describes a pilot study in the development and validation of a screening instrument designed to facilitate the placement of English learners into CEFR-aligned curricula. The screener is intended to efficiently place learners into two curriculum levels – above and below the A2/B1 threshold – with minimal error. This approach differs from that taken by the European Language Portfolio by focusing on self-assessment with specific stimuli rather than general statements of competence (Council of Europe, 2013). Self-assessment has typically been seen as a tool for the classroom and/or as a way of enhancing learner motivation (Black & William, 1998; Ross, 2005). Self-assessment tools have been shown to correlate moderately to highly with evaluation from teachers and learner peers (Oscarson, 1989; Heilenman, 1990) but the literature comparing learner self-assessment with independent measures such as standardized assessments is limited. Directed self-placement is now used more frequently to place US students into level-appropriate college composition courses (Gere et al, 2010; Jeffrey, 2009), but self-assessment of language skills other than writing for placement purposes remains uncommon. A small body of literature on the use of self-assessment as a predictor of reading comprehension (Le Blanc & Plainchaud, 1985; Hunter & Schmidt, 1990) exists but findings are inconclusive. Likewise, the efficacy of self-assessment as a tool for reliably and efficiently placing language learners into level-appropriate courses has not been established in the second language assessment literature. In an effort to build validity evidence on the use of self-assessment as an effective placement tool, the researchers undertook an exploratory study to examine the use of a placement instrument with an integrated self-assessment tool. Learners took a pilot screener composed of three tasks. The first task utilizes traditional multiple-choice vocabulary items. The second requires the learner to read a series of CEFR-leveled sentences and self-evaluate understanding of each sentence based on a five-point scale. The third employs the same sentences used in the self-assessment but requires learners to demonstrate their understanding of each sentence via multiple-choice reading comprehension items. These items probe learners’ sentence-level comprehension of reading for gist and paraphrasing skills. The research questions addressed were: 1 Does self-evaluation (via the Self-Assessment) predict performance on the Sentence Comprehension items? 2 Does self-evaluation (via the Self-Assessment) predict performance on the Vocabulary items? 3 Does performance on the screener correlate with scores on other tests (concurrent validity)? 4 Does performance on the screener result in accurate placement? Results from a pilot study (n > 200) are presented emphasizing the efficacy of self-assessment in predicting placement, potentially making learner placement more efficient without sacrificing reliability. Results are discussed in the context of a broader research program to validate an innovative self-assessment screening technique for use in raising learner self-awareness and enhancing learner-directed targeting of goals. The experimental design allows for testing of hypotheses (e.g., “Do learners provide more accurate self-assessment if they are confronted with actual comprehension tasks?”) and allows identification of the optimal choices for the design of self-assessment-based placement tests.
**Abstracts June 5**

**Work in Progress 25:** Thursday, June 5, 10:20

**Presenter:** Roxanne Wong and Carl Swartz

**Affiliation:** City University of Hong Kong

**Title:** An investigation into alignment of Lexile scores and analytical rubrics

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.

**Work in Progress 26:** Thursday, June 5, 10:20

**Presenter:** Candace Farris

**Affiliation:** McGill University

**Title:** Seeking the construct of interactional effectiveness in global aviation: A mixed-methods study

My objective in presenting a work-in-progress report is two-fold: First, to present the results of the first phase of a three-phase research project; second, to receive peer feedback regarding the integration of the results of the first and second phases of the study in the development of the methodology for the third phase. In 2004 the International Civil Aviation Organization (ICAO) introduced worldwide language proficiency requirements (LPRs). The ICAO LPRs stipulate that pilots and air traffic controllers (ATC) operating in contexts where the use of English may be required demonstrate proficiency in English. The goal of the LPRs is to ensure that ATC and pilots can communicative effectively in non-routine aviation situations. The LPRs apply to both native and non-
native or first and second-language English speakers in all 197 ICAO contracting states. While the LPRs mark an important step in recognizing the role of effective communication for air safety, the construct of aviation English remains unclear. ICAO has provided the LPRs assessment criteria in the form of six analytic rating scales based on linguistic constructs (pronunciation, structure, vocabulary, comprehension, fluency and interaction). Candidates must be awarded at minimum an operational level (the fourth of six levels) of proficiency on all six scales in order to be certified by their national regulator. Despite the high stakes nature of this assessment, validity evidence for the constructs and criteria of the scales in relation to the goal of the ICAO LPRs is lacking. As such, the objectives of my three-phase study are, 1) to identify skills essential to effective controller-pilot interactions based on empirical evidence, 2) to validate the findings in a multi-method study involving stakeholders, and 3) to provide recommendations for revisions to the ICAO LPRs based on empirical evidence. My study is conducted within a socio-cognitive theoretical framework that recognizes both the individual aspects of communication and the co-constructed nature of discourse. I investigate cognitive, environmental and social factors that impact communication in the operational environment. The first phase of my project is a mixed-methods (qualitative and quantitative) meta-synthesis of discourse studies in ATC-pilot communications. The research question guiding Phase One is the following: What communication and interactional skills are important for first and second language English speaking pilots and controllers in the global aviation context? The purpose of meta-synthesis is to synthesize the results of previous studies in a particular domain in order to respond to the researcher’s particular research question. Whereas quantitative meta-analysis relies on statistical procedures and inferences, qualitative meta-synthesis relies largely on narrative explanation. The synthesis conducted in Phase 1 of my study includes data gleaned from both quantitative and qualitative studies, and includes both quantitative and qualitative methods of meta-analysis. Phase 2 involves the validation of the findings of Phase 1 in a multi-method (focus groups, survey) study involving stakeholder participants. Finally, Phase 3 will involve providing recommendations for the operationalization of the construct interactional effectiveness in the global aviation context.

Work in Progress 27: Thursday, June 5, 10:20
Presenter: Bjorn Norrbom
Affiliation: NCA
Title: Developing the Arabic Placement Test (APT)
Every year a large number of students from a range of countries and with many different first languages arrive in Saudi Arabia to study in Saudi universities. A majority of these are not proficient in Arabic and are therefore required to take preparatory courses given by their university before being allowed to commence academic studies. In addition, a growing number of foreign professionals working in Saudi Arabia are taking Arabic language courses in order to interact more actively with Saudi society. This group is also highly heterogeneous with regards to country of origin and first language. The courses they take are given by a number of different institutions, including local colleges and universities, private language schools, and charity organizations. None of these are known to use any international or local scale/framework (ACTFL/CEFR/ISLPR) for placement, assessment, or teaching purposes. Consequently it is very difficult to assess what frequently used terms such as “level 1” or “intermediate” actually mean. Learners in both groups vary greatly in their levels of Arabic proficiency and therefore a reliable and validated tool is needed in order to assess their levels of proficiency and report these in relation to a recognized language proficiency scale. This will no doubt help increase the transparency of test results for all stakeholders. The present paper describes the development of such a tool, namely the Arabic Placement Test (APT). It aims to cater for both groups of learners mentioned above. Hence it should be able to correctly place both students of Academic and General Arabic into adequate levels. The APT is primarily aimed at CEFR levels A2 and B1 as these are believed to be the most relevant ones for placement
purposes, for both preparatory academic and general purposes. The test will focus on grammar and vocabulary following the rationale of the Password Test (University of Bedfordshire 2009). In developing English L2 exams related to the CEFR there is a host of resources available such as quality textbooks linked to the Framework, the Breakthrough, Waystage, Threshold, and Vantage specifications, English Profile etc. There are no equivalents for any of these for L2 Arabic. This, of course, presents a major challenge in developing test specifications. The APT will be developed in parallel to the creation of Reference Level Descriptors (RLDs) for Arabic. These will initially cover levels A1-B1 and the goal is to produce a learner-based online tool similar to English Profile. Therefore, the paper also provides a brief description of how the RLDs are developed in a three-phase process. Intuitive RLDs are developed by a group of experts, these will then inform initial test specifications and the learner data from the tests will be used to inform the final RLDs in the form of an online tool for L2 Arabic grammar, vocabulary, and, eventually, language functions. As work with the RLDs proceed, APT test specifications will be adapted.

Work in Progress 28: Thursday, June 5, 10:20

Presenter: Ricky Jeffrey

Affiliation: University of Nottingham

Title: Using tutor-to-student feedback comments as verbal protocols for performance descriptor creation

BACKGROUND/OBJECTIVE: While the primary method of creating performance descriptors is the elicitation of new verbal protocols through considering performance samples (Council of Europe, 2001) - a process which is demanding on resources - my study considers an alternative, already-existing data source which could be used instead: written teacher-to-student feedback comments. In the context of EAP courses, for example, it is normal for a written coursework assessment to result in both a score and a formative feedback comment of 100-200 words. Although these feedback comments usually have the primary purpose of communicating diagnostic areas for development to students, rather than simply describing performance, they can in some sense be regarded as verbal protocols, of the "non-mediated concurrent think-aloud" sub-type (Green, 1998), since their content relates directly to the quality of the performance samples. Also, since they were originally generated for a different purpose, they may overcome some of the artificiality of verbal protocols generated purely for the purpose of performance descriptor creation (Bowles, 2010). Many EAP programmes have electronic records of such feedback comments and their related scores going back several years, providing sufficient data to analyse feedback comments at each band of achievement, in order to identify performance descriptors. Because EAP programmes have the primary remit of teaching, and have limited resources to devote to assessment, this time-saving strategy, wherein pre-existing data is used rather than spending staff time on generating new verbal protocols, could enable some institutions to create performance descriptors in cases where it was not previously feasible. This method could also help generate more discipline-specific performance descriptors, since EAP tutors have often built up significant expertise about a learner’s target academic context (BALEAP, 2008), and the feedback comments are designed specifically to relate a student’s performance to a specific target academic context. METHOD: A bank of 600 tutor-to-student feedback comments given for a written EAP coursework assignment at the University of Nottingham's campus in Ningbo, China, will be sampled with the aim of finding commonly highlighted features at each band of performance. A sample of 30 comments at each of the 5 main performance bands will be analysed. If common features can be identified at each band, they will then be fashioned into performance descriptors for that corresponding band. Since the performance descriptors would relate to detailed levels of performance on a specific task, they should lend themselves to use in a rating scale (Council of Europe, 2001). To assess their suitability, a new rating scale will be developed using these newly generated descriptors, and this scale will be piloted with raters marking authentic coursework assignments from a prior assessment administration. Raters will then be interviewed to assess the suitability of the performance descriptors. A control group of raters will also mark
using the programme’s existing rating scale. The marks from both groups, as well as the real marks given for those essays during last year’s live administration, will be analysed for correlation and inter-rater reliability.

**Work in Progress 29:** Thursday, June 5, 10:20

**Presenter:** Glyn Jones

**Affiliation:** freelance

**Title:** The CEFR revisited: a replication study

This session will report on the current state of a study whose aim is to replicate the research undertaken by North (2000) which led to the calibration of the Common European Framework of Reference (CEFR) descriptors. The CEFR has been enormously influential, but has been the subject of considerable controversy. Among other perceived shortcomings, critics have pointed to • Inconsistencies in the terminology used in scale descriptors (Alderson et al., 2004) • Gaps in the coverage of descriptors (the paucity of descriptors at the C levels; the preponderance of descriptors for spoken language at the expense of written) • The limited cultural range of North’s informants • The fact that the calibration is based only on teachers’ shared beliefs, not observed features of learners’ performance (Fulcher, 2004; Hulstijn, 2007). In addition, there have been differences in understanding of its theoretical basis (do the descriptors constitute the framework, or are they merely illustrative of some underlying construct? To be “at” a level, does a learner have to match all corresponding descriptors or only some of them?) and wide differences in interpretation of the scales when it comes to putting them into practice (“Your B1 is nowhere near my B1”). In his study North collected a large pool of descriptors and had these rated by expert judges (teachers of EFL based in Switzerland). He then had the judges use the descriptors to rate video-recorded samples of spoken English produced by Swiss learners at a range of levels. The results of the rating exercises were used to calibrate the descriptors on a common scale. The present study will apply North’s method using a larger and culturally more diverse panel of expert judges. In addition it will incorporate elements aimed at addressing some of the issues enumerated above, notably • The inclusion additional descriptors designed to make for a more consistent and comprehensive framework • Analysis of samples of written and spoken production with a view to providing some empirical validation of the expert judgments by relating these to identifiable linguistic features. The presenter will particularly welcome participants’ suggestions regarding • Criteria for choosing, or principles for formulating, additional descriptors to supplement those in the CEFR • Approaches to the analysis of performance samples likely to be most fruitful in providing validation evidence in relation to expert judgments of level. References Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). The development of specifications for item development and classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Reading and Listening: Final report of The Dutch CEF Construct Project. Lancaster. Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. Language Assessment Quarterly: An International Journal, 1(4), 253-266. Hulstijn, J.H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency The Modern Language Journal, 91(4), 663-667. North, B. (2000). The development of a common framework scale of language proficiency. New York: Peter Lang. (485 words)
Abstracts June 5

Work in Progress 30: Thursday, June 5, 10:20

Presenter: Nathaniel Owen
Affiliation: University of Leicester
Title: An evidence-centred approach to reverse engineering

Like software systems, tests are designed to fulfill a specific purpose with particular users and purposes in mind. Similar to software companies, test companies are inclined to protect the commercial interests of their products and not publish their ‘source code’ (test specifications). A stakeholder with an interest in the source code of a software system may choose to ‘reverse engineer’ the publicly-available software to extract the source code which they may use to design complementary software programmes. This theoretical paper explores the concept of ‘reverse engineering’ in the context of language testing, using the language of software development as metaphor. Firstly, it describes the concept of ‘reverse engineering’ (RE) as it exists in the literature. Secondly, it outlines how this concept relates to existing ideas about the proposed content of test specification blueprints. Thirdly, it outlines a new holistic framework of reverse engineering by adopting an ‘evidence-centred design’ (ECD) approach, following Mislevy et al. (2003). ‘Reverse engineering’ is “the creation of a test specification from representative test items/tasks” (Davidson & Lynch, 2002: 41) and by Fulcher and Davidson as an “analytical process of test creation that begins with actual test questions and infers the guiding language that drives it” (Fulcher & Davidson, 2007: 57). An additional purpose of reverse engineering is to “clarify if a particular test is spec-driven” (ibid.: 42) by discerning the degree of congruence between items. The procedure of reverse engineering involves “analysis of a set of existing test tasks, to decide if they are similar, and then to induce a spec that could link them together” (ibid.: 42, emphasis added). Fulcher & Davidson define test specifications as a ‘generative blueprint’ that informs test design and content, and may be used iteratively in the creation of carefully-designed test items. Specs should have a descriptive and communicative function to “set up what a test score means… and lay out the behaviour domain being measured” (Popham, 1978: 114-115. Specifications are designed to continually foster dialogue and debate, and thus are consensus-based. They are designed to clarify the content of an assessment, and provide feedback to educational practitioners and test users – a validity narrative as part of an assessment argument (Bachman & Palmer, 2010). The framework forwarded in this paper encompasses Mislevy’s Conceptual Assessment Framework (CAF); the student model, the evidence model and the task model and how these are presented to test takers via the design elements – the presentation model, the assembly model and the delivery model. An ECD approach to RE can be used to ‘clone’ the original test, or guide a complete retrofitting process (surrogacy). The difference between these two procedures is informed by the extent to which one conducts the RE process. An RE framework relates constructs to score use. RE can usefully identify relevant constructs and their relationship to test score uses. This is an important step for limiting the claims that can be made on the basis of test scores and for creating a meaningful audit trail for stakeholders.

Parallel Session 9.1: Thursday, June 5, 11:25

Presenter: Hui Feng
Affiliation: Guangdong University of Foreign Studies
Title: Teacher-Student Rapport and Its Effect on the Assessment of Writing in the EFL Classroom

In writing tests in the EFL classroom, the students’ performance is typically rated by their own teacher, who has built a rapport with the students after teaching them for a period of time. In this context, it is inevitable that the teacher’s rating behavior will be influenced by classroom-related background factors, such as the teacher’s
expectations, the syllabus, and teaching experience. It is thus interesting to compare the teacher’s behavior in rating the writings of his own class, when he has built a rapport with his students, and his behavior in rating the writings of another class, with whom he has no rapport. The potential effect of the teacher-student rapport on the teacher’s rating behavior will provide further insight into how rater background brings about rater variability, which will then be interpreted from the angles of both summative and formative assessment. To understand the potential effect of the teacher-student rapport on the teacher’s rating behavior, an exploratory study was conducted in the context of a weekly test of English in two senior classes (N = 60) in a senior high school in China. While rating the writing samples of their own class as usual, the EFL teachers of the two classes were asked to rate the writing samples of the other class as well. The teachers also commented on the writing samples in the rating process, and were interviewed after the rating about their knowledge and expectation of their students, as well as their instructional emphases. The mean ratings given by the two raters to the two classes were compared through Mixed ANOVA. The results indicated a significant main effect of raters, but not of classes. No significant interaction between raters and classes was found. Both raters gave higher scores to their own class than to the other class. The raters’ comments were coded by two independent coders into four major categories: language, content, organization, and mechanism. The chi-square tests of independence of the frequency distribution of these comments indicated a significant three-way association between classes, teachers, and categories. One teacher consistently emphasized organization while the other played down the importance of organization. Subsequent interview showed that when the students’ identity was known, both teachers tended to give higher scores to their own students in order to “encourage” them. Furthermore, the emphases on different categories in the rating process reflected the general beliefs of the teachers in their teaching experience. In summary, teacher-class rapport was found to have an effect on both the process and product of essay rating in the EFL classroom. In terms of summative assessment, the differences attributable to teacher-class rapport could be interpreted as construct-irrelevant variance. However, from the perspective of formative assessment, the teachers’ rating practices are justified as part of their teaching practices intended for better learning and teaching. The tension between these two interpretations will be discussed in the argument-based framework of validity.

Parallel Session 9.2: Thursday, June 5, 11:25

Presenter: Bill Bonk

Affiliation: Pearson Knowledge Technologies

Title: Beyond pronunciation and fluency: automated evaluation of prosody and accentedness

Pronunciation and fluency are two common traits often evaluated in the tests that employ automated speech evaluation (ASE) systems (e.g., TOEFL Practice Online, PTE-Academic, and Versant). If other traits could also be evaluated, ASE could become more useful for describing the test-taker’s strengths or weaknesses or measuring progress. This paper reports on two studies on automatic evaluation of prosody and the degree of accentedness. The prosody research investigated methods for automatically measuring L2 speakers’ ability to apply appropriate prosody when reading short passages aloud. Prosody is often characterized by syllable length, loudness, and pitch. Therefore, in this study, prosody was measured by modeling variation in three components in spoken responses: duration, energy, and fundamental frequency (F0). Duration included segmental duration for phonemes and inter-word silences. A total of 85 passages were selected from an English for Academic Purpose (EAP) test. For each passage, 80 learner data and 15 native English speaker data were used to train the three models. Separately, 340 responses were set aside for fine-tuning the trained models. For an independent validation purpose, responses from 158 subjects were further kept. Trained human raters provided ratings using a 0-5 scale. Regression analyses showed that while F0 and energy were both strong predictors of the human prosody ratings, the duration measures were the best (r=0.76 for segment
duration and \( r=0.54 \) for inter-word silences). The final multiple linear regression model yielded machine scores that correlated at 0.80 with the human ratings. The result was better than the human-to-human correlation \( (r=0.75) \). The second research study investigated the feasibility of developing an automated evaluation system on the degree of accentuatedness in Indian English. Trudgill & Hannah (2008) identified 13 phonemes as the features of Indian English. Based on this study, an Indian English phoneme group was created including labiodental fricative /v/, bilabial approximant /w/, plosives /p/, /t/, /k/, alveolars /t/, /d/, /s/, /z/, /l/, /r/, etc. Three other phoneme classes were also created: all phonemes, all vowels, and all consonants. In addition to these phoneme classes, several other features such as confidence scores produced for each word as part of automatic speech recognition and spectral likelihoods were also extracted. The materials the second experiment were read-aloud passages from the same EAP test in Study 1. The data was augmented by including sentence-level read-aloud responses from another spoken English test. Responses were selected from 825 subjects including Indian English speakers, English learners from other L1s, and American and British English speakers. The response were divided into three sets: training \((n=411)\), development \((n=206)\), and validation \((n=208)\). A group of raters rated the responses using a 0-6 scale. The average of the inter-rater correlations at the response level was 0.77. Results indicate that a set of 18 features (16 phoneme-based features and 2 confidence scores) combined with a few spectral likelihoods can yield Indian accented scores correlated at 0.78 at the response level and 0.84 at the participant level. The results from the two studies suggest an encouraging possibility for providing new traits for test users.

**Parallel Session 9.3:** Thursday, June 5, 11:25

**Presenter:** Gad Lim and Ardeshr Geranpayeh

**Affiliation:** Cambridge English Language Assessment

**Title:** On Standards and Multilingual Frameworks: Evidence from Multiple Standard Setting Exercises

For a framework to function as a framework, it is important that users share a common understanding of it. This in part is the reason why, for example, the Common European Framework of Reference for Languages (CEFR) is accompanied by a manual detailing procedures for developing that shared understanding and how that understanding might be reflected in language examinations (Council of Europe, 2009). Where standardizing to multilingual frameworks in testing is concerned however, two related problems come to mind: First is that frameworks, to work across languages, typically need to be quite general and abstract, the result being that they are potentially inadequately specified, at least for the purposes of standard setting (SS) in testing (Alderson et al, 2006; Fulcher, 2004; Weir, 2005). The first problem might compound the second, which is the observation that multiple SS exercises often result in divergent outcomes. This is generally not thought of as a problem, the consensus explanation in the field being that SS is not an exercise in homing in on a “true”, objectively-existing cut score, but a values-driven enterprise arrived at for some policy-related end (Cizek, 1993; Kane, 1998; Zieky, 2001). That explanation is not satisfactory, however. First, multilingual frameworks such as the Canadian Language Benchmarks (CLB) specify what people can do at particular levels; that is, they are criterion related and have objective referents. In addition to values, therefore, there is at least another component that SS theory needs to account for. Second, actual test information is also often used in SS, making it not just values-driven but at least in part also a data-constrained exercise. Third, the explanation would imply that where the policy-related end is the same, the requirement for convergent outcomes would still hold (Jones, 2009; Reckase, 2009). Standard setting theory therefore needs to appropriately consider and balance the values/policy and the data/empirical dimension of the activity. The problem of inadequately specified frameworks and inadequately theorized procedures are considered in this paper using the results from a number of SS studies. First are the results of two independent SS studies relating the same exam to the same levels of the same framework (the CLB) for the same policy purpose. Second are several SS studies relating...
different exams to the same levels of the same framework (the CEFR). It will be shown that multiple groups of people can come to a common understanding of the frameworks involved. It will also be argued that the task of SS is to find the cut score that objectively corresponds to the given constraint. In such contexts, the activity is ‘expert informed’ rather than ‘values driven’. Multiple SS exercises should be seen as repeated measures where convergent outcomes are expected and divergent outcomes investigated. This approach brings SS closer in line with other areas of assessment, subject to validation and falsification, leading to better-supported validity arguments (Kane, 1992; 2006), and allowing it to be a useful tool in the service of frameworks, standards and standardization.

Parallel Session 9.4: Thursday, June 5, 11:25

Presenter: Johanna Motteram

Affiliation: The University of Adelaide

Title: Appropriate ‘tone’ and the rules of friendship; written language proficiency, letter tasks and ideology.

The term ‘tone’ is used in the public version of the scoring criteria for the IELTS General Training Module Written Task One (GTMWT1) under the heading Task Achievement, including and above band 5. The primary aim of the empirical study reported in this paper is to interrogate the term ‘tone’ through qualitative text analysis. The term ‘tone’ has been imported into language testing’s para-technical lexis from the field of literary criticism. In that context ‘tone’ relates to style, persona, voice and initially, the implicit expression of moral and social values in the text (Abrams 1971). In order to investigate ‘tone’ in the context of IELTS GTMWT1 responses (n=55) to the same task were collected from a diverse range of participants. The scripts were then double rated by senior IELTS markers. Following this, scripts which received a score of 5 or higher from at least one rater were analysed using a framework informed by Systemic Functional Linguistics (Halliday and Matthiessen 2004) and specifically Poynton (1985, 1990). From the perspective of Systemic Functional Linguistics, ‘tone’ can be located primarily within tenor, the expression of the interpersonal metafunction. Halliday sees the interpersonal component of meaning as a cumulative, prosodic colouring of the text (Halliday 1979) and as such, it is not possible to identify the interpersonal as discrete entities within text. Poynton’s work on the semiotics of social relations identifies the linguistic resources of the interpersonal which realise this cumulative, prosodic colouring as the combination of language choices which realise the tenor dimensions power, contact, and affect. Poynton represents these dimensions as clines. The expression of the power relationship between the author and addressee is recognised as being more or less equal or unequal, contact between the correspondents ranges in frequency, and affect may be expressed, suppressed, or somewhere in between. More than 20 specific lexicogrammatical linguistic realisations are identified by Poynton as being significant to the expression of tenor. This study proposes the inclusion of further realisations, such as the use of the interpersonal grammatical metaphor as a resource for authorial intrusion into the text, in the context of the letter task. The analysis is represented to demonstrate the cumulative, colouring nature of the interpersonal throughout the extent of the text. Discussion of the analysis is related to the patterns of language choices which match the tenor patterns dictated by western culture’s ‘rules of friendship’ and the extent to which candidate selection of these tenor patterns are rewarded or penalised by the raters. In some cases it can be argued that the tenor patterns expressed in the texts cause confusion to the raters and thus result in unreliable rating. This research is useful for transparent communication of the meaning of scoring criteria to test candidates and teachers. It should also be useful for rater training. Ultimately, the incorporation of an enriched understanding of the lexicogrammatical realisation of the interpersonal metafunction should lead to expansion of the construct of written language proficiency.
Abstracts June 5

Parallel Session 10.1: Thursday, June 5, 12:00

Presenter: Liying Cheng and Yongfei Wu

Affiliation: Queen’s University

Title: What Assessment Tasks Are Associated With Classroom Assessment Environment?

This study explores the relationships between students’ perceptions of assessment tasks and classroom assessment environment within the university context of teaching English as a foreign language (EFL) in China. As part of the worldwide movement to combine assessment of learning with assessment for learning to promote student learning and supporting classroom teachers to structure their teaching, the Chinese Ministry of Education put forward the College English Curriculum Requirements (CECR) (CMoE, 2004; 2007) to introduce this balanced assessment concept. However, teachers and students are caught between a long examination history and continued large-scale national English language testing, and this recent curriculum reform promoting formative classroom assessment practices. Recent research demonstrates that translating the CECR initiative into classroom practices is complex and involves many factors such as teacher’s beliefs, students’ perceptions, institutional differences, and educational tradition (Chen et al., 2013). Concern also remains as to whether this worldwide (‘foreign’) assessment movement works within the local Chinese university context, which is highlighted as test-centered, textbook-centered, and teacher-centered (Cheng, 2009). How do students perceive their classroom assessment tasks? What is the existing classroom assessment environment? What is the relationship, if any, between the assessment tasks and the classroom assessment environment? This study employed a survey design. The questionnaire was designed based on Dorman & Knightley’s (2006) Perceptions of Assessment Tasks Inventory (PATI) and Alkharusi’s (2011) scale to measure students’ perceptions of the classroom assessment environment. PATI included five subscales: congruence with planned learning, authenticity, student consultation, transparency, and diversity. Alkharusi’s scale comprised two subscales: learning-oriented classroom assessment environment and performance-oriented classroom assessment environment. Participants were 620 Chinese students from three universities in China. Descriptive statistics showed that the students perceived that their English assessment tasks were highly related to what they were learning and these tasks assessed their ability in English. However, students were not involved in developing the criteria for assessment. Exploratory factor analyses confirmed the original five-factor PATI and Alkharusi’s two-factor scales. Multiple regression analysis showed that congruence with planned learning, authenticity, student consultation, and diversity significantly predicted the learning-oriented classroom assessment environment explaining 48% of the variance. Congruence with planned learning was significantly yet negatively associated with the performance-oriented classroom assessment environment along with diversity explaining 12% of the variance. This indicates allowing diversity in the assessment tasks was associated with both learning-oriented and performance-oriented classroom assessment environment.

Diversity is a double-edged sword. On the one hand, allowing students to choose the assessment tasks provides students with opportunities to do what they are most interested in, which might intrinsically motivate students to learn. On the other hand, it also gives students chances to pursue the highest scores and compare their scores with their peers. The findings of the study fills the research gap in our limited understanding of the relationship between classroom assessment tasks and assessment environment and supports teachers to structure their day to day classroom assessment practices.

Parallel Session 10.2: Thursday, June 5, 12:00

Presenter: Summer Loomis

Affiliation: The George Washington University
Title: Investigating "Advanced" in Oral Proficiency Interview (OPI) data from L2 Arabic Speakers

Previous studies of the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) have largely focused on examiner behavior (Halleck, 1996; Thompson, 1995; Shohamy, 1983; Surface & Dierdorff, 2003), and the validity and applications of the OPI (Dandonoli & Henning, 1990; Fulcher, 1996; Henning, 1992; Herzog, 2003; Johnson, 2000; Meredith, 1990; Swender, 2003), with recent contributions on learners of Hindi (Ilieva, 2012), Japanese (Watanabe, 2003), and Russian (Fedchak, 2007; Isurin, 2012; Kagan & Friedman, 2003; Mikhailova, 2005; Rifkin, 2003; Robin, 2012). In contrast, the present study is the first to focus on second language (L2) Arabic speakers and their vocabulary production in the context of the ACTFL OPI. Iwashita et al. (2008) found measures of fluency and vocabulary distinguished best between speech samples from L2 speakers of English at different levels. Likewise, results from Malone 1999 and Read & Nation 2006 indicated that English L2 test takers produced more words as they increased in ability level, although Read & Nation noted significant variation. The present study examined L2 Arabic speech in order to determine if the test takers rated at the higher ACTFL levels of Advanced and Superior produced more words per minute (WPM) and more varied vocabulary, as measured by type-token ratios (TTR), than ACTFL Intermediate speakers. The study also included qualitative observations of test takers’ narrations. Twenty-eight full-length tests were transcribed and used to produce WPM values and TTR for ACTFL Intermediate-Mid, Advanced-Mid, and Superior test takers. Independent-samples t-tests revealed a significant difference between the Advanced-Mid WPM (M=46.87, SD=10.62) and the Intermediate-Mid rating level WPM (M=28.97, SD=4.24); t(11.79)=4.95, p<.001. Likewise, the difference between the Advanced-Mid TTR (M=45.14, SD=4.19) and the Intermediate-Mid TTR (M=41.34, SD=3.54) was significant; t(18)=2.19, p<.05. In contrast, there was no significant difference between Advanced-Mid WPM (M=46.87, SD=10.62) and Superior WPM (M=47.44, SD=12.97); t(16)=0.10, p=0.91 nor between Advanced-Mid TTR (M=45.14, SD=4.19) and Superior TTR (M=49.5, SD=4.39); t(16)=1.57, p=.137. In the qualitative review, 20 narration attempts were found in a data pool of 115 test recordings (because ACTFL OPIs do not have a set format, many test takers were not asked to narrate). According to William Labov, narration must contain events considered “reportable,” rather than quotidian, with the optional inclusion of an orientation or evaluation (Labov, 1972, p. 370). All nine ACTFL Advanced test takers produced reportable events; however, lexical gaps obscured or delayed most of their narration attempts. In contrast, the two ACTFL Superior test takers were both able to produce narrations that fulfilled Labov’s requirements while only three of the nine ACTFL Intermediate speakers did so. The WPM and TTR results suggest that the difference between ACTFL Intermediate and ACTFL Advanced is a salient one in terms of the quantity and diversity of vocabulary being produced. This supports ACTFL’s validity claims regarding the OPI’s ability to differentiate between intermediate and more advanced L2 Arabic speakers. However, despite the result of the qualitative review, the quantitative results do not differentiate between the higher ACTFL Advanced and ACTFL Superior levels respectively.

Parallel Session 10.3: Thursday, June 5, 12:00

Presenter: Spiros Papageorgiou

Affiliation: Educational Testing Service

Title: Towards a universal framework? Issues in alignment of assessments to external frameworks and standards

Towards a universal framework? Issues in alignment of assessments to external frameworks and standards The demand for alignment of assessments to external standards and frameworks has increased worldwide, because of educational reforms which push for accountability, including close monitoring of students’ progress and use of standardized tests (Deville and Chalhoub-Deville, 2011). Alignment in the educational measurement literature typically deals with the extent to which the content of an assessment covers the skills and abilities described in content standards (Webb, 2007). The Common European Framework of Reference (CEFR; Council
of Europe, 2001) has dominated relevant research in the field of second language testing, and has triggered alignment attempts in two interrelated ways (Council of Europe, 2009; Figueras et al., 2005): • content alignment, demonstrating how the content of a language test covers communicative language activities, contexts, text types and other aspects of language ability described in the CEFR • alignment of test scores to support classification decisions in relation to the CEFR levels through the setting of minimum scores (cut scores), which indicate that a test taker has demonstrated performance expected at a given CEFR level. Aligning test content and scores to the six main levels of the CEFR has been advocated as a way to increase transparency of test results for learners, teachers and parents across different educational contexts (Council of Europe, 2009). However, such alignment attempts have attracted considerable criticism regarding the use of the CEFR as a policy tool (McNamara, 2006) or as a tool for test providers to gain wider recognition for their tests for commercial purposes and to substitute procedures for validation (Fulcher, 2004). The aim of this paper is to problematize the theme of the 2014 Language Testing Research Colloquium and in particular the idea of “a common, universally valid descriptive system” in three ways: First, by considering issues in aligning tests to standards and frameworks such as the CEFR or the ACTFL Guidelines. Second, by examining the position of alignment procedures within theories of validation, in particular those that follow an argument-based approach (Chapelle, 2012; Kane 2006, 2013). Third, by reviewing the results of alignment studies that attempted to establish an empirical link between different language frameworks. The paper will conclude by arguing that a universal framework might not always be useful in supporting score-based decisions and interpretations.

Parallel Session 10.4: Thursday, June 5, 12:00

Presenter: Sathena Chan and Chihiro Inoue

Affiliation: CRELLA, University of Bedfordshire

Title: Reading-into-writing: construct definition, task design and scale development

Recent years have seen growing interest in the integrated assessment of language skills, particularly reading-into-writing. This trend is evidenced by the inclusion of reading-into-writing tasks in several English tests used for academic purposes, e.g. Pearson’s Test of English (PTE) Academic, LTTC’s General English Proficiency Test (Advanced Level GEPT), the Georgia State Test of English Proficiency (GSTEP), as well as more general English language tests elsewhere in the world. The development and use of reading-into-writing tasks is enabling closer investigation of the reading-into-writing construct, including consideration of how appropriate assessment criteria and rating scale descriptors can be developed for such tasks. In the direct assessment of second language writing ability, the use of measurement scales with verbal descriptors attached to each scale level describing quality of performance (also referred to as ‘scoring rubrics’) is well-established. Much has been written concerning the relative merits of using holistic and analytic scales for rating written performance. Issues of rater training, monitoring and evaluation have also received considerable attention. Less attention has been directed, however, towards the development of rating scale descriptors for assessing and reporting reading comprehension ability, perhaps because, unlike a writing task, the reading process does not generate a comparable visible outcome that lends itself to observation and evaluation. Assessing reading typically depends upon using objectively-scored test formats, e.g. MCQ, or a written task, e.g. summary. However, the former test method risks interfering with the constructed mental representation that results from reading, while the compositional nature of the latter poses challenges on grounds of cognitive and scoring validity (Taylor 2013). In theory, a reading-into-writing task should offer a new opportunity to address some of these issues. In practice, however, the task of identifying and evaluating the contribution of reading ability to the writing process and product and then reflecting that contribution in a rating scale is not a straightforward matter. This paper reports on attempts to describe and define the construct of reading-into-writing ability through a series of recent test development projects to design integrated tasks for operational language proficiency measures. In particular, the presenters will discuss their experience of establishing the assessment criteria for reading-
into-writing tests based upon the theoretical construct definition underpinning task design as well as on an analysis of features of test taker performance that appear to contribute to task success. Methodologies utilised included questionnaires, verbal protocol analysis, expert panel judgements and automated textual analysis. Findings from such studies can inform decisions about how best to frame reading-into-writing activity through the task setting and instructions, as well as how to account for the reading as well as the writing dimension of test taker performance in the rating scale descriptors, using either a holistic or an analytic approach. This paper has particular relevance to the LTRC2014 conference theme with regard to understanding the differential development and integration of language skills, and providing effective diagnostic support for learners in terms of their relative strengths and weaknesses.

Parallel Session 11.1: Thursday, June 5, 14:00

Presenter: Adam Gesicki and/or Jake Stone Amery Wu

Affiliation: Paragon Testing Enterprises / The University of British Columbia

Title: Classification Consistency and Accuracy of CELPIP-G Test for Canadian Immigration Application

The Canadian English Language Proficiency Index Program® General Test (CELPIP-G Test) is one of the English language tests designated by the Department of Citizenship and Immigration Canada to evaluate an individual’s language proficiency. The main purpose of this study is to evaluate the consistency and accuracy in the classification of test takers’ language proficiency indicated by a CELPIP-G score into the levels described by the 2012 Canadian Language Benchmarks (CLB, 1-12 levels). The decisions on CLB classification have pivotal implications on individual test takers and the Canadian society as a whole, but classification errors are inevitable; hence evaluating classification consistency and accuracy of the CELPIP-G cut scores set for the CLB is of extreme importance. In fact, classification consistency and consistency are viewed as evidence of reliability and validity of test score use (Lee, Hanson, & Brennan, 2002; Wu, Wehrung, & Zumbo, 2012). Method Instrument and Cut Scores for the CLB The CELPIP-G Test consists of four parts: listening, reading, speaking and writing tests. As a test referenced to the criterion of the CLB, cut scores for each of the tests in CELPIP-G Test were set through standard setting. Data Pilot test data were collected for two forms of each of the four tests. The two test forms were randomly assigned to 501 test takers, resulting in 237 test takers for Form A and 264 for Form B. Analysis Many methods (e.g., Huynh, 1976; Livingston & Wingersky, 1979) available for estimating the accuracy and consistency of classifications are based on the assumption that the test consists of equally weighted items, and the test score is the number of items answered correctly. However, test takers’ scores on each of the CELPIP-G subtests are based on the unequally weighted sums of scores of items/dimensions. Therefore, we adopted a more general method presented by Livingston and Lewis (1995), which can be applied to any test score as long as a reliability coefficient can be estimated. To estimate reliabilities, Cronbach’s $\alpha$ coefficient (1951) was used for the selected response tests of listening and reading, while generalizability $\phi$ coefficient (for absolute decision) was used for constructed response of the speaking and writing tests. The analyses were conducted through BB-CLASS (Brennan, 2004). Conclusions and Implications The results show that both forms of the listening and reading tests achieved excellent accuracy and consistency across CELPIP levels 4 to 10. The average probability of correct classification was above 0.93, while the average probability of consistent classification was above 0.89. The performance of the cuts for the speaking test was very good with classification accuracy ranging from 0.89 to 0.98, and classification consistency ranging from 0.85 to 0.97. The classification accuracy and consistency for the writing tests was also good although the corresponding statistics were somewhat lower than the speaking tests. This study provides strong evidence for the CELPIP-G test scores when used for classifying test takers into different CLB levels. Also, the almost identical result between Form A and B helped to verify the credibility of the analyses.
Parallel Session 11.2: Thursday, June 5, 14:00

Presenter: Rachel Brooks

Affiliation: Federal Bureau of Investigation

Title: Comparing Native and Non-native Raters of US Federal Government Speaking Tests

Previous language testing research has largely reported that although many raters’ characteristics affect their evaluations of language assessments (Reed & Cohen, 2001), being a native speaker or non-native speaker rater does not significantly affect final ratings (Kim, 2009). Fundamental differences between native and non-native speakers are also debated in Second Language Acquisition. Some researchers conclude that performance and perception differences exist between native and non-native speakers, while others contend that there is little conclusive evidence to support end state differences. The US Government requires speaking test raters to be both native and high-proficiency speakers of the test language due to the high-level context in which it operates (FBI, 2009). An exploration of how the native speaker construct is operationalized in research reveals a lack of common understanding of the native speaker, referring both to an ideal speaker and a native acquirer of language. This study built on previous research by expanding the breadth of proficiency levels rated to include highly articulate examinees, regrouping the raters to represent three ideas of nativeness (native/non-native speakers, speaking proficiency, and first language), and examining final and linguistic category ratings to reveal the raters’ scoring construct. Thirty FBI speaking testers, native and non-native speakers of English, rated 25 English Speaking Proficiency Tests. They assigned ratings for the overall test and linguistic categories, including functions, organization, structures, vocabulary, fluency, pronunciation and social/cultural appropriateness. Using ANOVAs and MANOVAs, the results indicated no significant difference between the native and non-native speaker groups. When raters were grouped by English proficiency level, lower proficiency raters gave significantly lower ratings, both in the final and in many linguistic category ratings, although with a small effect size. The results suggested that rater training organizations should consider rater proficiency level rather than whether or not they are native speakers. Additionally, they supported the theory that non-native speakers can demonstrate language acquisition equivalent to native speakers, at least when evaluating language. Finally, it was recommended that researchers and testing practitioners that use native speakers should clearly define and justify their use, or avoid the native speaker term altogether.

Parallel Session 11.3: Thursday, June 5, 14:00

Presenter: Eunice Eunhee Jang and Margaret Dunlop

Affiliation: Ontario Institute for Studies in Education, University of Toronto

Title: A Mixed Methods Investigation into the Role of Cognitively Diagnostic Feedback (CDF) on Young ELLs’ Cognition and Goal Orientations

As we witness the unprecedented role of testing in education and society at large, teachers seek diagnostic information about their students’ strengths and areas for improvement in target skills (Alderson, 2005; Brindley, 1998; Spolsky, 1992). Cognitive diagnostic assessment (CDA) departs from the focus of traditional testing on the amount of knowledge learners have acquired toward cognitive skills and strategies that they use to process linguistic knowledge required for successful performance on an assessment task. Diagnostic profiles generated from CDA serve as feedback. Cognitively diagnostic feedback (CDF) is intended to elicit and collect the traces of learners’ cognitive processes and their perceptions about cues taken from the task (Butler & Winne, 1995). It can re-orient learners to self-regulate own learning by evaluating the effectiveness of cognitive strategies. Currently, there is a paucity of empirical research that explains the mechanism of how feedback from assessment interacts with the learners’ mind. It is partly because previous feedback research has been predominantly focused on final products without considering cognitive processes and strategies. In addition,
feedback appears to be treated as a fixed stimulus that learners are expected to respond to passively (Ferris, 2003). Not all learners respond to feedback from assessment effectively or unequivocally. There is a great need of insights into how learners configure feedback and how their affect and goal orientations (Dweck, 1986) influence that configuration procedure and a resulting outcome. In this paper, we report a mixed methods investigation into the effect of CDF on learner cognition and psychology. We offer maximal details of how young learners’ minds interact with CDF and how the interaction is mediated through their goal orientations. We applied the bracketing strategy to the same individual learners (Creswell, 2007) in order to construct their profiles from multiple angles including cognitive, psychological and internal attributes. The cognitive profiles of Grade-6 students (N = 92) were constructed psychometrically based on the application of cognitive diagnostic modeling to the population’s test performance. Their goal orientations were determined using latent trait profiling while learners’ perceived abilities were profiled using their self-assessments. These three different learner profiles were triangulated and further confirmed through think alouds with 21 students in order to understand qualitative differences in their cognitive processes. The results showed that although the majority of students were oriented toward mastery, performance-oriented students’ responses to CDF showed a marked difference from those with the mastery orientation by attributing their success or failure to own ability rather than tasks and efforts. Meantime, students’ goal orientations did not show statistically significant correlations with their overall reading achievement levels. However, when students perceived their parents to be oriented toward performance, they tended to show lower skill mastery levels, suggesting that it is not students’ goal orientations but their perceived parents’ goal orientations that predict students’ cognitive skills in reading. In our presentation, we use students’ verbal accounts to illustrate a typology of learner profiles with distinct responses to CDF.

Parallel Session 11.4: Thursday, June 5, 14:00

Presenter: Guoxing Yu and Shwu-wen Lin

Affiliation: University of Bristol

Title: Test takers’ cognitive processes in completing writing tasks using graph prompts: Exploring evidences of test comparability

IELTS Academic Writing Task One (IELTS-AWT1) and GEPT-Advanced Writing Task Two (GEPT-AWT2) use similar graph prompts, but differ in various aspects, e.g., the amount of verbal and non-verbal information available in the task directions, time allocated to complete the tasks, and the requirement of summarization and interpretation of non-verbal information. It has been a continuous endeavour for mapping the tests to CEFR levels by their respective owners, often from institutional perspectives. This current independent research investigated the comparability between the two tests’ writing components, in terms of test takers’ cognitive processes. In particular, it explored: any similarities and differences in test takers’ cognitive processes when completing such tasks, and the extent to which cognitive processes are affected by test takers’ graphicacy (Wainer 1992), English writing abilities, and the use of different types of graphs (bar, line graph, pie chart etc)as prompts. Thirty-two potential GEPT-Advanced test takers in Taiwan participated in this study. They completed six writing tasks (3 IELTS-AWT1 and 3 GEPT-AWT2) in randomized order, while thinking-aloud their writing processes. Post-test interviews were conducted individually with all participants. Baseline data on the participants’ writing abilities and graphicacy were also collected. The think-aloud protocols were analysed using a qualitative analysis software to identify patterns of cognitive processes, with close reference to the participants’ graphicacy and writing abilities. In the previous study by AUTHOR (2012, IELTS Research Reports No.11) which investigated only IELTS-AWT1 tasks, a framework for analysing cognitive processes in completing IELTS-AWT1 tasks was developed empirically. This framework was employed to analyse the data from the present research. A strikingly similar cognitive process in completing IELTS-AWT1 tasks was noted between
AUTHOR (2012) and the present research. However, differences between IELTS-AWT1 and GEPT-AWT1 were evidenced, in particular, towards the second half of the GEPT-AWT2 tasks which required the participants to interpret and make comments on the data observed from the graphs, while IELTS-AWT1 required the participants to "describe". Findings from this research contribute to our understanding about the comparability of GEPT-AWT2 and IELTS-AWT1, the ongoing validation and development of the two tests as well as other language test tasks that use graphs as prompts (e.g., the Pearson PTE speaking task to orally describe a graph). Methodologically, it highlights the importance of investigating test takers’ cognitive processes when we conduct test mapping or alignment exercise. (382 words)

Parallel Session 12.1: Thursday, June 5, 14:35
Presenter: Sathena Chan and Rachel Yi-fen Wu
Affiliation: CRELLA, University of Bedfordshire
Title: Collecting context and cognitive validity evidence of the GEPT Advanced Level Writing Test

Language tests for academic purposes have an important gate-keeping function, namely, to provide information about whether prospective students have achieved a linguistic threshold necessary to cope with tertiary level education conducted in English. To increase the transparency of test results, it is essential that such language tests provide evidence of ability in language skills that resemble the skills people actually use in real-life academic contexts. The GEPT Advanced Writing test is a test of proficiency in English for Academic Purposes and is designed to simulate academic reading-into-writing tasks. Candidates are expected to extract information from the input provided in the test paper and transform what they have read into written arguments which can effectively meet the expectations of their audience. This paper will report on an empirical study which investigated the context and cognitive validity of GEPT Advanced Writing Task 1. The study aimed to investigate the extent to which the GEPT Advanced writing task reflects the ‘situational and interactional authenticities’ (Bachman and Palmer, 1996) of real-life academic writing tasks. To investigate context validity, the study applied both expert judgement and automated textual analysis to examine the degree of correspondence between both the overall task setting and input text features of the GEPT task and those of target academic writing tasks done in real life university courses in the UK. To investigate cognitive validity, the study examined, by means of a cognitive process questionnaire, the cognitive processes elicited by the GEPT task as compared with the real-life academic writing tasks. Data were collected from 160 students who were studying at a British University and from 192 test takers under live test conditions in Taiwan. The results showed that the GEPT writing task demonstrated both context and cognitive validity: the overall task setting of the GEPT Advanced test task resembled those of real-life tasks that university students would normally encounter, and the test tasks elicited most cognitive processes which were elicited by real-life writing tasks in a UK academic context. The researchers will present the study and discuss how the findings of this validation study can further improve the transparency of test results of GEPT Advanced Writing Test Task 1.

Parallel Session 12.2: Thursday, June 5, 14:35
Presenter: Troy Cox
Affiliation: Brigham Young University
Title: Assessing language proficiency: Effect of speaking test method on oral proficiency exam scores
This presentation will examine the effect of test method on foreign language speaking proficiency test scores on the Oral Proficiency Interview-Computer (OPIc) versus the standard Oral Proficiency Interview (OPI). Additionally, this presentation shows how the scores are affected by test preparation and the context in which the language was learned. While studies have been done to rate the validity of the OPI and OPIc independently, a limited amount of research has been done to compare the inter-exam reliability of the two exams that purport to measure the same levels of proficiency (Surface et al., 2008; SWA Consulting Inc., 2009; Magnan, 1987). Although ACTFL and other individuals have done studies looking at the reliability and validity of the OPIc with learners of English, studies have yet to be conducted that compare the results of Spanish language learners who take both types of exams and how the format of the exam may influence the ranking of the students’ proficiency (Surface et al., 2008; Surface et al., 2010). This research fills that very important gap and provides empirical data to determine not only if equal measurement of proficiency exists across exams for learners of Spanish but also if the familiarity with the OPI and manner in which a language was learned affect the overall proficiency results on both exams. For this study, 125 Spanish language learners of various proficiency levels were divided into two groups and were administered both the OPI and an OPIc. To control for an order effect, a counter-balanced design was used in which one group took started with the OPI and the other with the OPIc. Prior to the administration of the tests, learners were given a survey about their language learning background, their familiarity with the tests and what test preparation steps they had completed. After the tests, the learners were given a survey about the strategies they used when responding and their feelings about the different methods. To examine the effect of the test method, a mixed model ANOVA will be reported. To explore the relationship between the learning environments, test familiarity, test scores, correlations will be reported. These results are important in being able to understand how these factors can influence learners at all levels and particularly learners of Spanish. Future research needs to consider how other languages and learners may perform on the OPI and OPIc given the differences found in this study. Since these exams are used to make high stakes decisions in business, education, and government, it is important to understand how students perform on the different assessments especially given the weight of these measures.

Parallel Session 12.3: Thursday, June 5, 14:35

Presenter: Youngsoon So and Ikkyu Choi

Affiliation: Educational Testing Service

Title: Influences of the difficulty of a listening stimulus on integrated speaking task performance

Successfully completing an integrated task involves multiple skills. For example, in a Listening-Speaking (L-S) integrated task, test-takers should comprehend a listening stimulus and then produce a spoken response based on the comprehension. However, the complex relationship between these required skills is not fully understood. In particular, the impact of comprehension on the quality of subsequent language production has not been thoroughly studied. This knowledge limitation is partly due to the fact that most previous studies on integrated tasks have investigated performances based on fixed stimulus conditions. Consequently, there has been minimal examination into the extent to which the performance is influenced when the difficulty of the stimulus changes. To address this research gap, this study has investigated the impact of listening comprehension on the quality of spoken responses in L-S tasks, through a direct manipulation of the difficulty of listening stimuli. The specific research question addressed is “To what extent does the difficulty of a listening stimulus influence integrated task performance on two dimensions – language and content?” A total of nine listening stimuli were developed by creating three different versions for each of three topics. The three different versions were operationally defined to include differing degrees of elaboration within each stimulus. The elaboration mostly involved repetition and reiteration of points discussed in a stimulus, and therefore, it was expected that the difficulty of a stimulus would increase with a lower degree of elaboration. Next, three
forms consisting of three L-S tasks were produced by combining different versions of the three topics. Then, each of the three forms was administered to 100 TOEFL iBT test-takers; therefore, a total of 900 responses (i.e., three forms × three tasks per form × 100 test-takers) were collected. The collected responses were scored separately on language and content dimensions. The language score was given based on three features including delivery, language use, and cohesiveness. On the other hand, the accuracy and completeness of the information from the given stimulus were the criteria for scoring the content dimension, such that the resulting score can represent a test-taker’s comprehension of the stimulus. Two separate analyses of variance were conducted with different stimulus versions as independent variables and each of the two scores (i.e., language and content) as a dependent variable in order to examine whether the test-taker scores varied across different versions. The results indicated that the content scores were influenced more by the particular listening stimuli administered than the language scores. Specifically, it was found that test-takers received significantly lower content scores on L-S tasks with the less-elaborate versions. This implies that the characteristics of a listening stimulus could have a non-negligible influence on comprehension. It further suggests a potential for a scoring model that separates listening comprehension from speaking, particularly when the scores from L-S tasks are interpreted as indicators of speaking.

Parallel Session 12.4: Thursday, June 5, 14:35

Presenter: Rebecca Present-Thomas

Affiliation: VU University Amsterdam

Title: Enriching the Framework for academic writing: a corpus-based analysis of developmental linguistic features

Despite its wide international adoption as a common tool for describing language proficiency, the CEF is lacking in language-specific descriptive power, particularly with respect to higher proficiency levels and written skills in general. Academic writing skills, which are particularly under-represented in the CEF, are increasingly included in large-scale assessments worldwide, where they often contribute to high-stakes decisions such as access to higher education. Within the functionally-oriented Framework, this study aims to identify linguistic (lexical, syntactic, and cohesion/coherency) patterns characteristic of written academic English text at increasingly higher CEF levels. How do the academic writing skills of English language learners in tertiary educational settings develop over time? Are productive lexical features, such as type/token ratio, mean word length, or use of academic words, good indicators of academic proficiency? Which syntactic features or frames best represent the academic genre, and how (and when) do learners acquire them? Which features do proficient academic writers use to promote cohesion and coherence? In order to shed light on these research questions, a collection of academic texts representing a range of degrees of academic proficiency was assembled and analyzed quantitatively using various computational algorithms. First, a corpus of essays written by bachelor students of English from a Dutch university was collected longitudinally over a period of 3 years; the student corpus contained over 600,000 words from 1400 essays written by more than 300 students. Each essay was rated by trained raters according to the CEF levels and the corpus was split accordingly into sub-corpora representing levels B1, B2, and C1. In order to better represent the highest level of the CEF, which was only sparsely represented in the student corpus, a supplemental corpus of published academic texts (assumed to be representative of level C2) was compiled from the section J “learned” sub-corpora of the ICAME written corpora (over 1 million words from more than 500 authors representing various regional dialects of English). For each of the four CEF level-based (sub-)corpora, several features of lexical makeup, syntactic complexity, and coherence/cohesion are being investigated, the findings compared to existing functional, language-independent reference level descriptors, and new English-specific, developmental linguistic descriptors proposed. Initial findings based on a subset of the student texts (Present-Thomas et al., 2013) suggest that
more proficient writers rely less on the most frequently used English words and more on words from the Academic Word List (Coxhead, 2000). They also use longer clauses, more coordination, and more complex nominals. The published texts show an even further developmental stage. They are clearly distinct from the learner texts, both confirming these patterns, and highlighting previously insignificant trends in the student data: longer sentences, fewer clauses per sentence, and less subordination being characteristic of more proficient texts.

**Parallel Session 13.1:** Thursday, June 5, 15:10

**Presenters:** Daniel Elias and Jane Lockwood

**Affiliation:** Future Perfect

**Title:** ‘Weak’ and ‘strong’ language for specific purpose (LSP) performance assessments: a case study in Asian call centres

LSP performance assessment has become a special focus for language testers in recent years where experts have debated how testing tools and processes can be strengthened to more accurately and more validly assess professional communication at work. Suggestions to achieve this include ethnographic studies of the target language situation; authentic discourse analyses and subject matter experts (SMEs) being informants on what they view as successful communication at work. These are all proposed as solutions in building LSP performance assessment validity (see for example Jacoby & McNamara, 1999). Language testing researchers have also suggested there may be some merit in distinguishing LSP performance as ‘weak’ and ‘strong’ (see also McNamara, 1996; Douglas, 2001), although clear lines of distinction between these have not yet been made in the studies to date. We propose a distinction between a ‘weak’ and ‘strong’ version of LSP performance assessment based on the empirical data collected over the last ten years of embedding the Business Performance Language Assessment Scales (BUPLAS) into Asian and South Asian call centres. A distinction that has been generally unclear to date is whether workplace LSP spoken assessments FOR work (meaning those assessments that gauge entry levels and a ‘predictive’ success for work) is the same as workplace LSP spoken assessments AT work (meaning those assessments that gauge quality levels on-the-job and an ‘observed’ success at work). For the Asian call centre industry this distinction is important because the purpose of recruitment (exclusion) is different from the purpose of quality assurance (appraisal and coaching feedback). We propose that this distinction may suggest a number of key characteristics for a ‘weak’ versus a ‘strong’ LSP spoken performance definition in this context. We argue that the essential characteristics of a ‘weak’ version of LSP performance assessment at recruitment, is where summative scores are generated when assessing candidates on the phone interview. The phone assessors are typically communication trainers with language assessment training. In contrast to this however, we argue a ‘strong’ version of LSP performance assessment occurs on the call centre floor where direct observations are made by quality assurance personnel of real or recorded call exchanges on the phones. We explore how LSP assessment in each of these recruitment and quality assurance sites is carried out and how the specific needs in these two sites impact the number, the selection, the description and the weightings of the criteria used for LSP spoken performance assessment in this industry. The development of BUPLAS in the call centre context extends and relates to other studies LSP performance assessment (see for example, work with health professionals Elder, Pill, Woodward-Kron, McNamara, Manias, Webb & McColl, 2011); and aviation professionals (Alderson, 2011) and perhaps opens up a further research issue as to whether a ‘universal framework’ for LSP spoken performance is more applicable to a ‘weak’ definition as described in this paper.

**Parallel Session 13.2:** Thursday, June 5, 15:10

**Presenter:** Jessica R W Wu and Chia-Lung Lee
Affiliation: Language Training and Testing Center

Title: Development and Validation of the Self-Assessment Scales for Taiwanese EFL Learners

This paper reports the development and validation of the self-assessment scales based on the GEPT framework. The GEPT is a five-level criterion-referenced EFL testing system widely implemented in Taiwan to assess general English proficiency of EFL learners at all levels. More than half a million Taiwanese learners take the GEPT each year, and GEPT scores are used in both educational and professional domains. In common with most self-assessment scales, the GEPT self-assessment scales are intended to encourage autonomous learning by giving learners greater control over their learning and enhancing their awareness of their learning and proficiency (e.g., Bachman, 2000; Harris, 1997; Oscarson, 1989). They also allow learners to estimate their ability by using their self-assessment results before registering for a GEPT level which best matches their ability.

To develop the self-assessment scales, the research first produced can-do statements for each language skill, each of which was matched against a GEPT level using expert judgment (Ross, 1998; 2006). Furthermore, the Rasch model was used to calibrate the can-do statements on a common scale of difficulty. Both types of evidence suggest that the can-do statements optimally correspond to the GEPT levels. Subsequently, the research investigated empirically the relationship between learners’ ability estimated by their self-assessment results and their ability assessed by their actual GEPT test results (Alderson & Huhta, 2005). Statistical analysis by Rasch and Ordinal Logistic Regression was based on a sample of 10,000 learners who were invited to take part in the GEPT and to respond to the self-assessment statements. Results show that the self-assessment scales have achieved acceptable accuracy in estimating learners’ GEPT levels, ranging from 65% to 78% across language skills. The self-assessment scales, accompanied by practice tests and feedback, have been provided as a free online learning resource accessible to Taiwanese EFL learners. User feedback on the resource was collected to inform the improvement of the self-assessment tool.

Parallel Session 13.3: Thursday, June 5, 15:10

Presenter: Mikyung Kim Wolf and Jane Shore

Affiliation: ETS

Title: Developing and Validating an Academic Reading Assessment to Improve Teaching and Learning for English Learners

In the context of K-12 public schools in the United States, the annual, standardized summative assessments are criticized for their limited usefulness. Conversely, there is an increasing need for sound assessments to inform teaching and learning toward the goal of improved student achievement. In the case of English learner (EL) students who are still developing English skills to access content learning, teachers need a tool to both systematically gather evidence of their students’ linguistic progress and support English language development for successful participation in academic contexts. The assessment tool developed by this government-funded research project facilitates the assessment of students’ diverse stages of learning at regular intervals. This is particularly challenging when considering the development of academic literacy in light of the ever-increasing complexity of academic texts encountered in higher grades. To address the need for improved methods of supporting EL students’ academic reading comprehension development in the middle grades, this research project takes a formative assessment approach. We have developed a tool called English Learner Formative Assessment (ELFA), consisting of a set of reading assessment forms for classroom use, a Teacher Guide, a teacher observation protocol, and a score management program. Focusing on persuasive genre comprehension, the forms measure three reading sub-skills: (1) prerequisite reading skills (e.g., interpreting sentence meanings using grammatical competence, recognizing lexical structures), (2) comprehending literal meaning (e.g., details and the main idea), and (3) performing higher-order reading tasks (e.g., synthesizing
information across texts, outlining an author’s argument). These sub-skills were confirmed to be critical for ELs by the teachers who participated in a focus group and pilot studies. This construct definition was also well-integrated into the diverse standards and curricula across schools. The Teacher Guide and observation protocol materials were designed to facilitate teachers in using probing questions, collecting learning evidence, and conducting instruction in real-time while using the ELFA assessment. In this session, we will present the findings and lessons learned from the pilot and usability studies. Approximately 160 EL students from 7 ESL classes and their teachers participated in the pilot study. Additionally, two teachers and their four classes participated in the usability study where the researchers observed how the teachers utilized the ELFA tools over two weeks. Students’ scores, observation notes, teacher surveys, and teacher interviews were analyzed both qualitatively and quantitatively. The data provided empirical evidence of ELs’ reading comprehension difficulties as well as insights into the effective use of assessments for formative purposes. Preliminary results suggest that both foundational and higher-level reading comprehension skills are important constructs to be measured for intermediate ELs. It was found that ELs who completed some higher-order reading skill items correctly did not necessarily answer foundational reading items correctly. It was even observed that two teachers used the assessments differently, indicating the pivotal role of teachers in the use of assessments for formative purposes. One of the implications from the results was the need for professional development on assessment literacy and formative assessment for the success of the ELFA tool.

Parallel Session 13.4: Thursday, June 5, 15:10

Presenter: Mehdi Riazi and Jill Murray

Affiliation: Macquarie University

Title: Modeling real-life academic writing: Implications for assessing writing proficiency

The inclusion of direct assessment of L2 writing in large-scale high-stakes English language proficiency tests such as IELTS, TOEFL, and Pearson is considered to be a big leap forward in the assessment of productive skills. Notwithstanding this breakthrough, the construct validity of the writing modules of these high-stakes proficiency tests remains the subject of substantial research. A main concern for different stakeholders is the extent to which successful performance on the writing module of these proficiency tests can predict test takers’ real life academic writing performance. This qualitative study, which is part of a larger scale research project funded by the Educational Testing Service (ETS), reports on postgraduate students’ description of the processes they went through and the strategies they used to write their academic assignments. Data were collected from 17 postgraduate students studying at four major universities in New South Wales, Australia in 2011 and 2012. Participants were interviewed about an academic task they had undertaken for an enrolled course. The stimulated recall interviews followed a semi-structured format in which participants’ academic assignments were used to trigger recollections of their experiences of composing each text in terms of the processes and strategies. Each participant provided the researchers with a copy of an assignment they had submitted as one of the requirements of one of their academic courses. All the interviews were transcribed by a professional company and were uploaded to NVivo. The whole dataset was then coded using a coding scheme which was developed from the data and related literature in writing research. The coding scheme included three broad categories of cognitive, metacognitive, and social/affective each with its own sub-category codes. Matrix queries were run in NVivo to investigate the processes and strategies participants used to complete their academic assignments. Results show interesting findings in terms of the macro processes of “planning”, “composing”, and “revising” and the specific strategies these participants used to complete their academic assignments. Accordingly, a revised version of the Hayes and Flower (1980, 1983) model of writing will be presented along with implications for the construct of “academic writing” and discussion of the assessment of writing proficiency. References Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of
A Survey of Models for Language Performance Tasks Scored by Multiple Raters

Abstracts June 5


Parallel Session 14.1: Thursday, June 5, 16:10

Presenter: Ikkyu Choi

Affiliation: Educational Testing Service

Title: A Survey of Models for Language Performance Tasks Scored by Multiple Raters

A Survey of Models for Language Performance Tasks Scored by Multiple Raters Language performance assessments typically consist of multiple tasks to properly represent their target domain and employ multiple raters to ensure accurate scoring (Luoma, 2004; Weigle, 2002). Notwithstanding the wide use of multi-task, multi-rater scoring procedures, they yield data that pose challenges to test taker proficiency estimation, including residual dependencies among multiple scores and rater effects. Many-facet Rasch measurement models (MFRM; Linacre, 1989) can provide estimates of test taker proficiency and rater effects, and therefore, have often been employed to analyze language performance assessment data. However, it has been pointed out that MFRM could overestimate the precision of the test taker proficiency estimates due to its simplified assumption for rating data structures (Bock, Brennan, & Muraki, 2002; Patz, 1996). While several new models have been proposed to fully appreciate the complex data structure from multi-task, multi-rater scoring procedures, the technical nature of these models has hindered their widespread use. This study attempts to bridge this gap by presenting a survey of models proposed to handle data resulting from multi-task, multi-rater scoring procedures, which are often collectively called rater models. In particular, this study provides a brief summary of the characteristics of several rater models, and compares them in terms of structural assumptions, parameters, and estimation results. Rater models presented in this study include the rater bundle model (Wilson & Hoskens, 2001), the hierarchical rater model (Patz, Junker, Johnson, & Mariano, 2002), and the second-order graded response model (Choi, 2013), as well as approaches proposed by Verhelst and Verstralen (2001) and Maris and Bechger (2007). The presentation of these models is facilitated by a motivating example of multi-task, multi-rater data taken from an English oral proficiency test that is used for ITA screening at a large U.S. university. Factor-analytic diagrams of the presented rater models are provided to illustrate their structural assumptions. These models are compared as different approaches to conceptualizing rating procedures involved in the motivating example. It shows that the presented rater models differ in terms of the following dimensions: the unit of analysis and the assumption of “error-free” rating. The models are then applied to the data from the motivating example. The parameter estimates from the different models are compared. The results show that the models yielded comparable estimates for test taker proficiency and rater effects. This implies that the choice among the presented models could be made depending on the degree of correspondence between the model assumption and a given testing situation, as well as practical concerns such as estimation time and available software. This study concludes with a list of remaining challenges and potential extensions.

Parallel Session 14.2: Thursday, June 5, 16:10

Presenter: Cecilie Hamnes Carlsen and Ann-Kristin Helland Gujord

Affiliation: Vox, Norwegian Agency for Lifelong Learning

Title: External and internal variables affecting language learning and language test scores.
External and internal variables affecting language learning and language test scores. Adult language learning is characterized by variation both in terms of rate of acquisition and in terms of the final level of proficiency obtained. Such variation is caused by internal factors, such as age, language aptitude, sex, L1, and motivation, as well as external factors, such as the amount and quality of instruction, the amount and quality of the input and time of residency in the target language community (Ellis 1994, Sanz 2005, Dörnyei 2005). The variation among L2 learners manifests itself in the scores obtained at language tests. How various factors influence the learning process and learning outcome is a question of interest to both SLA-researchers and to language test researchers. This research paper presents the aim, design, results and implications of a collaborative study between both. The aim of the study is to identify those salient background variables that influence a learner’s success or lack of success in an L2 as measured by two different CEFR-based language tests of Norwegian as a second language (levels A2 and B1). Many prior studies of external and internal factors focus on one variable at the time (simple correlations) and include small numbers of informants. The present study was designed specifically to overcome these two common sources of error. It includes as many as 10 different internal and external factors and uses regression analysis to estimate the relation between the dependent variable (test score) and the distinct different independent variables (internal and external factors). This allows us to identify which factors have the greatest impact on test scores and therefore warrant further investigation. Since the current study is based on the data of 15 000 informants, it overcomes the problems that come with small sample sizes and enables us to see clear patterns. In the paper we will present the results of the study and we will show what factors appear to explain most of the variation in language learning success and hence the chance for a given candidate or group of candidates to pass the two test in question. By identifying salient internal and external factors, this study may help refine frameworks such as the CEFR by supplementing it with data that indicate what can realistically be expected of a learner at a certain level given his/her background variables. The current study bring new insight into the role of learner-related difference in second language learning, hence the paper briefly reflects upon the advantages of interdisciplinary collaboration between SLA-researchers and language test researchers.

Parallel Session 14.3: Thursday, June 5, 16:10

Presenter: Chihiro Inoue

Affiliation: CRELLA, University of Bedfordshire

Title: Measuring spoken language: A comparative study on general variables of syntactic complexity and accuracy

Variables of complexity, accuracy and fluency have been extensively used in investigating learner spoken language, especially in validation projects where level descriptors are closely examined against specifications and the elicited learner performance. However, a serious concern is that some variables are used conventionally without empirical justifications. It is crucial for language testers to know how results might be different if different variables are used, and choose them adequately to suit their purposes; otherwise, the results may not be as informative as they wish. Some studies have been conducted on variables of fluency (e.g. Kormos & Dénes, 2004) and of lexical complexity (e.g. Jarvis, 2002), but there is little empirical research on variables of syntactic complexity and of accuracy. This is the research gap that this paper aims to fill, by examining them using the spoken performances by 65 Japanese EFL students on two picture-based narrative tasks, which were rated as CEFR A2 to B2 according to Rasch-adjusted ratings by 7 human judges. The tasks were very similar, but had different degrees of task-essentialness (Loschky & Vle, 1993) for subordinate clauses. The first strand of this paper examines if the suggestions on variables of syntactic complexity (in writing) by Norris and Ortega (2009) applies to spoken performance. They suggest that the amount of coordination may be suitable for beginner level learners, the amount of subordination for
intermediate, and the mean length of clauses for advanced levels. The research questions here are: RQ1-1. How do the variables of syntactic complexity correlate with one another on the two tasks? RQ1-2. Does the discriminant analysis show that the amount of coordination, the amount of subordination, and mean length of clauses best predict the beginner, intermediate, and advanced levels respectively? If not, why? The second strand of this paper examines different variables of accuracy with a view to find out which variables may be valid, which researchers seem to disagree. The variables of accuracy in question are: the percentage of error-free clauses, the number of errors per AS-unit and the number of errors per 100 words, and research questions are: RQ2-1. How do the variables of accuracy correlate with the ratings of accuracy of the spoken performances on both tasks? RQ2-2. Do the t-tests on the variables of accuracy reveal the same results as on the ratings on both tasks? If not, why? It was found that, on both tasks, the variables of syntactic complexity correlated highly with one another (RQ1-1), and discriminant analysis showed inconsistent results with the suggestions by Norris and Ortega (RQ1-2). For accuracy, all the variables correlated very highly with the ratings on both tasks (RQ2-1), but only the errors per 100 words was in line with the human ratings in capturing the differences of the performances on the two tasks (RQ2-2). Analysis of transcripts revealed that these results were strongly influenced by the differing degrees of task-essentialness for subordination of the two tasks, which warrants scrutinisation of the characteristics of test tasks and careful pilotng of different variables accordingly.

Parallel Session 14.4: Thursday, June 5, 16:10

Presenter: Michelle Raquel and Jane Lockwood

Affiliation: The Hong Kong Polytechnic University

Title: Exploring the use of an automated essay scoring system for a diagnostic writing test

As part of a larger English language assessment being introduced to Hong Kong universities, a Diagnostic English Language Tracking Assessment (DELTA) writing component is being developed. DELTA Writing will need to (1) assess the academic writing skills of first year HK tertiary students across multiple disciplines; (2) be administered to large numbers of students across all Hong Kong universities and report results very quickly; (3) provide diagnostic feedback (i.e., report on strengths and weaknesses); and (4), enable students to track their own progress over time. To fulfill the diagnostic and tracking objectives, the DELTA writing component needs to produce fine-grained multiple-trait score profiles for each student, and to do so with enough reliability to make tracking over time feasible. Also, as students progress through their studies, the DELTA Writing will need to offer flexible elements to account for the growing demands of academic literacy placed on them, and for their varied disciplinary experiences. To address this complex set of demands, the DELTA writing research team conducted a three-phase study that (a) investigated the writing needs of first year HK tertiary students; (b) explored how a ‘genre-based’ approach might frame the assessment of academic writing; and (c) investigated the suitability of automated essay scoring (AES) systems to rate Hong Kong first year undergraduate academic essays. This paper will report the third phase of the research project. The recent growth of interest in the use of machine scoring of essays, and the related development of new AES systems, has encouraged the DELTA Writing project team to explore the feasibility and potential value of using automated essay scoring (AES) as a solution or partial solution. The project team explored the capability of AES systems to provide detailed domain scoring of the writing of large numbers of undergraduates, and provide rapid, reliable and valid diagnostic reports. Six AES systems were evaluated and three were invited to participate in a machine scoring trial. One method of evaluating AES systems is to check for high agreement of human scores with AES scores. Five hundred first-year HK tertiary undergraduate students were asked to write a 450-word academic essay in response to a reading-into-writing prompt. A pilot rater training with the DELTA Writing rubric was carried out with a team of twelve English language teachers from Hong Kong universities. Raters took part in standardization and then marked the scripts. The 500 essays and the in-house scores were provided to the
Abstracts June 5

three AES systems that participated in the study. The automated scores were then compared with the human scores. This paper will report on how human raters and AES systems scored these 500 scripts. It will then discuss the strengths and limitations of each AES system in scoring Hong Kong first year undergraduate academic essays. Finally, the feasibility of using an AES system as part of the DELTA writing component will be discussed.

**Plenary Session 4: Thursday, June 5, 16:45**

**Presenter:** Talia Isaacs and Pavel Trofimovich

**Affiliation:** University of Bristol

**Title:** Modelling comprehensibility in an oral production scale for L2 learners of English: Which linguistic factors generalize across L1s?

Pronunciation can be considered the skill most amenable to diagnostic assessment. Indeed, pronunciation-focused discrete-point items were prominently featured in Lado’s Language Testing (1961), designed to assess “language problems” based on differences between first and second language (L1-L2) inventories. Modern approaches to pronunciation assessment also demonstrate how systematic testing can pinpoint learner perception/production errors (e.g., Celce-Murcia et al., 2010). However, pronunciation is also arguably the most difficult skill to model in diagnostic scales designed for learners from diverse L1 backgrounds. This is, in part, because differences between learner productions often occur due to L1 transfer errors, with L1 effects on L2 speech tending to be more perceptually salient to listeners for pronunciation than for other skills (e.g., grammar, lexis). Thus, pronunciation contributes to listener perceptions of speech even if pronunciation isn’t explicitly what is being measured, making it difficult to specify, in rating descriptors, the phonological features that are universally applicable to learners from different L1 backgrounds. To complicate matters further, not all pronunciation errors “count” the same in pedagogical terms, with some errors being more detrimental to comprehensibility, or listeners’ ability to understand the L2 speech, than others (Derwing & Munro, 2009). The growing consensus among applied linguists is that the linguistic features most likely to impact comprehensibility should be emphasized in L2 instruction, as in rating descriptors, and that the features that contribute to an L2 accent but that are inconsequential for comprehensibility should be left aside (Authors, XXXX). The Common European Framework of Reference (CEFR) demonstrates problems inherent in modelling pronunciation in a common scale. The mandate of including language-generic features in the descriptors—so that the scale can be used for any (European) target language—is prohibitive for generating descriptors that are specific enough to isolate the source of individual learners’ pronunciation difficulties. Pronunciation is thus not included as a criterion in the global CEFR scales, largely due to erratic statistical modelling of pronunciation indicators during scale development (North, 2000). The CEFR Phonological Control scale assumes that as accent decreases, L2 speech will become more comprehensible, conflating these interrelated but distinct dimensions (Harding, 2013). Such shortcomings in modelling pronunciation are replicated in other commonly-used L2 speaking scales. In light of these gaps, this presentation reports on the development of a pedagogically-oriented L2 comprehensibility scale designed for use on university campuses to guide teachers’ identification of the linguistic components most conducive to learners’ production of comprehensible English so that these can be targeted in instruction. A synthesis of findings from Authors’ empirical studies (based on instrumental/auditory analyses of L2 speech; statistical analyses of ratings; and content analyses of focus group debriefings and introspective reports) shape the evolution of a “crude” empirically-derived 3-level global comprehensibility scale to a 6-level analytic scale, with a focus of the generalizability of linguistic criteria across L1 background and task type. Overall, comprehensibility cuts across a wider range of linguistic domains than previously expected, with pronunciation and lexicogrammatical dimensions differentially contributing to comprehensibility ratings as a function of L1 background.
Beyond the limits of universal criteria: Using indigenous criteria in assessing clinical communication in ESL

The notion of indigenous assessment (Jacoby, 1998; Jacoby & McNamara, 1999) assumes that criteria for judging performance may be specific to a particular setting, and that using those criteria rather than more general, universal criteria may lead to different decisions about the relevant communicative abilities of individuals in that setting. Dissatisfaction with decisions about the readiness of individual test takers to enter clinical practice safely based on performance on the Occupational English Test (OET), a four-skills test of English as a second language for communication in clinical settings, was the trigger for the three-year study reported in this paper of the criteria used to judge performance in the OET speaking sub-test. While the content of the OET speaking sub-test role-play materials reflects the communicative tasks with patients that health professionals are likely to face in the work setting, the criteria by which performance on these tasks is judged currently reflect general criteria for evaluating language proficiency: Intelligibility, Fluency, Appropriateness of Language, and Resources of Grammar and Expression. It was therefore decided to investigate to what extent workplace practitioners when evaluating the performance on such tasks in the work setting valued these criteria, and whether there were additional criteria to which they were oriented. The data for the study consisted of feedback by clinical supervisors to trainees in medicine, nursing and physiotherapy on their communication with patients. The settings included live, real-time feedback in the case of physiotherapy, captured on the ward, and simulated feedback on the performance of trainees in all three professions using video recordings of trainee–patient interaction. Thematic analysis of the feedback yielded a rich, bottom-up, empirically grounded model of what supervisors value in interactions with patients in clinical settings. The integration of clinical knowledge and communication skill featured strongly. While supervisors’ feedback showed that they did indeed value the features of performance reflected in the current criteria, they placed more emphasis onto the clinician’s engagement with the patient and the ability to manage the interaction, each implemented through language and therefore assessable in a language test. A checklist derived from supervisors’ comments was used to orient existing raters to the new criteria (‘Clinician Engagement’ and ‘Management of Interaction’) and descriptors, with raters then judging a sample of current OET performances using both the existing and the supplementary criteria. Rasch analysis of the resulting data showed the new criteria had good psychometric properties and compatibility with existing criteria. In a complementary standard setting exercise with clinical supervisors, grades awarded to individuals judged under the existing and proposed new OET criteria were compared. While some individuals were now classified as not ready for clinical work, others were judged more highly using the new criteria, suggesting that context-specific criteria yielded different and arguably more valid judgements. Overall, the study shows that universal, general scales and accompanying criteria are inappropriate for specific purpose contexts, particularly where the stakes are high – in this case, in terms of patient safety and clinician employability.
Abstracts June 6

Samuel J. Messick Memorial Lecture Award

Keynote 3: Friday, June 6, 9:15

Presenter: Denny Borsboom

Affiliation: University of Amsterdam

Title: Validity as the realist views it

I will argue for a realist framework that grounds validity in a causal analysis of measurement: a test is valid if and only if variation in the test scores is causally dependent on variation in the targeted attribute. If no such attribute exists, or variation in test scores is caused by some other attribute, then the test in question is invalid. I will show that this conception nicely fits paradigm examples of measurement across the sciences, and also pinpoints the problematic status of measurement in psychology: validity problems arise for a simple yet informative reason, namely that we do not understand what psychological attributes are or how they could cause variation in test scores. I will sketch some ways forward in advancing our understanding of validity in standard cases of psychological testing, but will also consider the possibility that the measurement framework is simply inappropriate for thinking about the construct-observation relation in psychology.

Denny Borsboom:

I am professor at the Psychology Department of the University of Amsterdam. My research is on the methodological foundations of psychology – in particular, on conceptual problems in psychological measurement. My current efforts are mainly directed to the construction of network models for psychopathology. Summary of activities: In the educational programme of the Department of Psychology, I teach an undergraduate course (Foundations of Psychology) and participate in two Bachelor courses (Introduction to Measurement and Network Analysis). In addition I have taught Research Master courses (Philosophy of Psychology) and postdoctoral courses (What is Psychometrics?). In organizational work, I am involved in developing the curriculum at various levels, from Bachelor to postdoctoral programmes. I currently oversee the implementation of a tutor group system in the second year of our curriculum. Service to the field has included membership of the Psychometric Society Board of Trustees; Expert Panel membership for the Flemish Organization for Scientific Research (FWO); and Board membership for the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS). Most of my current research is devoted to the Psychosystems project (http://www.psychosystems.org), which gravitates around the development of new methodologies for psychological research using complex systems theory and network models.
Plenary Session 6: Friday, June 6, 9:45

Presenter: James Purpura and Sara Woodson

Affiliation: Teachers College, Columbia University

Title: Tracking the development of learning outcomes in a learning-oriented approach to classroom-based language assessment

Many assessment researchers (e.g., Shohamy 1998; Turner, 2012; Hill & McNamara, 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. As a result, research in this area has examined: (1) teacher practices and processes in using L2 assessments (Leung & Teasdale, 1997; Rea-Dickins, 2003; Colby-Kelly & Turner, 2007); teacher assessment decision-making processes in using rating scales (Brindley, 1998); (3) the role of teacher knowledge, experience, and beliefs in planning and implementing assessments (Brindley, 2001, Gardner & Rea-Dickins, 2001, 2007); (4) the role of diagnostic or dynamic assessment in promoting teaching and learning (Alderson, 2005; Lantoff & Poehner, 2011); (5) the effects of standards-based and outcomes-based assessment on teaching, learning and policy (Davidson, 2007); and (6) the value of self and peer assessment for promoting self-regulation, autonomy, motivation, and learner outcomes (Patri, 2002; Saito, 2008). Aiming to further this discussion, Purpura and Turner (2013, Forthcoming) have proposed an approach to classroom-based assessment that, in prioritizing learning and learning processes, seeks to determine how well students have benefitted from assessment in narrowing achievement gaps. More specifically, this approach describes how planned and unplanned assessments are conceptualized and implemented from a learning perspective, as well as how planned assessments, together with those occurring spontaneously through social interaction, contribute to the advancement of L2 processing and the attainment of learning outcomes. In highlighting the integration of models of cognition, socio-cognition, and learning with L2 instruction and assessment, this approach, referred to as learning-oriented assessment (Purpura, 2004), is concerned with the contextual, cognitive, socio-cognitive, dispositional, and interactional dimensions that underlie the design, implementation, and use of assessments and their potential for facilitating learning. The current study uses a learning-oriented approach to investigate the nature of planned and unplanned assessments in an ESL classroom, and the role that these assessments played in learning the passive voice. The study examined specifically how the use of planned and unplanned assessments promoted learners’ L2 processing, and how the assessments, performed individually or collaboratively through social interaction, contributed (or not) to the ability of the learners’ ability to use the passive voice to describe operational processes (e.g., desalination). Three intermediate ESL classes were videotaped, using three cameras, 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 and transcribed. Instances of planned and unplanned assessments were then identified and examined iteratively and recursively through several lens (e.g., interactional features, proficiency features, processing and learning features). Learning patterns in the data were then tracked across the lessons and related to ultimate learning outcomes. The results showed that assessment, whether planned or unplanned, played a pervasive role in the teaching/learning process. The also showed a complex mix of patterns related to how spontaneous assessments may or may not contribute to the achievement of ultimate learning goals.
Parallel Session 15.1: Friday, June 6, 10:35

Presenter: Gina Park and Eunice Eunhee Jang

Affiliation: Ontario Institute for Studies in Education, University of Toronto

Title: Investigating text genre effects on young English language learners’ reading test performance using differential bundle functioning

Genre refers to a range of different writing styles, each of which differs by text purpose and structure (Littlefair, 1991). As young readers develop their literacy skills, they encounter various genres, such as narrative, expository, informational and procedural, which are crucial for developing literacy competence. Accordingly, most large-scale reading tests employ various types of genres. Considering that text genres reflect different community norms and cultural knowledge (Purcell-Gtes, Duke, & Martineau, 2007), English language learners (ELLs) may have difficulty with a certain text genre if they have insufficient experience and knowledge about it. This difficulty may influence ELLs’ test performance (Zheng, Cheng, & Klinger, 2007). The present study investigated the extent to which text genres cause reading test items to function differently across different language subgroups. The study employed a confirmatory multidimensionality-based differential bundle functioning (DBF) framework (Douglas, Roussos & Stout, 1996). It was hypothesized that item bundles with different text genres function differently among subgroups of students from different language backgrounds. We further conjectured that text genre was considered part of a primary dimension, and cultural knowledge associated with certain text genres (e.g., narrative, poetry) was considered a secondary dimension. Differential bundle functioning would occur if an item bundle were sensitive to both the primary (i.e., reading ability) and secondary dimension (i.e., cultural knowledge) against ELL groups in favor of non-ELLs. Three groups, including two focal groups and one reference group, were formed from a large-scale data using a provincial reading test for tracking Grade-6 students’ reading achievement. To test if cultural knowledge associated with certain text genres is the cause of DBF, we identified two focal groups, recent immigrant ELLs with their residence in Canada less than three years (n=1,425) and domestic ELLs who are born in Canada but speak mostly other languages at home (n=3,349). The DBF analyses confirmed our hypothesis that the item bundle with the narrative text genre exhibited a statistically significant DBF effect against both the focal groups in favor of the reference group. This finding confirms the effect of different text genres on students’ reading performance (Stampoltzis & Pumfrey, 2010) and further reveals that the narrative text genre presents ELLs with difficulty because it conveys meanings implied rather than explicitly stated and details that require a great deal of attention (Emmitt et al., 2003). Interestingly, the item bundle with the narrative genre functioned differentially against both the focal groups, suggesting that it is not the immigration status but home language environments that influence students’ cultural knowledge required for developing the knowledge of text genre. However, the item bundle with poetry did not show a statistically significant DBF effect. Considering its use of elliptical, metaphorical, and allusive language, this finding calls for further research on the effects of different text genres on ELLs’ reading performance and effective instructional strategies that helps ELLs develop the knowledge of text genres.

Parallel Session 15.2: Friday, June 6, 10:35

Presenter: Kellie Frost and Cathie Elder

Affiliation: The University of Melbourne

Title: Marching out of step: Task-based language competencies for military purposes and a general proficiency scale
This paper describes a recent project considering the feasibility of mapping the outcomes of a new, competency-based teaching and assessment framework to a pre-existing general proficiency scale. The project was undertaken as part of an external consultancy reviewing assessment practices at the Australian Defence Force School of Languages (DFSL), based in Victoria, Australia. Language courses at DFSL are designed to provide selected military personnel with the linguistic skills needed to perform specific military duties in locations around the world where the Australian Defence Force is deployed. A shift in the DFSL curriculum from general proficiency courses to task-based teaching and assessment aimed to target more efficiently the highly specific nature of the language demands faced by military personnel in contexts where languages other than English are used within the limited time frame available for teaching and learning before deployment. To satisfy existing administrative and reporting requirements, however, it remained necessary to map competency-based assessment outcomes against the established Australian Defence Language Proficiency Rating Scale (ADLPRS). A review of the claims underlying competency-based assessment on the one hand and proficiency scales on the other indicated that it might be difficult to generalise from performance on a particular assessment task to general descriptions of ability as characterised on the ADLPRS, or indeed on other generic scales adopted in Language for Specific Purposes (LSP) teaching contexts. To test this contention, two empirical studies were conducted comparing performance outcomes on (a) the end-of-course assessment associated with two competency-based courses (Tetum for Tactical Interaction and Indonesian for Military Communication) and (b) general proficiency tests designed to test at the level on the ADLPRS scale corresponding to the designated levels of the two courses. Results of this comparison revealed a misalignment between the two types of assessment, and highlighted complexities inherent in interpretations of assessment outcomes in this high-stakes specific-purpose context, regardless of the reporting framework adopted. The paper concludes that while the prospect of linking language competencies to general proficiency scales is bureaucratically appealing in that it offers a mechanism for mapping pathways between course offerings, any formal attempt to link the outcomes of context-specific competency-based assessments to more general proficiency scales, whether in military or other LSP contexts, is unlikely to be empirically defensible, and is therefore open to question in terms of its validity and reliability. The usefulness of reporting such ‘equivalencies’ is consequently limited to responding to administrative and policy practicalities rather than providing meaningful information about the language abilities of test takers.

Parallel Session 15.3: Friday, June 6, 10:35

Presenter: Daniel Isbell and Maria Nelly Gutierrez Arvizu

Affiliation: Northern Arizona University

Title: Is Equating Reasonable for Small-Scale Language Testing Programs?

Intensive language programs routinely administer placement and exit tests. A placement test in the beginning of a term is used to group students of similar ability levels into classes. An exit test at the end of the term is used to determine the degree to which students have achieved certain levels of language proficiency. These placement and exit tests are repeatedly administered, so different, yet parallel, forms of the test are developed to address issues of test familiarity and security. These parallel forms should be written to the same content and statistical specifications and should produce interchangeable scores, which will lead to increasing the transparency of test results. Providing interchangeable test scores is an important consideration for effective and confident placement decisions, and for helping learners accurately gauge progress. However, a question of fairness arises in respect to varying difficulty among forms. To address this question, a rationale and evidence should be provided for the interchangeability of these scores (Standard 4.10, AERA, APA, NCME, 1999). Large scale testing programs routinely use equating procedures to adjust for differences in form difficulty. Should small scale testing programs in language institutes also use equating procedures to ensure that scores from
different test administrations are interchangeable? The present study aimed to determine 1) whether or not equating introduces more error than it accounts for in a real-world, small-scale testing program and 2) how placement decisions in that setting were affected by test equating. Our study used non-IRT test equating with a non-equivalent anchor test (NEAT) design in a language testing program at an American university’s intensive English program. Data for the study consisted of two dichotomously scored sections (reading and listening) of two administrations (173 and 88 test-takers) of a placement test battery. Descriptive statistics and internal consistency reliability estimates were computed for each test section, including anchor sets, in order to explore six different equating methods for each section: identity, linear Tucker, linear Levine, circle-arc with a low point of 0, circle-arc with a low-point equal to the chance score, and pre-smoothed (to two moments) equipercentile (Kolen & Brennan, 2004; Livingston 1993; Livingston & Kim, 2009). To compare equating error, the standard error of equating (SEE), equating bias, and root mean square deviation (RMSD) are currently being calculated for each method. Scores of test takers in the second administration were adjusted to simulate alternative level placements in the intensive English program. Initial results indicate that using equating had a considerable effect on placement decisions. It is anticipated that the error introduced by equating will be smaller in magnitude than that of the change in difficulty between test forms. It should be noted that the existence of open-source equating software (the statistical program R and the package equate) lends to the practicality of test equating in small-scale programs. Due to its ability to model non-linear distributions across forms with small samples, the circle-arc equating method is preferred. Implications for other small-scale intensive language testing programs include equating’s effects on placement decisions.

Parallel Session 15.4: Friday, June 6, 10:35

Presenter: Sara Weigle and Eric Friginal

Affiliation: Georgia State University

Title: Comparing linguistic features of essay test writing with writing in the disciplines

One important validity question related to the assessment of writing for academic purposes is the degree to which writing on an impromptu essay test can be generalized to academic writing in different disciplines. One way of addressing this question is to look at the linguistic features of test essays as compared to successful student writing in content areas. Using Biber’s (1988) multidimensional analysis framework, our study compares the linguistic features of test essays written by native and non-native speakers with a corpus of successful student writing across a range of disciplines. Our study is based on a study by Author2 et al (forthcoming), who analyzed 829 papers in the Michigan Corpus of Upper-level Student Writing (MICUSP) (Author 2 et al., forthcoming), a corpus of papers written by advanced undergraduate and graduate students in a variety of disciplines. Four dimensions were found in the MICUSP data, and scores on these four dimensions differed across disciplinary areas (humanities & education, social sciences, biological and health sciences, physical sciences). Using these same four dimensions, we analyzed TOEFL iBT Independent essays on two prompts written by native (N=150) and non-native speakers (N=386). Our study reveals that the linguistic features of student writing on impromptu essays are similar in some respects to writing in humanities in terms of the four dimensions but differ in important ways from writing in other disciplines, particularly the sciences. Furthermore, results demonstrated that language background (i.e., native vs. non-native), essay prompt, and language proficiency (i.e., essay scores for the non-native speakers) were systematically related to scores on all four dimensions. Implications for writing pedagogy and language assessment are discussed.
Poster 1: Friday, June 6, 11:30

Presenter: Lyle Bachman and Adrian Palmer

Affiliation: UCLA

Title: Using an On-line Tool for Implementing an Assessment Use Argument-based Approach to Language Test Development


Poster 2: Friday, June 6, 11:30

Presenter: Maggie Dunlop and Maryam Wagner

Affiliation: Ontario Institute for Studies in Education, University of Toronto

Title: Validating a French for Academic Purposes placement test designed using cognitively diagnostic assessment principles

Language tests no longer need to only provide information about students’ proficiency for accountability purposes. Tests based on cognitive diagnostic assessment (CDA) principles have potential to provide rich information about students’ skill mastery across cognitive knowledge and skills. Tests developed according to these principles share several qualities: 1) the content of the test (and the basis for diagnosis) are based on learning goals specific to the instructional context; 2) test items are designed to elicit diagnostic information based on specified learning goals; 3) the scoring system affords the opportunity to diagnose fine-grained knowledge and skills; and 4) the feedback mechanism allows for maximizing diagnostic feedback (Author). There are multiple examples of the application of CDAs to existing tests (e.g., Li, 2011; von Davier, 2008), but currently no examples of tests developed along these principles exist. This poster reports on the development and validation process of a French as a second language (FSL) test developed according to CDA principles. The
computer-based test is intended to serve diagnostic and placement purposes by providing information about Anglophone university students’ French skills. We adopted an evidence-based design framework (Mislevy, Steinbery & Almond, 2003) which facilitated utilizing argument-based (Kane, 2006) validation approaches. The theoretical framework for assessing FSL for academic purposes was developed based on the systematic review of relevant literature and complemented through FSL classroom observations and curriculum review. The resulting framework operationalized the construct of FSL for AP into six domains of linguistic knowledge and cognitive processing skills in multi-modal discourses: Grammar; Vocabulary; Textual knowledge; Explicit Information Processing; Making Connections Using Implicit Information; and Critical Appraisal. A total of 54 test items were developed to cover all the essential knowledge and skills identified in a Q matrix (Tatsuoka, 1983). A Q matrix is a fundamental component of CDA-principled tests; they are used to make connections between skills tested and items. The core principles for designing items included: (1) fulfilling the diagnostic potential of the test by maximizing cognitive discrimination power, (2) enhancing test authenticity by using multi-modal tasks relevant to academic language functions, and (3) eliciting cognitively rich processes by maximizing the diagnostic potential of selected responses. A total of 157 students participated in a field test. Firstly, the items were checked to verify their appropriateness for a Q-matrix, which requires moderate difficulty and high discrimination indices. In classical testing theory analyses, 60 percent of p-values were in the .4-.7 range, and 91% were in the .3-.8 range. Discrimination indexes ranged from .22 to .67 for 89 percent of items. A 2PL item response model fitted best. The results indicated the majority of items performed adequately. Q-matrices also require test multidimensionality. DIMTEST (Stout, 1987) and hierarchical cluster analysis indicated the test was multidimensional. DETECT (Zhang and Stout, 1999) clearly identified five dimensions. Presently, large-scale implementation of the test is being organized to fit a cognitive diagnostic model (CDM) that provides test-takers with individualized diagnostic feedback. This test exemplifies an innovative development and validation process with much potential for teaching and learning.

Poster 3: Friday, June 6, 11:30

Presenter: Pablo Garcia Gomez and Anne Katz

Affiliation: Educational Testing Service

Title: English-for-Teaching: Reframing the assessment of teacher English proficiency for the classroom

English language proficiency for teachers has long been promoted as a key component in improving the quality of classroom instruction for EFL students across global contexts. The premise is that the higher the level of teacher language proficiency, the better the quality of instruction. As a result, a limited command of the English language is perceived by non-native English-speaking English language teachers around the world as one of the most important factors undermining their confidence within the classroom (Butler, 2004). Efforts to help teachers increase their overall English proficiency constitute a commendable goal; however, in addition to considerable practical challenges to this goal, we argue that general proficiency may not actually provide teachers with the immediate tools they need to gain more confidence in using English in their classrooms. Such approaches offer a generalized view of language and of teaching that often do not connect with what teachers are doing in the classroom on a daily basis. Indeed, focusing on the specific language and language skills needed to teach English in English has not typically been part of professional development programs (Sešek, 2007). This poster session will describe the design and development process of a professional development program that contains learning materials and assessments and that focuses specifically on the actual language and skills typically needed to prepare and conduct lessons in English. The poster will show how a combination of English for Specific Purposes (ESP) and Evidence Centered Design (ECD) frameworks helped to both define the English-for-Teaching construct (Freeman, et al. 2013) and to create innovative learning and assessments tasks. ESP approaches tend to focus on ensuring that the design of learning materials and assessments is centered on learners’ needs and that practical outcomes are emphasized (Douglas, 2000). ECD, on the other hand, provides an overarching framework that requires all components of an assessment program to have

Poster 4: Friday, June 6, 11:30

Presenter: Yuanyuan Guan and Alan Urmston

Affiliation: James Cook University

Title: The impact of educational background on the DIF of the DELTA listening component

A test item is considered to exhibit Differential Item Functioning (DIF) if it performs differently for different subgroups of test takers with equal ability. DIF results are an important indication of item bias and used as evidence for validity arguments and test fairness (ETS, 2002). Previous DIF studies have been conducted on learners of different genders, nationalities and disciplines. However, there has been scarce research on the impact of regional/educational backgrounds on DIF performance. This paper reports on a preliminary study of regional DIF analysis of the listening component of the Diagnostic English Language Tracking Assessment (DELTA) and exploration of possible causes for the DIF. The DELTA listening component assesses a range of listening sub-skills employed in understanding different spoken genres on a variety of topics. In the 2013-2014 academic year 276 text-based multiple-choice questions were administered to the incoming Year One students. The 2000 students primarily comprised of three sub-groups: Hong Kong locals graduating from Chinese-as-medium-of-instruction schools (CMI), Hong Kong locals from English-as-medium-of-instruction schools (EMI), and non-locals from Mainland China where Mandarin is used as the primary means of instruction (MMI). It is assumed that due to socio-economic statuses, linguistic and educational backgrounds in the two regions, e.g., exposure to the English environment and various language curricula, these three subgroups of students have undergone distinctly different experiences in English language learning, thus showing diverse performance in the test. By applying Rasch measurement techniques, students with similar DELTA measures were selected for DIF analysis. The analysis compared the item difficulty estimates for the three subgroups, identified potentially biased test items, and compared the item characteristics curves (ICCs) for items with indication of DIF. Based on the DIF results, content analysis was conducted by expert judges to identify possible sources of group difference. The findings show how the listening items performed differently on assessing the listening ability of the three subgroups. Analysis of these differences has enabled the researchers to look at the relationships between regional variation in examinee background and the abilities of the examinees to answer correctly different listening item types. These relationships will be discussed together with implications of this for item revision and instructional practice.
Poster 5: Friday, June 6, 11:30

Presenter: Katharina Kley

Affiliation: University of Iowa

Title: **Defining the construct of interactional competence in a German paired speaking test setting**

The purpose of this study is to describe the construct of interactional competence in more detail. The study is set in a low-stakes, classroom-based context targeted at intermediate-level learners of German, who are in their second year of German instruction at the college level. Within a paired test setting, it is investigated what variation in co-constructed discourse arises from different paired tasks (jigsaw and discussion task) and different speaking ability-level combinations in the test-taker dyads (same ability vs. weak/strong combinations). Paired speaking tests seem to be particularly suitable to elicit test takers’ interactional competence, a concept that originated in the field of SLA and that refers to the candidate in interaction with other language users from a dynamic and social perspective (Kramsch, 1986). Kramsch (1986) argues that foreign language proficiency should be understood in terms of language learners’ skills to collaborate with one another by negotiating meanings, providing and requesting clarifications, and anticipating the interlocutor’s response to reach a common meaning and understanding. To support this view of social or behavioral interaction, reference is often made to Sociocultural Theory, which is based on Vygotsky’s work. In the given assessment, interactional competence has been operationalized as a set of features or conversational moves (e.g., referring to the partner’s ideas, providing words, asking for clarification, etc.), which indicate collaboration or co-construction between two test takers. These features, as operationalized in this study, are used by the partners to negotiate meaning, to signal comprehension, and to encourage an exchange of opinions and information. Even though construct definition and validation studies of interactional competence have been conducted in paired test settings (Brooks, 2009; Ducasse, 2010; Ducasse & Brown, 2009; May, 2006, 2009, 2011; Taylor, 2001), the potential influence of contextual factors has been largely ignored. However, investigating how the contextual dimension affects co-constructed discourse between test takers is of importance as it provides a more detailed definition of the construct of interactional competence. The research questions for this study are as follows: 1. What are the range and frequency of interactional features produced in a discussion task and in a jigsaw task by pairs where both members of the pair are at the same, a different, and a very different level of speaking ability? 2. How do the jigsaw task and the discussion task relate regarding the type and frequency of interactional features co-constructed in each of the test-taker dyad combinations? The findings will contribute to our knowledge of interactional competence and how it is constructed under different testing conditions. In other words, the findings will provide insights on the impact of task type and test-taker ability combinations on the resulting performance of the test-taker pairs. Hence, the findings have immediate practical application in the design of paired oral tests in German language courses.

Poster 6: Friday, June 6, 11:30

Presenter: Benjamin Kremmel and Tineke Brunfaut

Affiliation: University of Innsbruck; University of Nottingham

Title: **Exploring the role of phraseological knowledge in second language reading test performance**

Second language (L2) knowledge has been theorised to play a vital role in L2 reading (see e.g. Coady, 1979; Grabe, 2009), and has also empirically been shown to contribute significantly to L2 reading performance (see e.g. Brunfaut, 2008; Shiotsu, 2010; Van Gelderen et al., 2004). Different studies, however, have operationalized L2 knowledge in different ways, with several of these contrasting the contribution of L2 lexical and L2 syntactic knowledge. The results of studies investigating the relative significance of these two components in explaining variance in L2 reading test performance are mixed but with most studies showing greater contribution to
Abstracts June 6

variance from L2 lexical knowledge. The study presented in this poster builds on the existing knowledge base, but at the same time extends it by examining the impact of a broader construct definition of linguistic knowledge. More specifically, it explores whether also taking into account phrasal expressions provides a more comprehensive construct representation of L2 language knowledge and yields valuable insights into the role of phraseological knowledge in reading test performance. To this end, data was collected from 418 English second language (ESL) learners in a European country. The participants completed a reading comprehension test (the dependent variable), and a syntactic knowledge test, a vocabulary knowledge test, and a test of multi-word expressions (the independent variables). The relative contribution of these three L2 knowledge variables to L2 reading test performance was evaluated by means of structural equation modelling. The analyses showed that the superiority of syntactic knowledge in predicting reading test performance, as for example found by Shiotsu and Weir (2007), could not be corroborated; vocabulary knowledge appeared to be a stronger predictor. Importantly, however, the test of multi-word expressions outperformed both the syntactic and the vocabulary knowledge measure. Additional insights into the role of phraseological knowledge were gained by means of a qualitative follow-up study. 15 ESL learners were asked to provide verbal protocols while answering reading comprehension test items that targeted the understanding of multi-word units within written context. Coding and analysis of these data provided further insights into how advanced language learners deal with such lexical items in the comprehension process. It will be proposed that these findings might hint at (1) an important but to date underestimated role of phraseological knowledge in reading ability, and (2) the need for language testers to adopt the notion of “lexicogrammar” (Sinclair, 2004) in their measures, viewing lexical and syntactic knowledge as a cline rather than as a dichotomy. Adopting this view, understanding the role of phraseological knowledge in L2 reading would be an important contribution to a universal framework and such a framework needs to account for this factor.

Poster 7: Friday, June 6, 11:30

Presenter: Samantha Musser and Megan Montee

Affiliation: Center for Applied Linguistics.

Title: Overlap between task models and student production on an oral proficiency test for young learners

What is the influence of model responses on examinee-produced speech? In computer-based oral proficiency testing, there is no test administrator involved who can probe an examinee’s response if he or she fails to provide an answer that addresses the task completely. Is the lack of a full response a function of the examinee’s language proficiency or a misunderstanding of the task demands? In computer-based oral proficiency testing, pre-task modeling may be one way to scaffold an examinee’s full understanding of task expectations. Prior research in Second Language Acquisition suggests that pre-task modeling may have a positive impact on some aspects of student language production (Ellis, 2005; Ortega, 1999; Wigglesworth and Elder, 2010). In addition, some research indicates that modeling may help to reduce the cognitive load placed on students (Ellis, 2006; Skehan, 2006). However, there has been little research or attention to this issue in language testing contexts for young learners producing oral language. This study investigates the influence of a pre-task model in a computer-based English language speaking test for young learners. Throughout the test, examinees hear model responses to speaking tasks before responding to similar tasks. To examine the potential influence of the pre-task model, characteristics of examinee language and the language of the pre-task model are compared. In this poster, the following research questions are addressed: 1. What quantity of language do examinees produce after listening to model responses of varying lengths and proficiency levels? 2. To what extent do examinee responses reflect qualities of the model responses in their own answers? 3. How do examinees perceive the model responses? We address these research questions by analyzing test responses and cognitive interview data from English language learners (N=30) with differing levels of English proficiency in grades 4-5 in U.S. public schools. Data were collected as part of a multi-phase piloting process that included iterative changes to the pre-task models. Task responses and interview data were transcribed, coded and
analyzed for vocabulary and linguistic complexity. Interview results indicated that the model was integral to examinees’ task representation (Plakans, 2010) or how they understood the response expectations. Results also indicated that, while examinees may not use specific linguistic structures from the model, they frequently mirror reasoning structures. The results will be discussed in terms of implications for task modeling, computer-based oral proficiency testing and language elicitation.

**Poster 8: Friday, June 6, 11:30**

**Presenter:** Margaret van Naerssen

**Affiliation:** Immaculata University

**Title:** Language assessment and myths in legal cases

In working towards a Universal Theory, the possible application of principles should also include implications for a wide range of populations beyond those frequently addressed in language assessment. At a macro-level the language assessment community has raised important ethical concerns about various national policies regarding the use/ misuse of certain language assessment instruments and procedures impacting immigration and asylum (e.g., Language and Asylum Research Group). Less visible are the micro-level applications of language assessment in individual court cases where laws vary widely by judicial jurisdiction, by country, within a country, and case by case. At both the macro and micro levels serious awareness raising efforts are needed to overcome folk beliefs / myths many laypersons have about the communication skills of non-native speakers. Those holding such beliefs include politicians, program administrators, and forensically-related groups such as law enforcement officers, lawyers, judges, and jury members. These beliefs are also sometimes even perpetuated in judicial opinions. For professionals in forensic contexts, there is a need for increased understanding of the strengths and limitations of language assessment when it is used as evidence. This, in turn, can contribute to the credibility of language assessment when it is done appropriately. However, the situation is aggravated when well-meaning language testers (but not well-trained) do not apply the basic principle of validity, in the choice of language assessment instruments. Additionally they may not apply evidence argumentation appropriately. They may, at times, only loosely (if at all) link language assessment descriptors to the actual language evidence in a case. Court rules of evidence also need to be applied. It should not be simply a matter of giving a test, reporting a score or descriptors, and then jumping to an opinion about the legal question. These forensically-related groups, such as judges, are typically not included in the populations frequently addressed by many in the language assessment community. Certainly parents, teachers, learners, language training/ program administrators, academic-specific language trainers, and so on, are important groups. Yet in forensic contexts the outcomes of some cases can be life or death decisions, at least as critical or at times more so than decisions about, for example, highly valued university admissions. Judges may end up becoming language testing “experts” (Wennerstrom 2011). The focus of this poster then is at the micro-level in forensic contexts. The poster is organized around five myths. Related examples from judicial opinions (including Wennerstrom) illustrate “acceptance” of these beliefs, including examples related to language assessment. Counter evidence from second/ foreign language acquisition research and sociolinguistics is also briefly presented for each. Relevant principles / practice from language assessment are also introduced for each, along with relevant judicial variables. Handouts are provided. Reference: Wennerstrom, A. (2011). Why is this judge a language testing expert?” Presentation at IAFL Conference, University of Aston, Birmingham, UK.
Poster 9: Friday, June 6, 11:30

Presenter: Sibylle Plassmann

Affiliation: telc - language tests

Title: Assessment for the workplace: the new German test for doctors

The poster will present two new tests for medical personnel: telc Deutsch B1-B2 Pflege and telc Deutsch B2-C1 Medizin. Non-native speakers of German need at least a B2 certificate to practice medicine in Germany and at least B1 to work as a registered nurse. The demand for a specific test led to the development of two new job-related formats which reflect everyday language demands placed on doctors and nurses. A thorough needs analysis was conducted first. The findings then had to be matched with the relevant CEFR levels as indicated by law. On a practical level, question such as this arose: What does it actually mean to provide patient care documentation? And how can you find common ground considering the characteristics of such a text and the descriptors for B1 or B2? The challenge to interpret the universal reference system that the CEFR provides in view of language use in the hospital led to new descriptors for the language use of doctors and nurses. The poster will show examples of these descriptors for the special test purpose in the light of the CEFR scales. The test specifications with more in-depth information will be available for consultation.

Poster 10: Friday, June 6, 11:30

Presenter: India Plough

Affiliation: Michigan State University

Title: The Local Informing the Global

The local informing the global Despite federal initiatives to improve the level of foreign language proficiency and to increase the number of languages acquired by the people of the United States (U.S.), results of the 2006 and 2008 General Social Survey (Rivers and Robinson, 2012) indicate that the proportion of speakers of languages other than English in the U.S. has remained relatively unchanged for 35 years. Additionally, the overwhelming majority who self-assess their ability to speak that language “very well” learned that language at home and not in school. Within this national context, the faculty and students of a new residential college within a large land-grant university remain committed to making the study and use of world languages an essential part of the residential college experience. Since opening its doors in 2007, the college’s language proficiency program has faced a number of challenges. Two significant programmatic revisions have been initiated as a result of a comprehensive review (2012-13). First, implementation of a Cultures and Languages across the Curriculum (CLAC) program is underway; second, a performance-based test of speaking proficiency is under development to replace the currently required Oral Proficiency Interview. This poster presentation, organized into three main sections, describes the design, format, and rating of the test, which is necessarily situated within the context of the CLAC program, the larger curriculum, and the college’s mission. First, the motivations for the creation of an RCAH test and revisions to the program are explained by way of an overview of selected elements of the review process: identification of stakeholders; the quantitative and qualitative instruments developed to explore test purposes and uses of test results; and stakeholder recommendations for assessment of student proficiency. Presentation of data focuses on those interpretations that have led directly to the creation of an RCAH performance-based speaking test and an RCAH CLAC program. The college’s language proficiency program treats the processes of teaching, learning, and assessment as interdependent and symbiotic. That is, teaching and assessment must be informed by a model of learning; assessment must be aligned with learning objectives; and, this alignment is better accomplished if teaching and assessment are conducted in the same way (National Research Council, 2001). An interactionalist, socio-cultural model of acquisition and assessment has been adopted. The performance-based test and the CLAC program are built on
the same pedagogical principles and incorporate similar pedagogical practices. The second part of the poster elaborates on the incorporation of these shared principles and practices into the test. The final section describes the current version of the test and rating scale. Initial level descriptors of the scale are informed by test trials, the goals of the instructional units within the CLAC program, test score uses, and internationally recognized scales of speaking proficiency. Empirical validation and further revision of the scale begins with the first test administrations (Fall 2014). The need to develop a local test, with accompanying scale, in lieu of using existing, validated measures of speaking proficiency should contribute to the dialogue of a universal validity.

**Poster 11: Friday, June 6, 11:30**

**Presenter:** Natalia de Andrade Raymundo and Natalia Guerreiro Marques

**Affiliation:** ICEA

**Title:** The Validation Process on the Brazilian English Proficiency Exam for Air Traffic Controllers

This study aims at presenting initial research on the ongoing validation process of the Brazilian Air Force English proficiency exam for Air Traffic Controllers, named EPLIS. EPLIS is a performance test developed by language teachers and aeronautical subject matter experts in Brazil, in compliance with the International Civil Aviation Organization (ICAO) requirements. EPLIS, which is considered to be a high stake exam because of its relevance for society, assesses the performance of Brazilian ATCos in communicating in non-routine situations within an international community of users of the English language. This exam has been going through validation studies conducted by language testing experts so as to attest its reliability and validity. In this presentation, we will bring into question some criticism of the exam and of the very framework it set out to follow, by making a parallel between what the ICAO rating scale considers as operational in Air Traffic Control regarding English proficiency and what language skills Brazilian Air Traffic Controllers need in order to deal with language complications.

**Poster 12: Friday, June 6, 11:30**

**Presenter:** Daniel Reed and Heekyoung Kim

**Affiliation:** Michigan State University

**Title:** New grammar tasks that reflect textual complexity features associated with 2 CEFR levels

In the past few years, several studies have investigated the merit of using automated measures of textual complexity in the development of reading comprehension tests (cf., Green et al., 2013; Kim et al., 2013; Ohlrogge et al., 2013; Reed et al., 2013) and speaking tasks (Kang, 2013). The present study considers the implications of the results of such studies for the grammar sections of two examinations that target the B2 and C2 ranges of the CEFR, respectively. Reed et al. (2013) supported claims of greater complexity of the C2 texts relative to the B2 texts for many of the measured features, but with mixed outcomes for “real world” texts. Of particular relevance to the current study were the findings by Green et al. (2013) that ‘features of cohesion... and lexis... rather than syntax are critical in distinguishing between the texts used in the three highest levels of the Cambridge English Examinations,” which motivated the current investigators to consider redesigning the grammar sections of Michigan State University’s Certificate of English Language Competency (MSU-CELC), a four-skill examination in the CEFR B2 range, and MSU’s Certificate of English Language Proficiency (MSU-CELP), a four-skill examination in the CEFR C2 range. The redesign of CELC and CELP items included increased contextualization with a greater emphasis on lexis and features of cohesion (e.g., focusing on links between sentences). The current study used the tasks that performed best for Kim et al. (2013) and lengthened the pilot tests. These tasks included both cloze format and contextualized grammar and vocabulary and were administered to both high and low ability groups to see if a basic B2/C2 distinction was upheld and sharpened.
Statistical analyses included T-tests on the differences between mean values of complexity indices as well as correlations between complexity measures and item p-values. Implications for revising specifications and the validity arguments for these examinations will be presented based on the empirical results from these comparisons.

**Poster 13: Friday, June 6, 11:30**

**Presenter:** Caroline Riches  
**Affiliation:** McGill University  
**Title:** Language Assessment Development with Haitian Teachers

This poster will report on a week of workshops on EFL exam creation held with approximately 130 high school teachers from across the Northern Department of Haiti this past March. These workshops were supported in part with a Workshops and Meetings Grant from ILTA. We will share our concurrent research data, which reveal the success of this collaborative process as well as indicating future challenges. English National Examinations in Haiti are poorly constructed and unreliable, which is especially troubling considering that they are very high stakes: students cannot complete high school or access higher education if they fail. In a country where fewer than 25% of students even attend high school, school completion is already a tremendous challenge. Unfortunately, serious problems in the basic construction of these examinations conceivably affect the pass rate. A concrete way to improve the quality of these examinations is by training teachers in assessment, as teachers prepare and submit the National Examination questions. However, the great majority of teachers have no formal training in teaching or assessment. During this week of workshops, teachers analysed and suggested revisions to draft English national examinations provided by the research team. In addition, teachers learned the basics of language assessment and item writing, and practiced drafting their own items for the examinations. Research data were also collected during this trip. The research questions were the following: 1) How do the final exams created by the assessment specialists and teachers respond to needs in Haiti and improve upon current examinations? 2) What are the most beneficial and problematic characteristics of this collaborative process? Results will be presented from participant surveys, as well as from interviews with 10 teacher participants following the workshops (analysed using the constant comparative method). Results reveal the overwhelming benefits of the collaboration: our combined expertise resulted in the eventual creation of 5 secure examination versions and one practice version, as well as answer keys and scoring guides. These versions were judged by both parties to be a great improvement upon current exams in place. In addition, all participants viewed the collaboration positively and elaborated on the positive effects of the experience on their future teaching and assessment practices. However, during our collaboration other concerns about the national examination process arose. For example, teachers explained how the paid contracts for exam grading are often awarded as gifts to friends, not to people with appropriate qualifications. This means that English examinations are often graded by individuals with little to no English ability. This results in practices such as correct items being marked incorrect if they do not follow the response in the answer key verbatim. Therefore, while exam creation is a worthwhile beginning, any future work will have to address grading processes. In addition, it remains to be seen if the final exam versions will be accepted for consideration by the Ministry of Education of Haiti (the MENFP) to replace the current National English Examinations. Very persistent negotiations are ongoing.
Poster 14: Friday, June 6, 11:30

Presenter: Susan Sheehan

Affiliation: University of Huddersfield

Title: Testing Trainee Teachers

This poster proposal describes a new test which has been created for prospective teachers of English. The test has been developed according to the requirements of the agency which validates the teacher training course. The test takers are prospective English language teachers. They are adults and educated to degree level. They are students at a British University who have enrolled on Masters TESOL course. Some of the test takers so have English as their own language. They have to complete a test to demonstrate they are at C1 level in speaking, listening, reading and writing. Test takers who have English as their own language are required to take a test which demonstrates their language awareness. Test takers who have passed the first test are then expected to take the language awareness test. So, two different types of test have been developed and this poster will describe the challenges of developing simultaneously two inter-related but substantially different tests. The first section of the poster will focus on how the test developers choose to create a test which meets the criteria of demonstrating a C1 level. The test writers drew heavily on a published guide to the linguistic features of C1. Thus, the language used in the tasks was hoped to be representative of C1 and the test takers responses would also be judged to be C1. A possible weakness of the test is that by drawing on one source the test replicates the deficiencies of its source. The second part of the test covers language awareness. Here defined as the ability to recognise and describe language in terms of grammar, lexis and phonology. The validation authority stipulates that the test should be timed, handwritten and completed under examination conditions. The task must produce a stretch of writing long enough to demonstrate appropriate structure, spelling and punctuation. Thus, the candidates are being tested on both their level of language awareness and their ability to express this awareness in an appropriate manner. The task relates closely to their future work as English language teachers. This would seem to be reasonable but it could be argued that such a test is asking them to demonstrate these abilities before they have undertaken the training course which is designed to develop them. This project could be described as innovative in the approach taken to demonstrating that the first part of the test is a C1. Both parts of the test have to relevant for future teachers of English. This is relevant to the conference themes of developing descriptors for academic disciplines and developing job-related descriptors. The test is small scale in terms of test taker numbers but high stakes for the test takers as it relates to training and job prospects. This paradox will be explored in the poster’s conclusion.

Poster 15: Friday, June 6, 11:30

Presenter: Tom Velghe

Affiliation: Thomas More @ KULeuven

Title: VocabLab: an accurate test for Flemish learners’ English and French vocabulary size

This poster reports on an English and French vocabulary size test, i.e. the VocabLab tests, specifically designed for Dutch-speaking learners in Belgium. In the development of the English test, our purpose is to improve the validity of the well-known Vocabulary Levels Test (Schmitt et al. 2001) which turns out to be not sensitive enough to adequately measure the vocabulary size of Flemish undergraduate students. For the French leg of the VocabLab project, the vocabulary size test is created from scratch, as a valid French test is not available at present. Such a test is, however, of particular relevance to Flemish learners given the status of French as the second official language in Belgium. More specifically, we aim to develop form-recall and meaning-recognition tests for both English and French which (i) rely on up-to-date frequency information, (ii) pinpoint the vocabulary level of upper-intermediate learners in an accurate, valid way and (iii) are tailored to the Flemish
Abstracts June 6

context. Firstly, the item selection for the VocabLab tests is based on recent frequency lists extracted from the 450-million-word Corpus of Contemporary American English (COCA, Davies 2008-) and the Frequency Dictionary of French (Lonsdale & Le Bras 2009, based on a 23-million-word corpus). The detailed frequency data enable us to define a more adequate representative noun-verb-adjective ratio which, in contrast with Nation’s (1990) 3:2:1 ratio and Schmitt et al.’s (2001:58) 5:3:1 ratio for the English VLT test, differs per frequency band (e.g. the COCA data display a 6:3:2 ratio in the 1 to 1,000 band versus 3:1:1 in the 4,001 to 5,000 list). Secondly, we address the sampling rate and differentiation within the frequency bands, viz. 1-2,000, 2,001-3,000, 3,001-5,000, and 5,001-10,000, a concern also raised by Schmitt et al. (2013) regarding the Vocabulary Size Test (Nation & Beglar 2007). Preliminary results of the VLT test administered to Flemish undergraduate students show that the VLT might not be sensitive enough to measure learners’ vocabulary size of the frequency bands between 3,000 and 5,000. Therefore, the English and French VocabLab tests take in five different frequency bands, viz. 1-1,000, 1,001-2,000, 2,001-3,000, 3,001-4,000 and 4,001-5,000. Thirdly and crucially, our test design is adapted to Dutch-speaking learners in Flanders. The VLT contains a considerable number of cognates and loanwords, including items that are orthographically identical to Dutch (e.g. sport). Consequently, the VLT runs the risk of also measuring successful guessing strategies, resulting in an overestimation of learners’ vocabulary size (Schmitt et al. 2013). This may even more so be the case for Flemish, which is prone to English borrowings and known for its historical influence from French – especially in regional variants. Therefore, the item selection for both the English and French test was preceded by a careful screening for Germanic and Romance cognates. In this poster, we will discuss the design and item selection of the French and English VocabLab test in more detail. In addition, we will present the first results of the pilot study.

Poster 16: Friday, June 6, 11:30

Presenter: Roxanne Wong

Affiliation: City University of Hong Kong

Title: Feedback for Diagnostic Writing

In 2011 an online diagnostic writing assessment was undertaken in a joint project comprising 3 universities in Hong Kong. A team of language and assessment professionals designed a set of prompts and improved upon a pre-existing scoring rubric to be used by an AES system to score written essays of 450 words. Approximately 1500 students took part in the pilot study. The students wrote their essays using the MetaMetrics EdSphere online system and Lexile scores were generated for each student. While this is informative for teachers, the lexile scores are unhelpful for students. This poster reports on the initial phase of my PhD research on the types of feedback that can be provided in conjunction with the Edsphere writing platform. Providing useful feedback for the students to be able to improve their overall writing is an important part of 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." I will explain the rationale behind the score distribution breakdown and the designed feedback that was generated for the students who participated in the pilot study. I will also give some of the initial responses from both teachers and students on the effectiveness of the feedback generated.
Poster 17: Friday, June 6, 11:30

Presenter: Amery Wu and Jake Stone

Affiliation: The University of British Columbia

Title: Detection of Plagiarism in the Writing Component of the Computer-Administered CELPIP-G

Although the application of electronic plagiarism detection systems is used frequently in higher education, little literature can be found specifically detailing an application of such technologies in operational test use. The Canadian English Language Proficiency Index Program – General (CELPIP-G) test measures functional English language proficiency in four functional domains (reading, listening, speaking, and writing), using a computer-administered format. The CELPIP-G is considered to be high-stakes, as CELPIP-G scores can be used as mandated evidence of English language proficiency in Canadian citizenship and immigration applications. In these high-stakes situations, testing programs such as CELPIP-G must be vigilant in ensuring the test scores provided to score users indeed represents a test-takers’ independent effort and true ability. As neatly summarized by Cizek in 2005: “when cheating occurs, inaccurate inferences result” (p. 7). The detection technique used with the CELPIP-G was originally presented by Lyon, Malcom, and Dickerson (2001), and involves the calculation of a Similarity Index between two texts. This index is calculated by taking the proportion of three-word groupings (“trigrams”) that are shared between two texts. Texts that are highly similar – and thus potential cases of cheating – have a higher proportion of shared three-word groupings than that which is observed in natural language use. The simplicity of this technique allows it to be used as part of the CELPIP-G analytical battery at the conclusion of each testing session, without the need for elaborate information technology solutions. This is possible due to the computerized nature of the test, where texts do not need to be digitised before analysis is performed with a computer. To date, the CEPIP-G testing program has successfully identified a number of pairs of texts where the probability of cheating is high. This paper will explain the Similarity Index, describe how to implement this cheating detection technique in a large scale testing environment, present the distributions of the Similarity Index observed in actual test-data, and provide sample texts that have been identified as being more similar than expected without breach of confidentiality. It is the responsibility of every testing programme to ensure scores released under their name indeed reflect genuine effort by the test-taker named on the score report. The three-word trigram technique is a simple, practical, and easily programmed method that can be used in a computer-administered language testing application to detect potentially plagiarised texts. The use of this technique has revealed that despite active invigilation and tightly controlled conditions at the point of test delivery, plagiarism still can (and does) occur.

Poster 18: Friday, June 6, 11:30

Presenter: Hui-Chun Yang

Affiliation: National Hsinchu University of Education

Title: The nature of summary writing ability: Linking strategy to performance

Summarization skills are crucial to academic success in higher education (Braine, 1995; Maclellan, 1997; Woltersberger, 2008) because the ability to summarize is conducive to effective content acquisition (Coffin, 2006; Friend, 2002; Hood, 2004), reading comprehension (Cordero-Ponce, 2000) and written communication (Garner & McCaleb, 1985; Spack, 1988). Embodied in the notion of communicative language testing, assessing summarization ability has become one of the most promising areas in language testing (Cohen, 1994). A growing number of language tests (e.g., The Internet-based Test of English as a Foreign Language, Canadian Academic English Language Assessment, General English Proficiency Test, Georgia State Test of English Proficiency) have incorporated tasks which involve summarization in their assessment batteries because such tasks may promote test fairness by providing scaffolding content (Feak & Dobson, 1996; Leki & Carson, 1997).
and enhance positive washback effects on language education due to their closer approximation to real-life writing tasks (Cumming, Grant, Mulcahy-Ernt, & Powers, 2004). The complexity of such tasks, however, can jeopardize the construct validity of the test due to the threat of ‘muddled measurement’ (Weir, 2005) stemming from the joint influence of reading and writing abilities on summarization performance (Alderson, 1996; Charge & Taylor, 1997; Fox, 2003). These concerns deserve further investigations on the validity of summarization tasks. To this end, this study explores the construct of a summarization test task by means of single-group and multi-group structural equation modeling (SEM). It examines the interrelationships between strategy use and performance, drawing upon data from 298 EFL undergraduates’ summary essays and their self-reported strategy use. Single-group SEM analyses indicated that the task required multiple constructive strategies of comprehension and composition including planning, evaluating, discourse synthesis, and sourcing strategies. The planning and evaluating strategies provided administrative control over other strategies employed at varying degrees to improve summarization performance. Multi-group SEM analyses showed that the effects of the planning and evaluating strategies on the discourse synthesis and sourcing strategies appeared to be invariant across the high and low groups. However, the effects of the discourse synthesis and sourcing strategies on summarization performance were found to vary across groups. The significant positive loading from discourse synthesis to summary writing ability revealed that writers in the high group were more likely to be engaged in highly constructive act of discourse synthesis and such involvement had affected positively on their summarization performance. Their less skilled counterparts, however, were more likely to generate their summaries by manipulating surface-level discourse structures and linguistic forms of the source texts. An unexpected result was also found: sourcing strategies had a negative and a positive impact on the high-scoring and low-scoring writers’ summary writing performance respectively despite the fact that both paths were not statistically significant. The findings offer insight into the nature of summarization tasks, and provide useful validity information for the test task. Implications of the findings are discussed in relation to second language writing assessment and pedagogy.

Poster 19: Friday, June 6, 11:30

Presenter: Ying Zheng and Liying Cheng

Affiliation: Southampton University

Title: The Effects of Test Preparation on Chinese Test Takers’ Pearson Test of English Academic Performance

By 2008, nearly 1.4 million Chinese students study abroad for better education and future career (MoE, 2010). For these Chinese students, the English language tests (e.g., Pearson Test of English Academic (PTE Academic), TOEFL, and IELTS) are the gatekeepers for their university entrance. Therefore, many Chinese students take test preparation courses accordingly to seek improved test performance. However, only limited studies have been conducted to investigate how Chinese students’ test preparation practices influence their test performance (e.g., Chau, 2008; Xie, 2008; 2011; Yu, 2012; 2013). All of these studies have involved with the IELTS and the TOEFL; however, no studies have investigated the more recent PTE Academic, which is brand new to Chinese students in term of its test format and structure. This study examined Chinese students’ test preparation practices for PTE Academic and the relationship between their preparation practices and their PTE Academic performance. The test-takers (n=25) who took PTE Academic preparation courses at a test preparation centre in northern China and then the live PTE Academic were recruited as the participants for this study. Data were collected using four instruments: a) Common European Framework of Reference (CEFR) self-assessment grid, b) a questionnaire on PTE Academic test preparation practices, c) live PTE Academic tests, and d) a semi-structured interview guide on students’ PTE Academic preparation experience. The following statistical analyses were conducted to reach an understanding of the relationship between the participants’ test preparation practices and their PTE Academic performance: (1) the descriptives to depict the characteristics of the participants and their PTE Academic preparation practices; (2) regression analyses to investigate the effects of various aspects of test preparation practices on their PTE A performance. Given the small group of
participants for the statistical analyses, the qualitative data were complementary to the interpretations of the statistical analyses results. An inductive approach were used to analyze the interview data to understand the participants’ experience and perceptions of taking the PTE Academic test preparation courses. The preliminary results showed that the instructions that the Chinese students received from the PTE Academic test preparation courses were mostly test-oriented although these students mostly expected to improve English language abilities. With these expectations, the students attended more to the methods of building their English language skills and they perceived the improvement in their specific language skills after taking the courses. The students with higher PTE Academic scores significantly made better use of the logic in oral expression they learned from test preparation courses than those with lower scores. The further results from the ongoing regression analyses and interview data will be linked to explain the effects of the various dimensions of these Chinese students’ test preparation practices on their test performance. The results of this study provide empirical evidence of how test preparation influences Chinese students’ performance on PTE Academic. Accumulated with results from the investigations on this topic in other English language test contexts, this study will also contribute to reach an understanding of Chinese students’ test preparation for high-stakes language tests.
Symposium 3: Friday, June 6, 14:00

Presenters: Liz Hamp-Lyons University of Bedfordshire, Cathie Elder University of Melbourne, Dan Douglas University of Iowa, Jane Lockwood City University of Hong Kong, Yan Jin Shanghai Jiaotong University, Sally O’Hagan University of Melbourne, Tim McNamara University of Melbourne, John Pill University of Melbourne, Alan Davies University of Edinburgh

Title: The Universal and the Specific: must proficiency choose?

The papers in this Symposium will consider whether the aspiration for a universal framework of language proficiency is achievable and compatible with the assessment demands of particular occupational and academic contexts. Two questions are raised: 1. Is language proficiency always universal or always specific or both? 2. Who decides on the criteria for success: the more form-focused language testers or the more use-focused local stakeholders? After a brief introduction from the Symposium convenor, the four presenters will speak for 15 minutes each. The discussant will offer a 10-minute response to the papers and invite attendees to debate the issues raised. Paper 1 considers the quest for language for specific performance assessment validity from both linguistic and business viewpoints and is situated in the context of outsourced Asian call centres. It argues that such LSP performance assessment must be nuanced and co-constructed with language experts (LEs) and subject matter experts (SMEs), and furthermore crafted to satisfy the needs of quality assurance end-users. This highly tailored and collaborative approach makes the application of a universal language framework unlikely. Paper 2 outlines a study of the indigenous criteria oriented to by health professional supervisors when evaluating patient-trainee communication. A contextual, empirical, bottom-up model of what these clinicians value yielded new professionally-relevant criteria for the speaking component of the Occupational English Test (which uses simulated clinical role-plays but is assessed with general language proficiency criteria). Application of the new criteria led to different, more defensible admission decisions. The paper argues against a universal framework and universal criteria for assessing communicative competence in workplace settings. Paper 3 reports a needs analysis of English for Professional Purposes (EPP) for graduates in China which uncovered both common and profession-specific language needs. It argues that a universal framework for assessment of EPPs would necessarily be of limited use: a structure is needed that incorporates such features as domain-specific knowledge, context-based text types and activities, technical/semi-technical vocabulary, and other salient features in addition to a more generalized framework for lower-order language usage. Paper 4 argues that a failure to include native English speaking professionals in the International Civil Aviation Organization English assessment means that this group remains largely unaccountable for communication problems that bear heavily on air travel safety. It suggests that an aviation English test should include a focus on intercultural communicative abilities for both native and non-native English speakers. Such an approach would help achieve a balance between a universal framework and context-specific language assessments. The discussant argues that the issue addressed in these papers is an old one: setting grand narrative realists/modernists against deconstructionist/post-modernists/nominalists. Neither is wrong. Both are needed. A universal framework is essential to ensure that different LSPs communicate with one another; at the same time that universal framework must accommodate variation to take account of context-specific distinctions.
Abstracts June 6

Symposium 4: Friday, June 6, 14:00

Presenters: Claudia Harsch University of Warwick, Johannes Hartig German Institute of Educational Research, Tom Lumley ACER, Juliette Mendelovits ACER, Dara Ramalingam ACER, Tineke Brunfaut University of Lancaster, Gareth McCray University of Lancaster

Title: Towards a universal framework for modelling item difficulty characteristics: Increasing transparency by psychometrically robust definitions

Modelling item difficulty for tests targeting receptive skills is a recurrent theme in language testing research, since predicting item difficulties allows reporting test results in an understandable way, improves item writer guidelines and facilitates validation studies (Embretson 1998). Despite much research on which characteristics can best predict item difficulties, no unified framework has yet been established which could be employed across domains or (first and second) languages. Such a framework would be appealing since it could coherently describe key factors which contribute to item difficulties and which can increase the comprehensibility of reporting test results: If results are reported in terms of proficiency levels, these levels can be transparently described by the key factors contributing to item difficulties. Moreover, a framework of key characteristics which is applicable across domains and/or languages would greatly enhance communication with stakeholders.

The way towards such a universal framework, however, faces several challenges: Not only are there challenges in establishing the nature of the key item-difficulty characteristics (IDCs), there are also methodological questions about the most suitable way to model item difficulties. This symposium brings together existing research into IDCs for first and second language tests, for the domains of listening, reading, and vocabulary knowledge. It also presents Explanatory IRT models which are unexplored in the field of language testing but have the potential to offer novel insights into the feasibility of developing a universal IDC framework. The symposium begins with an introduction to the topic, followed by four papers focusing on conceptual as well as methodological issues. The first paper investigates factors influencing L1 reading item difficulty, reporting on studies that used human ratings to apply selected IDCs to print and digital reading items from PISA 2009. The findings of the regression analyses indicate that ten identified IDCs can be reduced to five factors for print reading, but differences were observed in the digital environment. The second paper explores the suitability of the LLTM+e Explanatory IRT model to synthesise judgement and performance data in a single modelling framework in order to explain item difficulties. By employing the above PISA-IDCs to an L2 test (PTE Academic), the study also aims at bridging L1 and L2 environments. A combination of methodological approaches is taken in the third paper, which employs an innovative test-design, operationalising four pre-selected IDCs into L2 reading and listening tests. The same stimuli and items are used for both domains to examine differential effects of the IDCs across domains. Regression and Explanatory IRT (EIRT) analyses show that two of the IDCs can explain substantial variation in both domains. The fourth paper further explores EIRT models for their potential to reveal the psychometric basis of item difficulties. EIRT is applied to vocabulary test items to illustrate the possibilities for comparing how variations in item content and design affect difficulty for distinct test-taker groups. The symposium concludes with the discussant critiquing the four papers, before discussion of implications and ways forward towards a universal framework for modelling item difficulty is opened to the audience.
Abstracts June 6

Parallel Session 16.1: Friday, June 6, 14:00

Presenter: Yasuyo Sawaki

Affiliation: Waseda University

Title: A qualitative analysis of tutor behaviors during practice reading exercise review sessions for test preparation

While providing detailed performance feedback to learners may promote test preparation, previous research (e.g., Hamp-Lyons, 1998; Wall & Horak, 2011) shows that language test preparation materials often lack information that helps learners and teachers identify strengths and weaknesses and plan on further studies. In reading test preparation materials, descriptive information about individual items (e.g., scoring results, correct answers and their rationales) is often provided as feedback. However, it is not necessarily clear how such information could be utilized in subsequent test preparation activities, suggesting the need for other types of feedback that supplement the information and promote learners’ engagement in further learning activities. To take a step forward in this direction, the present study examined behaviors of experienced instructors during one-on-one tutoring sessions for reviewing reading exercise results as part of TOEFL iBT® test preparation.

Analyzing the process of one-on-one tutoring, a type of instruction that generates extensive individualized feedback with demonstrated instructional effectiveness (e.g., a meta-analysis by Elbaum, Vaughn, Hughes, & Moody, 2000), could offer insights into how one might go about designing enhanced performance feedback for L2 reading test preparation. Participants were five experienced Japanese teachers of English who served as tutors and 12 adult Japanese learners of English. Each learner completed 18 TOEFL iBT practice reading sets individually online over a three week period. During this period, each learner met with a tutor in a weekly three-hour one-on-one tutoring session where they reviewed six sets the learner had completed in the week. The tutoring sessions were semi-structured. The tutors followed tutoring materials for reviewing the reading set content developed for the purpose of this study, while they were allowed to adapt the instructional procedure depending on their moment-by-moment assessment of individual learner needs. For all tutor-learner pairs, the tutoring sessions on the same four sets were videotaped. Transcripts of those videotapes were segmented, and the segments were coded for two categories: (1) skills, processes, and strategies for reading and test-taking being discussed and (2) tutor and student move categories based on previous tutoring process studies (e.g., Cromley & Azevedo, 2005; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Thompson, 2009). Double coding of approximately 25% of the segments showed an acceptable level of coder agreement. A qualitative analysis of the coding results showed the tutor-learner pairs’ engagement in discussion on various metacognitive, cognitive, and test management processes and strategies in response to learner comprehension difficulties. While instructional tutor moves (e.g., explaining, asking/answering questions, giving examples) accounted for a large portion of the tutoring time in general, the tutors differed in the types and frequencies of cognitive scaffolding moves (e.g., giving hints) and motivational scaffolding moves (e.g., giving positive, neutral, and negative feedback to the learner) that they employed. Various tutor strategies to accommodate reading needs of readers at different levels were identified as well. In this presentation the key study findings will be summarized, along with their potential implications for designing feedback for L2 reading test preparation materials.
Parallel Session 16.2: Friday, June 6, 14:00

Presenter: Seon-Hi Shin

Affiliation: Seoul National University

Title: Developing Performance Level Descriptions for a Writing Test

It is an important task in language testing to establish performance level descriptions which can distinguish examinees’ levels of achievement in order to assist examinees, teachers, administrators, and others in relating the test scores to examinees’ language skills and understanding. The need to have a list of well-defined and articulated descriptions for different performance levels has increased in recent years since language assessments providing information with more diagnostic and pedagogical value are in greater demand in the language testing community. Several approaches to developing performance level descriptions have been reported in the literature. Scaling anchoring is one method used to achieve this goal (ETS, 2005). The basic idea of this approach is as follows: a few carefully spread scale points, called anchor points or benchmark points, must be selected. Then, content area test specialists review the anchor items to which examinees on a pair of adjacent anchor points responded at different success rates, and describe what the examinees at anchor points on a scale know and can do. There are two approaches to scale anchoring. The direct method uses discrete item response functions, while the smoothing method uses a curve-fitting procedure to create smoothed empirical item response functions (Beaton & Allen, 1992). Another approach to establishing performance level descriptions was used by ACT for developing the College Readiness Standards (CRS). ACT’s content area test specialists wrote the CRS based on their analysis of the skills and knowledge that students needed in order to respond successfully to test items that were answered correctly by the majority of the examinees who scored within each score range (ACT, 2011). The CRS approach started with forming groups of examinees whose scores fell within each of the defined score ranges. Multiple forms of the test were reviewed to determine the knowledge, skills, and understanding that are necessary to answer the test items in each score range. Then, the additional knowledge, skills, and understanding that are necessary to answer the test items in the next score range are identified. This process is repeated for all the score ranges. All the lists of statements identified by each content specialist are merged into a composite list for individual reviews. The list is refined through several rounds of individual and group reviews and through discussions via a master list into the final list of skills and understanding accurately described for score ranges. ACT’s example in the literature (ACT, 2011a, 2011b) used multiple-choice items. This study expanded the ACT’s CRS method to embrace a performance test whose items were graded on four dimensions over a 0-to-5 scale by two independent judges. Performance level descriptions were constructed for an English writing test through multiple rounds of reviewing a list of items-by-score range, anchor papers, scoring rubrics, and anchor items. The test had three items. There were a total of eight performance levels. The current study introduces a feasible method to establish performance level descriptions for a language test consisting of constructed response items.

Parallel Session 17.1: Friday, June 6, 14:35

Presenter: Limei Zhang and Christine Goh

Affiliation: National Institute of Education, Nanyang Technological University

Title: Gender Differences in Strategy Use and Reading Test Performance: A Study of Chinese College Students

Language assessment researchers have perceived test-takers’ strategy use as an important factor in test validation and test performance (Cohen, 2006). For example, Bachman (1990) argued that a critical limitation to the quantitative approaches to construct validation is that only the products of test taking processes (i.e.,
Abstracts June 6

test scores) are investigated, but no means are provided to examine the test-taking processes themselves. Messick (1989) also pointed out that test takers “differ consistently in their strategies” (p. 54). This study was designed to make investigation into the differences in strategy use and its relationships to test takers’ reading test performance across gender groups through the use of multi-group structural equation modeling (SEM) approach. The findings of the study are expected to provide implications for test developers and language instructors. Bachman and Palmer (2010) believe that language knowledge (i.e., grammatical knowledge, textual knowledge, and pragmatic knowledge) is managed by a set of strategies which determine how language is realized in actual language use. Similarly, the role of metacognition in reading comprehension has long been recognized. The consensus view is that students’ reading comprehension is closely related to their appropriate use of strategies (e.g., Pearson, 2009; Paris, Lipson, & Wixson, 1983; Zhang, 2010). Further, researchers have conducted numerous studies addressing the differences in academic performance according to gender. This study examines how male and females students are similar or differ in their employment of different strategies when engaging in reading comprehension activities in the test context. 584 (274 male and 310 female) Chinese college test takers were invited to answer a 38-item strategy use questionnaire and sit for a 50-item reading comprehension test. The questionnaire was developed based on relevant theories and empirical studies (Bachman & Palmer, 2010; Pressley & Afflerbach, 1995) and validated before actual use. The reading test, the College English Test (CET-4) Reading subtest, is the most influential college English test from the mainland China (Jin, 2008). First, baseline models of strategy use and reading test performance were established for male and female groups. Then cross-validation was conducted across gender groups. The study supported the invariance of the factorial structure of the reading comprehension test, the relationships between test takers’ strategy use and reading test performance, and part of the strategy use questionnaire across male and female groups. The result showed that while gender differences can be perceived in the reading ability level and strategy use patterns, males and females are similar in the way of responding to reading comprehension tasks and employing strategies to tackle the tasks. Implications of the findings for classroom teaching and test development are discussed.

Parallel Session 17.2: Friday, June 6, 14:35

Presenter: Donald Powers

Affiliation: Educational Testing Service

Title: Modeling young learners’ growth in English language proficiency: Towards a validity argument for a new proficiency assessment

This study examines the extent to which an assessment of young learners’ (age 11 and above) proficiency in English as a foreign language is capable of reflecting anticipated learning effects. According to Messick (1989) and Chapelle et al. (2008), a validity argument can be based on the extent to which test scores reflect appropriate changes as the result of construct-relevant learning. Young learners’ language ability develops as the result of their learning both in and outside of formal schooling. It is therefore imperative for an assessment developed for young learners to exhibit an ability to reflect both (a) within-individual growth that accrues from learning, and (b) differential growth patterns across individuals with different characteristics and experiences. This study investigates the research question: To what extent does the test reflect changes in language proficiency (a) as a function of time elapsed between repeated observations within individuals, and (b) as a function of self-reported learning variables across individuals? The test measures Reading, Listening, and Language Form and Meaning (LFM). A total of nearly 3,000 students took the test at least twice. A subset of them, about 300, took the test three times. All participants were engaged in English language learning in various instructional programs during the course of the repeated observations. They also self-reported on four construct-relevant learning characteristics (LCs): years of studying English (LC1), hours per week studying
Abstracts June 6

English in school (LC2), hours per week studying English outside of school (LC3), and length of living in an English-speaking country (LC4). A multivariate repeated-measure model was fitted to the data from two time points. The within-subject growth effect was significant, showing that learners scored higher on the second test-taking. Test-takers were separated into three groups by the length of time between test-takings: less than 150 days, between 150-210 days, and more than 210 days. Group effect was significant. Test-takers in the third group performed differently from the other two groups, especially in LFM and Reading. This group also made the highest gains on all sections on the second test-taking. The results also showed that the effects of LC1, 2 and 4 were significant at differentiating test-takers in terms of the amount and rate of growth. A two-level hierarchical linear model was fitted to the data from three time points. Time had a significant effect. Three out of the four LCs (1, 2, and 4) contributed significantly to accounting for the variances in initial status and growth rate. We concluded that the test is capable of reflecting changes in language ability as a function of both the length of time elapsed between repeated measures, and between-subject variations in relation to learning. The findings therefore provide support for the claim that the test can be used to monitor growth for young English language learners. This lays open the possibility that information on score change can be incorporated, thereby enriching the content of score reports for test takers who take the test multiple times.

Parallel Session 18.1: Friday, June 6, 15:10

Presenter: Ari Huhta and Charles Alderson

Affiliation: University of Jyväskylä

Title: Diagnosing the development of reading in a second or foreign language at different CEFR levels

We report on a project on diagnosing reading in a second or foreign language (SFL) and its development, and how the CEFR has been utilized in this research. The project was an international 4-year (2010-2013) study that sought to deepen our understanding of SFL development and of the factors that affect it by identifying the cognitive, affective and linguistic features which predict a learner’s strengths and weaknesses in reading. The research comprised a by cross-sectional study with several hundred SFL learners and a 2-3 year longitudinal study of a somewhat smaller number of learners. The project brought together scholars from different theoretical orientations including language testing, applied linguistics, and L1 reading development and its problems. We first describe how the CEFR influenced the overall design of the study, the selection of data gathering instruments and the rating of performances. We then focus on the key findings of the longitudinal study of the development of reading in two groups of young learners (aged 9 to 13): (1) about 150 learners of a foreign language and (2) about 60 learners of a second language. We examined how different cognitive, linguistic and motivational tasks measured at the beginning of the study predict reading and its development in these two different languages and language learning contexts. Depending on the number of cases available, the predictive regression models were obtained either via structural equation modeling or via regular regression analyses. The longitudinal study involved the measurement of learners’ SFL writing and vocabulary skills, several aspects of motivation to learn and use the SFL, language learning aptitude, non-verbal intelligence, and reading strategies. Therefore, it was possible to study how changes in these skills / areas related to changes in the ability to read in the SFL. We discuss the structure and content of the predictive models of reading that were identified in the above analyses and the relationships between different aspects of language and learning (reading, writing, vocabulary, motivation, aptitude, intelligence, strategies) at different stages of learning, as well as how these stages might relate to the CEFR levels. Finally, we discuss the implications of the findings for a theory of the diagnosis of SFL reading and for elaborating the definitions of reading levels in the CEFR.
**Parallel Session 18.2: Friday, June 6, 15:10**

**Presenter:** Ching-Ni Hsieh  
**Affiliation:** Educational Testing Service  
**Title:** Using expert judgments to assess content representation of a young learner assessment, TOEFL Primary

Reporting content validity of newly developed tests is important because this type of validity evidence provides test users essential information regarding the extent to which test contents reflect the target constructs being measured and the validity of the inferences drawn from the test scores (D’Agostino, Karpinski, & Welsh, 2011). This test validation study aimed to provide content validity evidence for the new language assessment—TOEFL Primary—a test designed for young learners ages between eight and twelve who are learning English in English as a Foreign Language (EFL) contexts. The test focuses on core communication goals and enabling knowledge and skills represented in various EFL curricula. A panel of 17 experienced EFL teachers, representing 15 countries, participated in the study. The teachers completed a survey that was designed to facilitate the evaluation of the relevance and importance of the knowledge, skills, and abilities (KSAs) assessed by the test items. Content Validity Indices (CVIs) (Popham, 1992) was used to determine the degree of match between the test contents and the target constructs. Interrater agreement was used to determine the extent to which the panelists agreed that the KSAs assessed by the test were important for young EFL learners’ effective classroom performances (Davis, 1992). Results of the study showed that the majority of the test items had a CVI above the cut-off value of .90 and over 80% of the KSAs assessed were rated as important for effective classroom performances. The results suggested that the test content substantially represented the target constructs being measured and reflected the KSAs taught in different EFL classrooms—i.e. the target language use domains. Findings of the study provide validity evidence for the content representation of TOEFL Primary and have important implications for the development of language assessments for young learners.

**Parallel Session 19.1: Friday, June 6, 16:10**

**Presenter:** Erwin Tschirner and Olaf Bärenfänger  
**Affiliation:** University of Leipzig  
**Title:** Relative levels of difficulty of different question types across levels of proficiency in reading assessments

There are many research questions at the intersection between SLA and assessment that need to be addressed to provide a foundation for a potential universal framework. The more assessment is concerned with real-life linguistic abilities, the more it needs to take into account what is known about how language abilities develop. Seeking common ground between three tried and proven systems of scaling language proficiency needs to take into account what is known about language and about how languages are learned. This paper will address one such question related to reading proficiency, i.e., the relationship between various question types eliciting various types of reading strategies or reading depth. Assessments of reading proficiency commonly distinguish between global, detail, selective and inference questions. These types of questions may be interpreted using the parameters established by Weir and Khalifa (2008): global vs. local text comprehension and expeditious vs. careful reading approaches. The goal of the present study is to look at the relative difficulty of test items at various levels of a reading test that rates reading proficiency according to the ACTFL Proficiency Guidelines 2012. The results of 12 forms of a reading test consisting of a total of 900 items that had been administered to 1,708 test takers were re-analyzed to determine relative difficulty levels of four question types: global, detail, selective, and inference questions across five ACTFL proficiency levels ranging from Intermediate Low to
Superior. IRT-based test equating procedures had been used to make sure that all forms of the test were equally difficult. A one factor ANOVA revealed significant differences in the means of the four item groups with $F = 11.346$ (df = 3), $p < .001$. Levels tested were IL, IM, AL, AM, S. Global and detail questions were used at all five levels under consideration, selective questions were used at the lowest levels only (IL) and inference questions were used only at the two highest levels (AM and S). Both, classical separation indexes and calibrated Rasch item difficulty values show that IL testees found global questions the most difficult, followed by detail and selective questions. At the levels IM, AL, and AM global questions proved to be just as difficult as detail ones. At AM in addition, there was very little difference in difficulty between global, detail and inference questions. At Superior, detail and inference questions were similar to each other with respect to difficulty level, whereas global questions seemed to be somewhat easier. Overall, selective questions were found to be the easiest and inference questions the most difficult. At all levels except Superior, difficulty indexes were very similar for global and detail questions. One reason for this similarity may be the ACTFL guidelines’ emphasis on aligning text and task levels, e.g. requiring all Advanced level tasks to focus on information that is spread out across several sentences. This may suggest that aligning text and task levels may be a feature that should be considered carefully for a universal framework.

**Parallel Session 19.2:** Friday, June 6, 16:10

**Presenter:** Wai Lan, Winnie Shum Yuanyuan and Gwendoline Guan

**Affiliation:** Lingnan University

**Title:** Exploring the facets of feedback in a diagnostic listening assessment

Research in the field of second language (L2) listening proficiency has identified factors that are important to learner’s listening performance. Corresponding strategies specific to listening comprehension were therefore proposed in the belief that it could help learners to improve their listening ability. Vandergrift’s (1997) framework of listening strategies proposed three main types of strategies: metacognitive strategies, cognitive strategies and socio-affective strategies. Among the three strategies, metacognitive strategies are the ones that learners can be most conscious of. It is also suggested that metacognition plays an important role in many language-related cognitive activities (Flavell, 1976). Al-Alwan et.al (2013) believe that metacognitive strategies are employed to manage the overall learning process. Various studies have also shown that there is a positive influence of metacognitive strategies on learners’ language ability (e.g., Vandergrift, 2003; Chamot, 2005). Interestingly, Liu’s (2008) study showed that the choice of listening strategies, more specific metacognitive strategies, is related to learner’s learning styles. In addition, more proficient listeners were likely to be more flexible than less proficient listeners in their learning style. However, a detailed account of relating specific metacognitive strategies to specific learning styles was not reported. In diagnostic assessment, providing test-takers with useful and informative feedback that they can act on to improve their language proficiency is always a challenging but important focus. What feedback or evidence can we provide so learners can get a better picture of what to do with their diagnosed strengths and weaknesses? In a pilot study of the current research, relationships between listening sub-skills tested in the Diagnostic English Language Tracking Assessment (DELTA) and listening metacognitive strategies were first explored. It was found that it is possible to relate specific metacognitive strategies to specific listening sub-skills. In the present study, an attempt was made to investigate the relationship between listening proficiency, listening sub-skills, metacognitive strategies and learning styles. DELTA was used to test around 500 students’ listening proficiency. The metacognitive Awareness Listening Questionnaire (MALQ) by Vandergrift et al (2006) and the Learning Style Survey (Cohen et al, 2013) were also administered after the test. The relationships among the three sets of data were analyzed. Follow-up interviews were also conducted to give a more comprehensive understanding of the data. Hopefully, this study can shed some light on the exploration of how feedback of a diagnostic assessment can provide
The current study investigates random assignment of partial cross-over ratings as a way to reduce the impact of rating biases while maintaining reliability in assessing test takers’ speaking performance. Numerous studies in recent years have identified various rating biases, e.g., priority of criteria (Eckes, 2008), rater’s first language (Johnson & Lim, 2009), and favoring of certain accents (Winke & Gass, 2012). However, research on how to reduce the effect of rater bias remains scarce. The study by Elder, Barkhuizen, Knoch, and von Randow (2006) suggested that successful initial and on-going training and calibration may not substantively reduce bias. The CELPIP-G Test is a large scale, standardized, assessment of general functional communicative ability. CELPIP-G scores are used by the Canadian federal government to inform decisions about individuals’ immigration and citizenship status. The revised CELPIP-G test is computerized. Responses to the speaking test are rated through a centralized online rating system that is accessed by certified raters located across Canada. The revised CELPIP-G test has eight tasks. The eight tasks will be divided into 3 sets of 4 responses. Three raters will be randomly assigned to rate the three sets of the responses. The raters will be rating by distance through a secured online interface. Responses to tasks 1-4 (set-1) will be single rated by rater 1, and responses to tasks 5-8 (set-3) will be single rated by rater 3. Responses to tasks 3-6 (set-2) will be rated by rater 2. Hence, responses to tasks 3 and 4 will be double rated by raters 1 and 2 independently, and responses to tasks 5 and 6 will be double rated by raters 2 and 3. This double rating aspect of the design ensures that (1) agreement between raters can still be evaluated, (2) each test taker will receive 12 ratings from 3 raters - four more ratings than the 8 ratings allowed by a traditional single rating design, (3) reliability will be at least as high as the single rating design if not as high as a double rating design, and (4) each test taker is rated by three independent raters. It is expected that as well as reducing bias by increasing the number of raters, the random assignment of raters can potentially guard against biases. Data were collected through a pilot test in October 2013 and can be investigated under three conditions of rating assignment: single, double, and partial cross-over designs. The rating demand, rating effect, and score reliability will be compared across the three conditions. The demand of the rating will be recorded as estimated total number of hours. The rating effect and score reliability will be analyzed using methods of variance components and dependability coefficient under generalizability theory.
Abstracts June 6

human-rated according to rubrics involving a considerable degree of subjectivity. Additionally, over the last
decade, researchers such as McNamara have paid increasing attention to the social dimension of language
testing, raising concerns over the fairness and justice of assessment. A potential example of bias in perceptions
of non-native speakers in the business world may be found in customer perceptions of call center
representatives located abroad. Even in the case of representatives who are highly proficient speakers of
English, many "inner-circle" English speakers may claim to have difficulty understanding their instructions.

Given the fact that rater perceptions of differences in pronunciation, intonation, and lexical choice have been
shown by Gass and others to impact ratings of speaking proficiency, arguably the fairest test of a speaker’s
ability to relay information in English is not by subjective opinions of performance, but rather the successful
and expedient conveyance of information. In this study, the researchers devised a task-based, tablet-computer-
mediated, online test of “Objective Communicative Ability.” In a testing session, an examinee is presented with
information (a word, picture, or audio lecturette) on his/her tablet, which is to be explained to a rater who is
unaware of what has been presented to the examinee. On the rater’s tablet appear several options from which
to choose, based on the explanation provided by the examinee. When the rater selects an answer, it is written
to the server along with the time to completion, and the next item is pushed to the tablets, following a complex
pseudorandom pattern ensuring equal coverage. The fact that the rater is unaware of the correct response,
and is only required to complete the task successfully using the examinee’s instructions, mitigates the influence
of rater subjectivity from the measurements. However, a challenge of such a time-based scoring system is that
raters could exhibit differences in time to completion for reasons unrelated to language proficiency, resulting in
construct-irrelevant variance. For this reason, the researchers administered the test to a sample of native
English speakers in order to establish scoring norms, and subsequently to a sample of Japanese university
students with known TOEFL scores. The times were re-expressed via log transformation and standardized,
allowing native speakers’ mean time to serve as the highest rating on a 6-step rating scale from the mean
through 4 standard deviations slower. These ratings were then entered into a many-facet Rasch model for
scoring. Time-based scores reliably separated student examinees into ability strata as indicated by their TOEFL
scores, with even the most proficient non-native examinees (TOEFL ≥675) attaining scores more than 1
standard deviation slower than the native speaker mean. The results indicate that the test method shows
promise as an objective, easily interpretable measure of communicative ability. The talk will conclude with a
discussion of limitations and avenues for future research.

Parallel Session 20.1: Friday, June 6, 16:45

Presenter: Veronica Benigno and Katarzyna Pawlak

Affiliation: Pearson

Title: The use of academic collocations in essays in a test of academic English

Use of vocabulary by language learners has been extensively investigated within the broader area of lexical
acquisition, second language teaching, and language assessment. In recent decades there has been a shift from
a view of vocabulary knowledge as the ability to produce single words to a view of vocabulary knowledge as the
ability to use words appropriately in combination with other words and according to their pragmatic value. The
awareness of the crucial role played by formulaic language in contributing to the fluency of language learners
(Nattinger & DeCarrio, 1992; Ellis, 2002; Wray, 2002), supported by findings in psycholinguistics and
neurolinguistics (Jackendoff, 2002), has led to the proliferation of studies investigating both the theoretical
nature of phraseological units and the impact of their use on language performance. The fortunate
simultaneous development of corpus linguistics has created very fertile ground for studies on different aspects
of phraseology. Due to their idiosyncratic nature, collocations in particular have captured the attention of many
scholars. Many have pointed out that vocabulary knowledge is not only related to the size of vocabulary known
by speakers but also to its depth. A range of measures are commonly employed to measure lexical richness, an umbrella term including lexical variation, sophistication, density and accuracy (Read, 2000). For example lexical diversity, i.e. the number of different types used in a text, is traditionally measured by the type-token ratio (TTR), and more recently, by more sophisticated indexes (e.g. the HD-D measure, McCarthy and Jarvis, 2007) aimed at minimizing the effect of text length on the score. However, the drawback of these measures is their strong quantitative component, while the qualitative rating of test takers’ use of vocabulary is generally left to the judgement of expert raters (deciding, for example, whether test takers choose the right lexical item or whether their lexical choice is appropriate to the context). This paper investigates the use of academic collocations in 50,000 test takers’ essays in order to establish whether there is a relationship between their use of collocations and their proficiency. Our aim is to identify both quantitative and qualitative features discriminating between proficiency levels (B1 to C2), using the ACL (Academic Collocations List) developed by Ackermann and Chen (2013) as a reference list. This list includes 2,500 collocations extracted using PICAE, the Pearson International Corpus of Academic English. The list was compiled using a mixed approach which combines statistical analysis with experts’ judgement to identify a set of collocations for pedagogical use. In conclusion, our study represents an attempt to bring a more qualitative insight into assessment of vocabulary. We argue vocabulary is a complex sub-construct involving multiple dimensions, which therefore needs an integrated assessment model using multiple measures, especially qualitative measures of vocabulary use. This investigation of the use of academic collocations in a test of academic English seeks to contribute to these measures.

**Parallel Session 20.2: Friday, June 6, 16:45**

**Presenter:** Sible Andringa and Punya Hoek

**Affiliation:** University of Amsterdam

**Title:** Interactions between listener and text characteristics in second-language listening comprehension

Listening comprehension is the construction of a mental model of the spoken message and listening comprehension tests attempt to gauge the success of this process. Two factors determine whether comprehension is successful: the listener and the text that is listened to. This is reflected in the literature. On the one hand, there is a body of research that has investigated what skills and knowledge listeners require to be successful comprehenders (e.g., Andringa et al., 2012); on the other hand, there is research that has investigated what text characteristics influence a text’s comprehensibility (e.g., Revesz & Brunfaut, 2013). In addition, language testers have investigated how task characteristics influence test results, acknowledging that the characteristics of the task play a mediating role between the listener and the text (e.g., Buck & Tatsuoka, 1998; Kostin, 2004). The goal of the present study was to investigate how listener, text and task characteristics interact. To this end, 113 learners of Dutch as a second language were tested on a multiple choice listening discourse comprehension task and on several measures of linguistic (vocabulary knowledge and several other subskills of listening) and cognitive capacity (working memory capacity, IQ, non-verbal processing speed). In addition, for each listening text, a number of text characteristics were identified that are potentially predictive (measures of phonological, lexical and grammatical quality). Also, several task characteristics were operationalized (plausibility of distractors, the use of low frequency words, and explicitness of the correct answer in the text). Multilevel logistic modeling was then used in which both participants and items were jointly modelled as random effects. Significant effects were found for 1) listener, 2) text and 3) task variables. The odds of a correct response was found to be associated with 1) more vocabulary knowledge and higher IQ; with 2) higher lexical frequency, lower lexical density, lower grammatical complexity, lower speech rates, higher levels of explicitness, and with location of target information in the text; and with 3) the plausibility of the distractors. Interactions were also present between listener characteristics on the one hand, and text and
task features on the other. Two examples: First, there was an interaction between explicitness and IQ: people of high IQ were more likely to provide a correct answer when resolution of the task hinged on the ability to process implied rather than explicitly stated information in the text. Second, an interaction between Vocabulary and IQ suggested that having more vocabulary was particularly helpful when resolution of the task hinged on processing paraphrased information. Thus, this study was more than a replication of findings obtained within each of the three research traditions outlined above because the interactions obtained provide insights into why certain effects obtain. Such insights not only deepen our understanding of the listening comprehension process, they are also valuable tools in the hands of test constructors in that they provide a better handle on how to align texts and tasks with the skill that one intends to measure.

**Parallel Session 20.3**: Friday, June 6, 16:45

**Presenter**: Keita Nakamura

**Affiliation**: Eiken Foundation of Japan

**Title**: Examination of possible consequences of a new test of English for university entrance in Japan

Bachman and Palmer (2010) discuss the increasing importance of test developers being accountable to stakeholders. They argue that accountability means being able to demonstrate to stakeholders that the intended uses of an assessment are justified. In terms of the timing of data collection to demonstrate test validity, Briggs (2004) and Weir (2005) stressed the importance of doing so both before and after the official implementation of the test, in order that a test can be justified with respect to its design before its administration and its intended uses following administration. However, for practical and political reasons, it is usually hard to conduct a baseline study before the test is officially administered. This paper will describe an effort to collect data to validate a new test of English in Japan before its official administration. The possible consequences of the test, which is intended for university entrance purposes, were examined by collecting data from various stakeholders at various stages of test development. As a means of data triangulation, questionnaire survey studies were conducted with 3rd-year high school students, high school teachers, university lecturers of English, and incoming freshman university students. Overall, 3838 high school students, 423 high school teachers, 19 university lecturers, and 2633 freshman university students took part in this study. For the first three of these stakeholders—high school students, high school teachers, and university lecturers—5 types of questions were asked. These included questions about perceptions of university study and the English language skills required at university, current practices in high school English classes, and the washback effects of the current university entrance exams. Freshman university students were asked whether the test tasks used in the new test were covered in their high school studies, and whether the tasks also appeared in the other entrance exams they had experienced. Participant’s responses showed that the tasks covered in the new test mostly reflect the English skills required at university. The results also showed that disproportionate emphasis was placed on the practice of reading in high school, a skill which is a main requirement in most university entrance exams. Yet, it was also found that if English speaking and writing skills become mandatory in university entrance exams, high school teachers and students would start working on the practice of such skills in the classroom. In addition, the study also found that high school teachers felt social pressures in relation to the number of their students entering prestigious universities each year. Questions which asked about the innovative features of the new test, such as the number of times the test is offered each year, elicited some differences among the stakeholder groups. Overall, the results suggested that all stakeholders would generally support the test design of the new test. The studies also provided data that will act as a baseline which can be compared with future validation studies of the test. The implications are discussed in terms of the actual method to conduct baseline study as a part of the overall validity argument.
Abstracts June 6

Parallel Session 20.4: Friday, June 6, 16:45

Presenter: Ghazal Akhavan Masoumi and Karim Sadeghi

Affiliation: Urmia University

Title: Optimal number of options for multiple-choice tests across gender

The most important concern for testers throughout the history of language testing has been to find ways that will measure the quantity of a learner’s language knowledge in as unbiased and as fair manner as possible. While various options and solutions have been offered by scholars and researchers, the challenge still goes on since no body clearly knows the best type of test for measuring skills and components of language. However, multiple-choice tests are preferred mostly in educational assessment because of their objectivity, ease of scoring, and reliability (Haladyna, 2004), but the number of options for multiple-choice tests is a challenging and unsolved issue. Despite numerous studies done about optimal number of options for multiple-choice tests, few studies are conducted in EFL/ESL context (Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006) and especially in terms of examining vocabulary knowledge. This study aimed to investigate this issue by looking at the interrelationships between number of options as well as gender. To this end, researchers utilized an adapted Key English Test (KET) and 4 versions of (3, 4, 5, and 6-option) a stem-equivalent multiple-choice vocabulary test which differed only in the number of options. Initially, the adapted KET was administered to 205 pre-intermediate fourth grade high-school male and female students in order to homogenize them and later researchers administered four versions of the stem-equivalent multiple-choice test with 3, 4, 5, and 6-options to four groups of male and four groups of female students (one group, one format). Data from 194 students were analyzed through SPSS and results of one-way ANOVA suggested that altering number of options in multiple-choice tests does not affect item discrimination; however, the results of another one-way ANOVA showed significant differences between item facilities of 3, 5, and 6-option and 4, 5, and 6-option but not between 3 and 4-options, suggesting that 6-option test is the most difficult test. There were significant but moderate correlations between gender and performance of students in multiple-choice tests, and analysis of mean scores showed that males outperform females in multiple-choice tests. Results of the study demonstrate that 3-option multiple choice is optimal, since it has similar psychometric properties to others. Also it is easier to write 2 distractors than 3, 4 and 5 distractors and 3-option test takes less time to administer. The analysis also showed that there seems to be a gender bias in doing multiple-choice tests, since males scored better female in all versions of the test. It is suggested that combination of test formats be used to reduce gender bias. Haladyna, Thomas M. (2004). Developing and validating multiple-choice test items. Mahwah, N. J.: Lawrence Erlbaum Associates. Shizuka, Tetsuhito, Takeuchi, Osamu, Yashima, Tomoko, & Yoshizawa, Kiyomi. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. Language Testing, 23(1), 35-57. doi: 10.1191/0265532206lt319oa

Plenary Session 7: Friday, June 6, 17:20

Presenter: Mikyung Kim Wolf and Alexis Lopez

Affiliation: ETS

Title: Comparing the Performance of K-12 English Language Learners and Native English Speakers on Assessment Tasks

In the United States, public schools at Grades K-12 are mandated to assess the English language proficiency of the students whose home language is not English. Based on the assessment results, the students are classified as “English language learners (ELLs)” and entitled to receive appropriate instructional services including ESL or
bilingual programs. Misidentification of ELLs entails substantial consequences for individual students’ academic paths and use of school resources. Thus, it is vitally important to set a valid cut-off for students’ performance on English language proficiency (ELP) assessments to determine accurate ELL status. This study is part of a large-scale project developing a technology-enhanced ELP assessment to identify ELL students at Grades K-12. The large-scale project developed a number of technology-enhanced assessment tasks to better measure students’ English language proficiency needed in school settings and conducted multiple prototyping studies investigating how the tasks functioned for both ELL and non-ELL students. Utilizing the data collected from these prototyping studies, the present study aimed to examine the performance of ELL students and their native English-speaking peers in order to identify key differences between these groups. It is anticipated that the findings will provide valuable information for the refinement of the tasks as well as a future standard-setting study for the project. In this study, the items and tasks were designed for multiple scenarios to assess both independent and integrated language skills (e.g., listening, reading, listen-speak, read-write, etc.). These scenario-based tasks were intended to better elicit students’ communicative language abilities in school settings with increased student engagement and simulated communicative environments. The tasks were divided into the six grade levels: Grades K, 1, 2, 3-5, 6-8, and 9-12. Regarding the data, the responses from a total of 249 students were examined including 141 ELLs and 108 native English-speaking students. The data included students’ scores on selected-response questions in listening and reading comprehension tasks as well as scores from spoken and written responses. A sample of students also participated in cognitive labs where researchers interviewed students individually. Other students participated in small-group administration with researchers taking detailed notes. Descriptive statistics of ELL and non-ELL students’ scores were first compared. In addition to the scores, the spoken and written responses were closely analyzed with respect to the amount and diversity of the language use at the lexical, syntactic, and discourse levels. The observation notes and interview transcripts were analyzed, focusing on students’ engagement and any difficulties students encountered in the technology-enhanced tasks. Preliminary findings indicate that the advanced-level ELL students sometimes performed better on certain items than average non-ELL students. Particularly, selected response scores on listening and reading comprehension were similar for advanced-level ELL and non-ELL students. However, spoken and written responses displayed greater discrimination between advanced ELL and non-ELL students. In this presentation, major distinctions between ELL and non-ELL students’ performances will be presented, along with their test-taking behaviors. Given the conference theme toward a universal framework, this study will provide useful insight into defining ELL students in K-12 school settings.
## Wednesday June 4

### Room Auditorium

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00-09:05</td>
<td>Opening of LTRC 2014 by Prof. Michel ter Hark, Dean of the Faculty of Arts</td>
</tr>
<tr>
<td>09:05-09:45</td>
<td><strong>Keynote Session: The British Council John Trim Lecture</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Douglas Biber:</strong> Predictable patterns of linguistic variation across proficiency levels and textual task types</td>
</tr>
<tr>
<td>09:45-10:15</td>
<td>John Read</td>
</tr>
<tr>
<td></td>
<td>The Potential of Vocabulary Measures in Informing Proficiency Frameworks</td>
</tr>
</tbody>
</table>

### Room HG-05A05

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:20-10:50</td>
<td>Hyejeong Kim</td>
</tr>
<tr>
<td></td>
<td>Using a general [...] for radiotelephony communication?</td>
</tr>
<tr>
<td></td>
<td>Cathie Elder</td>
</tr>
<tr>
<td></td>
<td>Tensions between general [...]languages in Australia</td>
</tr>
<tr>
<td></td>
<td>Yi Mei</td>
</tr>
<tr>
<td></td>
<td>Scoring Fairness in [...]Matriculation English Test (NMET)</td>
</tr>
<tr>
<td></td>
<td>Gerriet Janssen</td>
</tr>
<tr>
<td></td>
<td>Building a better [...] academic writing proficiency</td>
</tr>
</tbody>
</table>

### Coffee break

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:25-11:55</td>
<td>Anthony Green</td>
</tr>
<tr>
<td></td>
<td>No descriptor available: plugging gaps in the CEFR</td>
</tr>
<tr>
<td></td>
<td>Yan Jin</td>
</tr>
<tr>
<td></td>
<td>Developing the Common [...]and Micropolitical Levels</td>
</tr>
<tr>
<td></td>
<td>Wei Yan</td>
</tr>
<tr>
<td></td>
<td>Determinants of Chinese [...]Korean universities</td>
</tr>
<tr>
<td></td>
<td>Yuanyuan Gu</td>
</tr>
<tr>
<td></td>
<td>Using MCQs and [...]listening comprehension processes</td>
</tr>
</tbody>
</table>

### Room HG-02A05

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00-12:30</td>
<td>Junghee Byun</td>
</tr>
<tr>
<td></td>
<td>The L1-L2 [...]the Same Construct?</td>
</tr>
<tr>
<td></td>
<td>Lieve De Wachter</td>
</tr>
<tr>
<td></td>
<td>The construction and [...]to academic achievement.</td>
</tr>
<tr>
<td></td>
<td>Beverly Baker</td>
</tr>
<tr>
<td></td>
<td>Investigating Language Assessment [...]Canadian University Admissions</td>
</tr>
</tbody>
</table>

### Lunch break (Language Testing Editorial Board Meeting)

### Room HG-01A05

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00-14:30</td>
<td>Kinnie Kin Yee Chan</td>
</tr>
<tr>
<td></td>
<td>Computer vs. Human [...]Students’ English Essays</td>
</tr>
<tr>
<td></td>
<td>Kirby Grabowski</td>
</tr>
<tr>
<td></td>
<td>Describing pragmatic [...]speaking test</td>
</tr>
<tr>
<td></td>
<td>Symposium 1</td>
</tr>
<tr>
<td></td>
<td>Gary Ockey</td>
</tr>
<tr>
<td></td>
<td>Critical Issues in Assessing L2 Listening Ability: Construct and Consequential Validity</td>
</tr>
</tbody>
</table>

### Room HG-05A06

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00-15.40</td>
<td>Yan Jin</td>
</tr>
<tr>
<td></td>
<td>Developing the Common [...]and Micropolitical Levels</td>
</tr>
<tr>
<td></td>
<td>Wei Yan</td>
</tr>
<tr>
<td></td>
<td>Determinants of Chinese [...]Korean universities</td>
</tr>
<tr>
<td></td>
<td>Yuanyuan Gu</td>
</tr>
<tr>
<td></td>
<td>Using MCQs and [...]listening comprehension processes</td>
</tr>
</tbody>
</table>

### Room HG-05A05

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:35-15:05</td>
<td>Bart Deygers</td>
</tr>
<tr>
<td></td>
<td>Using the CEFR [...]rating scale.</td>
</tr>
<tr>
<td></td>
<td>Chaitanya Ramineni</td>
</tr>
<tr>
<td></td>
<td>Impact of Response [...]Spoken Responses</td>
</tr>
<tr>
<td></td>
<td>Symposium 2</td>
</tr>
<tr>
<td></td>
<td>Tobias Haug</td>
</tr>
<tr>
<td></td>
<td>Issues and challenges related to signed language assessment</td>
</tr>
</tbody>
</table>

### Room HG-05A05

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:10-15:40</td>
<td>Lorena Llosa</td>
</tr>
<tr>
<td></td>
<td>The Role of [...]Content Assessments</td>
</tr>
<tr>
<td></td>
<td>Megan Montee</td>
</tr>
<tr>
<td></td>
<td>Computerized oral [...]two contexts</td>
</tr>
</tbody>
</table>

### Coffee break

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:45-16:05</td>
<td>Xiangdong Gu</td>
</tr>
<tr>
<td></td>
<td>A Pilot Study [...]to the CEFR</td>
</tr>
<tr>
<td></td>
<td>Jonathan Schmidgall</td>
</tr>
<tr>
<td></td>
<td>The impact of [...]of oral proficiency</td>
</tr>
<tr>
<td></td>
<td>Maria Stathopoulou</td>
</tr>
<tr>
<td></td>
<td>Shuttling between languages of research results</td>
</tr>
<tr>
<td></td>
<td>Hye Won Shin</td>
</tr>
<tr>
<td></td>
<td>Measuring L2 vocabulary [...]knowledge scale tasks</td>
</tr>
</tbody>
</table>

### Room HG-05A05

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:45-17:15</td>
<td>Colin Finnerty</td>
</tr>
<tr>
<td></td>
<td>Balancing statistical evidence [...]of language ability.</td>
</tr>
<tr>
<td></td>
<td>Rachel Brooks</td>
</tr>
<tr>
<td></td>
<td>Diagnosing Proficiency and Performance Abilities</td>
</tr>
<tr>
<td></td>
<td>Maria Tagarelli De Monte</td>
</tr>
<tr>
<td></td>
<td>The development of [...]perspectives and issues.</td>
</tr>
<tr>
<td></td>
<td>Michelle Raquel</td>
</tr>
<tr>
<td></td>
<td>The Excel@English [...]after diagnostic testing</td>
</tr>
</tbody>
</table>

### Room Auditorium

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>17:20-17:50</td>
<td>Carolyn Turner</td>
</tr>
<tr>
<td></td>
<td>Process and performance [...]for French-speaking nurses</td>
</tr>
<tr>
<td>17:50-18:00</td>
<td><strong>Poster Pitches 1-10</strong></td>
</tr>
</tbody>
</table>

---

118
<table>
<thead>
<tr>
<th>TIME</th>
<th>ROOM</th>
<th>SESSION</th>
<th>SPEAKERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00-09:45</td>
<td>Room Auditorium</td>
<td>Keynote Session 2: Cambridge/ILTA Distinguished Achievement Award Lecture</td>
<td>Cyril Weir: Three lessons from the historiography of language testing</td>
</tr>
<tr>
<td>09:45-10:15</td>
<td>Room HG-05A05</td>
<td>John H.A.L. de Jong</td>
<td>Standards &amp; Scaling</td>
</tr>
<tr>
<td>10:20-11:05</td>
<td>Room HG-05A05</td>
<td>Work in Progress Session</td>
<td>Meg Malone, Jacob Barrows, Simon Dawson, Rie Koizumi, Ute Knich and John Pill, Troy L. Cox, Simona Sunara, Yuichiro Yokouchi, Claudia Harsch, Yehbeen Yang, Payman Vafaee, William Pellowe, Bill Bonk, Toshihiko Shiotzu, Yumiko Moore, Bart Deygers, Kris Buyse, Ying Zheng, Ying Zheng, Timothy Farnsworth, Romula Guedez-Fernandez, Maggie Dunlop, Daniel Reed, Robin Stevens, Roxanne Wong; Candace Farris, Bjorn Norrbom, Ricky Jeffrey, Glyn Jones, Nathaniel Owen</td>
</tr>
<tr>
<td>11:05-11:25</td>
<td>Room HG-05A05</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11:30-12:30</td>
<td>Room HG-05A05</td>
<td>Hui Feng</td>
<td>Teacher-Student Rapport</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bill Bonk</td>
<td>Beyond pronunciation and prosody and accentedness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gad Lim</td>
<td>On Standards and [...]Standard Setting Exercises</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Johanna Motteram</td>
<td>Appropriate ‘tone’ and [...] letter tasks and ideology.</td>
</tr>
<tr>
<td>12:00-12:30</td>
<td>Room HG-05A05</td>
<td>Liying Cheng</td>
<td>What Assessment Tasks [...]Classroom Assessment Environment?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Summer Loomis</td>
<td>Investigating ”Advanced” in [...]L2 Arabic Speakers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spiros Papageorgiou</td>
<td>Towards a universal [...] frameworks and standards</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sathena Chan and Chihiro Inoue</td>
<td>Reading-into-writing [...] and scale development</td>
</tr>
<tr>
<td>14:00-14:30</td>
<td>Room HG-05A05</td>
<td>Adam Gesicki and/or Jake Stone</td>
<td>Classification Consistency and [...]Canadian Immigration Application</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rachel Brooks</td>
<td>Comparing Native and [...]Government Speaking Tests</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eunice Eunhee Jang</td>
<td>A Mixed Methods [...]and Goal Orientations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guoxing Yu</td>
<td>Test takers’ cognitive [...] of test comparability</td>
</tr>
<tr>
<td>14:35-15:05</td>
<td>Room HG-05A05</td>
<td>Sathena Chan</td>
<td>Collecting context and [...] Level Writing Test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Troy Cox</td>
<td>Assessing language proficiency [...] proficiency exam scores</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Youngsoo So</td>
<td>Influences of the [...]speaking task performance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rebecca Present-Thomas</td>
<td>Enriching the Framework [...] developmental linguistic features</td>
</tr>
<tr>
<td>15:10-15:40</td>
<td>Room HG-05A05</td>
<td>Daniel Elias</td>
<td>‘Weak’ and ‘strong’ [...] Asian call centres</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jessica R W Wu</td>
<td>Development and Validation [...]Taiwanese EFL Learners</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mikyung Kim Wolf</td>
<td>Developing and Validating [...]for English Learners</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mehdi Riazi</td>
<td>Modeling real-life [...] assessing writing proficiency</td>
</tr>
<tr>
<td>15:45-16:05</td>
<td>Room Auditorium</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>16:10-16:40</td>
<td>Room HG-05A05</td>
<td>Ikkyu Choi</td>
<td>A Survey of [...]by Multiple Raters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cecilie Hamnes Carlsen</td>
<td>External and internal [...] and language test scores.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chihiro Inoue</td>
<td>Measuring spoken language [...] complexity and accuracy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Michelle Raquel</td>
<td>Exploring the use [...] diagnostic writing test</td>
</tr>
<tr>
<td>16:45-17:15</td>
<td>Room Auditorium</td>
<td>Talia Isaacs</td>
<td>Modelling comprehensibility in an oral production scale for L2 learners of English: Which linguistic factors generalize across L1s?</td>
</tr>
<tr>
<td>17:20-17:50</td>
<td>Room Auditorium</td>
<td>John Pill</td>
<td>Beyond the limits of universal criteria: Using indigenous criteria in assessing clinical communication in ESL</td>
</tr>
<tr>
<td>17:50-18:00</td>
<td>Room Auditorium</td>
<td>Poster Pitches 11-19</td>
<td></td>
</tr>
<tr>
<td>19:00-Late</td>
<td></td>
<td>IN DE WAAG: Banquet</td>
<td></td>
</tr>
</tbody>
</table>
### Friday June 6

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:15-10:00</td>
<td><strong>Keynote Session 3: Samuel J. Messick Memorial Lecture</strong>&lt;br&gt;&lt;br&gt;<em>Denny Borsboom:</em> Validity as the realist views it</td>
</tr>
<tr>
<td>10:00-10:30</td>
<td><strong>James Purpura</strong>&lt;br&gt;Tracking the development of learning outcomes in a learning-oriented approach to classroom-based language assessment</td>
</tr>
<tr>
<td>10:35-11:05</td>
<td><strong>Gina Park</strong>&lt;br&gt;Investigating text genre [...] differential bundle functioning</td>
</tr>
<tr>
<td>11:05-11:30</td>
<td><strong>Coffee break</strong></td>
</tr>
<tr>
<td>11:30-12:35</td>
<td><strong>Poster Session in the Hallway near the Auditorium</strong>&lt;br&gt;&lt;br&gt;<strong>Natalia de Andrade Raymundo,</strong> <strong>Benjamin Kremmel,</strong> <strong>Pablo Garcia Gomez,</strong> <strong>Samantha Musser,</strong> <strong>Caroline Riches,</strong> <strong>Ying Zheng,</strong> <strong>Lyle Bachman,</strong> <strong>Tom Velghe,</strong> <strong>Hui-Chun Yang,</strong> <strong>India Plough,</strong> <strong>Sibylle Plassmann,</strong> <strong>Susan Sheehan,</strong> <strong>Yuanyuan Guan,</strong> <strong>Roxanne Wong,</strong> <strong>Amery Wu,</strong> <strong>Margaret van Naerssen,</strong> <strong>Daniel Reed,</strong> <strong>Katharina Kley,</strong> <strong>Maggie Dunlop</strong> and <strong>Maryam Wagner</strong></td>
</tr>
<tr>
<td>12:35-14:00</td>
<td><strong>Lunch break (Language Testing Quarterly Editorial Board Meeting)</strong></td>
</tr>
<tr>
<td>14:00-14:30</td>
<td><strong>Yasuyo Sawaki</strong>&lt;br&gt;A qualitative analysis [...] for test preparation</td>
</tr>
<tr>
<td>14:35-15:05</td>
<td><strong>Limeng Zhang</strong>&lt;br&gt;Gender Differences in [...] Chinese College Students</td>
</tr>
<tr>
<td>15:10-15:40</td>
<td><strong>Ari Huhta</strong>&lt;br&gt;Diagnosing the development [...] different CEFR levels</td>
</tr>
<tr>
<td>15:45-16:05</td>
<td><strong>Coffee break</strong></td>
</tr>
<tr>
<td>16:10-16:40</td>
<td><strong>Erwin Tschirner</strong>&lt;br&gt;Relative levels of [...] in reading assessments</td>
</tr>
<tr>
<td>16:45-17:15</td>
<td><strong>Veronica Benigno</strong>&lt;br&gt;The use of [...] of academic English</td>
</tr>
<tr>
<td>17:20-17:50</td>
<td><strong>Mikyung Kim Wolf</strong>&lt;br&gt;Comparing the Performance of K-12 English Language Learners and Native English Speakers on Assessment Tasks</td>
</tr>
</tbody>
</table>
## Reviewers

| Lyle Bachman | Ardeshir Geranpayeh | James Purpura |
| Beverly Baker | April Ginther | David Qian |
| Jayanti Banerjee | Kirby Grabowski | John Read |
| Khaled Barkauoi | Tony Green | Dan Reed |
| David Belgar | Liz Hamp-Lyons | Yasuyo Sawaki |
| Vivien Berry | Luke Harding | Rob Schoonen |
| Bill Bonk | Claudia Harsch | Toshihiko Shiotsu |
| David Booth | Talia Isaacs | Masanori Suzuki |
| Sarah Briggs | Eunice Jang | Lynda Taylor |
| Tineke Brunfaut | Dorry Kenyon | Randy Thrasher |
| Gary Buck | Youn-Hee Kim | Dina Tsagari |
| Nathan Carr | Ute Knoch | Carolyn Turner |
| Micheline Chalhoub-Deville | Yong-Won Lee | Alan Urmston |
| Yeonsuk Cho | Young-Ju Lee | Alistair VanMoere |
| Martyn K. Clark | Gad Lim | Elvis Wagner |
| Christine Coombe | Lorena M. Llosa | Sara C. Weigle |
| Deborah Crusan | Sari Luoma | Bert Weltens |
| Alister Cumming | Meg Malone | Gillian Wigglesworth |
| Fred Davidson | Tim McNamara | Paula Winke |
| John De Jong | John Norris | Mikyung K. Wolf |
| Nivja de Jong | Gary Ockey | Guoxing Yu |
| Barbara Dobson | Spiros Papageorgiou | Bok Yung |
| Dan Douglas | Lia Plankans | Ying Zheng |
| Jamie Dunlea | India C. Plough | |
| Janna Fox | Rebecca Present-Thomas | |

## Volunteers

| FreiJa Camps | Nerie Kamphuis | Amber Witsenburg |
| Donja Ceelie | Jasmin Keijzer | |
| Rianne Dolmans | Aurora Mazzocchi | |
| Shanta Duijvelaar | Selena Schipper | |
| Mariëlle Fieret | Mandy de Waal | |
## Index

### A
- Akhavan Masoumi, Ghazal, 116
- Alderson, Charles, 109
- Andringa, Sible, 114

### B
- Bachman, Lyle, 10, 90
- Baker, Beverly, 24, 38
- Bärenfänger, Olaf, 110
- Battie, Aaron Olaf, 112
- Benigno, Veronica, 113
- Biber, Douglas, 15
- Bonk, Bill, 51, 65
- Booth, David, 54
- Borsboom, Denny, 85
- Brau, Maria, 36
- Brooks, Rachel, 36, 72
- Brunfaut, Tineke, 93
- Buyse, Kris, 54
- Byun, Junghee, 23

### C
- Carlsen, Cecilie Hamnes, 80
- Chan, Kinnie Kin Yee, 26
- Chan, Sathena, 70, 74
- Cheng, Lijing, 68, 102
- Choi, Ikkyu, 75, 80
- Cox, Troy L., 42, 45, 74

### D
- Dawson, Simon, 43
- De Jong, John H.A.L., 41
- De Monte, Maria Tagarelli, 37
- De Rijk, Emma, 54
- De Wachter, Lieve, 23
- Deygers, Bart, 29, 53
- Dixon, Daniel, 10
- Dunlop, Maggie, 52, 90
- Dunlop, Margaret, 72

### E
- Elder, Cathie, 17, 84, 87
- Elias, Daniel, 77
- Evans, Jeremy, 42

### F
- Farnsworth, Timothy, 56
- Farris, Candace, 60
- Feng, Hui, 64
- Finnerty, Colin, 35
- Friginal, Eric, 89

### G
- Garcia Gomez, Pablo, 91
- Geranpayeh, Ardeshir, 66
- Gesicki, Adam, 71, 112, 119
- Goh, Christine, 107
- Grabowski, Kirby, 26
- Green, Anthony, 20, 35
- Gu, Xiangdong, 32
- Guan, Yuanyuan, 22, 92
- Guédez-Fernandez, Romulo, 57
- Gujord, Ann-Kristin Helland, 80

### H
- Han-Ting Liu, Heidi, 51
- Hamp-Lyons, Liz, 104
- Harsch, Claudia, 105
- Hartig, Johannes, 105
- Haug, Tobias, 28
- Hauser, Peter, 28
- Heeren, Jordi, 23
- Hoek, Punya, 114
- Hsieh, Ching-Ni, 110
- Huhta, Ari, 109

### I
- In’nami, Yo, 11, 44
- Inoue, Chihiro, 70, 81
- Isaacs, Talia, 83
- Isbell, Daniel, 88

### J
- Jang, Eunice Eunhee, 72, 87
- Janssen, Gerriet, 19
- Jeffrey, Ricky, 62
- Jin, Yan, 20
- Jones, Glyn, 63

### K
- Katz, Anne, 91
- Kim, Heekyoung, 97
- Kley, Katharina, 93
- Knoch, Ute, 17, 44
- Koizumi, Rie, 44
- Koo, Jungyeon, 23
- Kremmel, Benjamin, 93

### L
- Lake, J., 50
- Lee, Chia-Lung, 77
- Lim, Gad, 66
- Lin, Shwu-wen, 73
- Llosa, Lorena, 30
- Lockwood, Jane, 77, 82
- Loomis, Summer, 68
- Lopez, Alexis, 116

### M
- Malone, Meg, 31, 42
- Mann, Wolfgang, 28
- Marques, Natalia Guerreiro, 97
- May, Lyn, 44
- McNamara, Tim, 16
- Mei, Yi, 18
- Meier, Valerie, 19
- Montee, Megan, 31, 94
- Moore, Yumiko, 52
- Motteram, Johanna, 67
- Murray, Jill, 79
- Musser, Samantha, 94

### N
- Nakamura, Keita, 115
- Norrbom, Bjorn, 61

### O
- Ockey, Gary, 27
- Ohlrogge, Aaron, 58
- Owen, Nathaniel, 64

### P
- Palmer, Adrian, 10, 90
- Papageorgiou, Spiros, 69
- Park, Gina, 87
- Pawlak, Katarzyna, 113
- Pellowe, William, 50
- Pill, John, 44, 84
- Plassmann, Sibylle, 96
- Plough, India, 96
- Powers, Donald, 108
- Present-Thomas, Rebecca, 76
- Purpura, James E., 86

### R
- Ramineni, Chaitanya, 2
- Raquel, Michelle, 37, 82
- Raymundo, Natalia de Andrade, 97
- Read, John, 16
Index

Reed, Daniel, 58, 97
Riazi, Mehdi, 79
Riches, Caroline, 98
Sadeghi, Karim, 116
Sawaki, Yasuyo, 106
Schmidgall, Jonathan, 32
Schmitt, Norbert, 16
Schönström, Krister, 28
Sheehan, Susan, 99
Shin, Hye Won, 34
Shiotsu, Toshihiko, 51
Shore, Jane, 78
Shum, Winnie, 111
So, Youngsoon, 75
Stathopoulou, Maria, 12, 33
Steele, Jeffrey, 46
Stevens, Robin, 59
Stewart, Jeffrey, 112
Stone, Jake, 71, 101, 112
Sunara, Simona, 46
Swartz, Carl, 60
T
Trofimovich, Pavel, 83
Tschirner, Erwin, 110
U
Urmston, Alan, 92
V
Vafaee, Payman, 49
Van Gorp, Koen, 29
Van Naerssen, Margaret, 95
Velghe, Tom, 99
Volkov, Alex, 112
W
Wagner, Maryam, 25, 90
Wai, Lan, 111
Weigle, Sara, 89
Weir, Cyril, 40
Wolf, Mikyung Kim, 78, 116
Wong, Roxanne, 60, 100
Woodson, Sara, 76
Wu, Amery, 71, 101
Wu, Jessica, 77
Wu, Yongfei, 68
Y
Yan, Qiaozhen, 32
Yan, Wei, 21
Yang, Hui-Chun, 101
Yang, Yehbeen, 48
Yokouchi, Yuichiro, 47
Yoon, Su-Youn, 29
Yu, Guoxing, 73
Z
Zhang, Limei, 107
Zheng, Ying, 54, 55, 102
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.


Copyright © 2014 by Educational Testing Service. All rights reserved. ETS, the ETS logo, LISTENING. LEARNING. LEADING., TOEFL, TOEFL Junior, TOEFL iBT and TOEFL YL are registered trademarks of Educational Testing Service (ETS) in the United States and other countries. TOEFL PRIMARY™ is a trademark of ETS. 26149
Progress!

The first ever online, fully automated and institutionally administered progress test.

- Objectively and accurately measure students’ progress throughout their course
- Fully automated speaking and writing sections
- Measure small increments in progress on the Global Scale of English

Save time and see results. See for yourself:  
www.pearsonelt.com/progress

Follow the conversation!  
#testingprogress