Table of Contents

Welcome from Conference Host – Cambridge ESOL 3
LTRC Cambridge 2010 Conference Organising Committee 4
Message from Chair of LTRC Cambridge 2010 Organising Committee 4
Acknowledgements 6
Address of ILTA Business Office 6
Proposal Reviewers 7
Award Winners 8
Award Committees 9
TIRF Awards for the Language Assessment Research Priority 10
New BAAL SIG for Testing, Evaluation and Assessment (TEA) 10
SAGE Online Access 10
About ILTA 12
Message from the ILTA President 13
Sponsors and Exhibitors 16
LTRC 2011 – University of Michigan 23
Programme Overview 25
Full Programme 30
The Messick Award 44
UCLES/ILTA Lifetime Achievement Award 47
Abstracts 51
Pre-Conference Workshops 51
Symposia 54
Research Papers 61
Posters 90
Works in Progress 100
Presenter Index 111
Additional Information for Delegates 114
In Memoriam – Caroline Clapham 118
Notes 121
Maps 126
EnglishProfile
Setting professional standards for English language learning worldwide

English Profile is a ground-breaking research programme which is shaping the future of English language learning, teaching and assessment worldwide.

Endorsed by the Council of Europe, English Profile will provide a unique benchmark for progress in English by clearly describing the language learners need at each level of the Common European Framework (CEFR).

- Coming in 2010, English Profile Journal - a new online journal publishing the latest research related to the English Profile programme.
- Coming in 2011, English Profile Wordlists - a fully-searchable listing of words and phrases in English at each level of the CEFR, starting with levels A1-B2.

“This project brings to the description of learner levels insights gained through new technology, as well as the combined wealth of dozens of practitioners from all over the world. It is truly innovative and radical in its approach.”

Professor Michael McCarthy, University of Nottingham and English Profile Journal Editor

To get involved visit www.englishprofile.org and find out how.
Welcome from the Conference Host – Cambridge ESOL

Dr Michael Milanovic, Cambridge ESOL

On behalf of the University of Cambridge and the Local Organising Committee in my own department, Cambridge ESOL Examinations, I am very pleased to welcome the 32nd Language Testing Research Colloquium to Cambridge.

Rather I should say “welcome back” to the many old friends who are here again for the second LTRC in Cambridge.

The last time we had the honour of hosting this event was in the Summer of 1993 when the 15th LTRC was hosted in Cambridge and in Arnhem (Netherlands). This was a double first; it was the first time it was held outside North America, and the first (and only time) it was held in two countries!

The subsequent internationalisation of LTRC has been a major success, and since 1993 the event has been held in nine countries as far apart as Finland and Australia. I think this is testimony to the worldwide interest in language testing and the importance now placed on high quality tests and professional standards. This is a very important aim for our University and we are very pleased to support this trend.

Over the last 17 years we have seen a shift of emphasis in our field, so that the technical aspects of testing which were a main focus in the 1980s and early 1990s are now balanced alongside a growing concern for the social and educational aspects of language assessment.

In the Local Organising Committee we wanted to come up with a theme which could encompass this broad perspective and allow for a multi-faceted conference with something to appeal to everyone, no matter what angle you come from.

The title we chose, Crossing the threshold: investigating levels, domains and frameworks in language assessment most obviously conjures up the literal notions of language levels and proficiency frameworks – such as the CEFR. This is clearly represented in the programme and in the pre-conference workshops. I am pleased to note, however, that papers have also been accepted which treat this theme in a more figurative sense, where “crossing the threshold” refers to the crossing of a physical boundary (e.g. entry into a country or into a profession), and where the domain or framework in question forms part of a social or political system in which language and language assessment play an important part. In this respect too, I think the programme promises to be very varied and no doubt stimulating.

I wish you a very pleasant and successful stay in Cambridge and I look forward to meeting many of you during the conference.

Michael Milanovic
CEO, University of Cambridge ESOL Examinations
LTREC Cambridge 2010 Conference Organising Committee

Conference Organisers
Dr Lynda Taylor (Chair), University of Cambridge ESOL Examinations (Consultant)
Dr Michael Milanovic, University of Cambridge ESOL Examinations (CEO)
Dr Nick Saville, University of Cambridge ESOL Examinations
Sarah Brydie, University of Cambridge ESOL Examinations
Keri Ridout, University of Cambridge ESOL Examinations
Xi Xiaoming – ILTA Liaison, Educational Testing Service
Sara Cushing Weigle – ILTA Liaison, Georgia State University
Robert Ranieri, Prime Management
Anna Trott, Prime Management

Message from Chair of LTRC Cambridge 2010 Organising Committee

Dear friends and colleagues,

As anyone who has ever been involved knows, organising our annual Language Testing Research Colloquium (LTRC) can be a complex and time-consuming endeavour. This year was no exception, and the large and unexpected increase in proposals submitted for the 2010 conference made it even more of a challenge. People sometimes wonder how decisions are made when scheduling the event and selecting the programme content. This year, in the interests of transparency and accountability, I thought it might be helpful to summarise some of the behind-the-scenes planning and decision-making that went on this year which we trust will result in an enjoyable and stimulating conference for all of us.

The LTRC 2010 Call for Proposals was launched at the March 2009 event in Denver and appeared on the conference website in April 2009. In that Call for Proposals the Local Organising Committee provided some guidance on the chosen theme, identifying potential questions of interest, highlighting thematic strands and suggesting related sub-themes. Our intention was to offer a broad background canvas with plenty of scope for interpretation but with the aim of achieving a reasonably coherent and cohesive programme. We also offered guidance on what differentiates a symposium from a session of research papers, in the hope that this would avoid the frustration that can arise when a symposium becomes simply a sequence of loosely connected, individual research papers with little time for plenary discussion. We included in the Call for Proposals explicit evaluation criteria, based partly on those used by AAAL, to assist colleagues in conceptualising and crafting their proposal. With an eye on publishing a potential conference proceedings after the event, we also invited people to theme-code their abstract proposals on submission to help us get ahead with the advance planning of such a volume.

By the proposal submission deadline on 30 September 2009, we had received a total of 298 abstracts. Though we imagined Cambridge might be an attractive venue, we had certainly not anticipated a 50% increase on submissions from the previous year! The overall submission figure breaks down as follows: 11 symposia, 159 research papers, 76 works-in-progress (WiPs) and 52 posters. This necessitated assembling a team of 40 willing members from right across the international language testing community to peer review and grade up to 16 mixed abstracts each over a period of 3–4 weeks. The LTRC 2010 reviewer team is listed on page 7 and we should be grateful to them for their service to our community. It was particularly good to be able to welcome some first-time reviewers to the team and each of these was paired with a more experienced LTRC reviewer.

When assigning abstracts for review at the outset, we took care to avoid any potential conflict of interest; reviewers themselves were also asked to alert us to any case of this on receipt of their
allocation but in only one instance did a reviewer ask us to reassign one of the abstracts allocated to them. Every research paper, WiP and poster submission was blind reviewed and graded independently by 2 reviewers. Each of the 11 symposia, however, was reviewed and graded by 4 independent and experienced reviewers. The whole process was done online; all reviewers were sent a set of guidelines, which reiterated the evaluation criteria together with the usual 4-point rating scale used for grading LTRC abstracts: 1 = Reject; 2 = Possibly accept if space permits; 3 = Good, accept; 4 = Outstanding, definitely accept. Given the large number of submissions for LTRC 2010, reviewers were encouraged to use the full length of the rating scale to help discriminate between stronger and weaker proposals.

Following completion of this stage, reviewer grades were averaged to generate a final overall grade for each proposal. All proposals were then rank ordered according to this final grade, i.e. 4, 3.5, 3, 2.5, etc, and also by presentation category, i.e. symposia, research papers, WiPs and posters. The highest scoring proposals in each of the 4 categories were considered eligible for inclusion in the programme. At this point it became clear that adopting the plenary-only format for LTRC 2010 meant there would not be space to include some high-quality proposals in the programme.

For this reason the Local Organising Committee (in consultation with the ILTA Board) decided to schedule several parallel research paper and symposium sessions on the Thursday and Friday. This allowed us to increase the number of research paper slots from about 22 to 40, and to increase the number of symposium slots from 2 to 4. Our venue facilities meant we could accommodate around 20 posters and 18 WiPs. We were sensitive to the fact that there is a longstanding affection for the LTRC plenary format and there have been mixed reactions to parallel sessions at past LTRCs. However, as well as allowing as many people as possible to present and take part in the event, we believed this decision would enable us to offer a broad and stimulating conference program for all participants, who now come to LTRC from many parts of the world and represent a growing range of professional interests.

In order to compile a provisional programme we began by accepting all proposals which had scored an overall average grade of 4 and 3.5. We added to these the ‘safe 3’ proposals, i.e. those where both graders had assigned a 3 to the abstract. Thus all the slots for posters and WiPs, and most of the research paper slots, were filled with such proposals. The remaining research paper slots were filled with proposals averaging a 3 overall but graded as a 2 and 4; these were also selected for content, i.e. they were chosen to match with thematic clusters of research papers and to ensure a balanced programme overall. This means that all research papers, WiPs and posters in the programme scored 3 or above. The 4 highest scoring symposia were also accepted; these scored 3.5, 3.25, 2.75 and 2.75 respectively (remember that all symposium proposals were evaluated by a team of 4 experienced reviewers).

When all the available programme slots had been provisionally filled, acceptance notifications were sent out in early December with an invitation to present. At the same time, a waiting list was drawn up for each presentation category of all the remaining proposals scoring 3. Notifications were sent out to those placed on a waiting list and to those whose proposals were not accepted. In the event, almost everyone accepted the invitation to present, and in only a couple of cases was it possible to extend an invitation to someone on a waiting list.

Once presenters had confirmed their intention to attend and present, the 40 research papers were grouped according to topic/theme and were scheduled as sensibly and as sensitively as possible. The four symposia were paired up – one with a slightly more social/cultural orientation, the other with a slightly more technical/linguistic focus. We made efforts this year to enhance the pre-conference support and in-conference profile of all those presenting WiPS and posters.

Choosing what pre-conference workshops and related events to offer at LTRC is always a challenge. This year’s programme was selected to address some contemporary issues in the field arising from our conference theme, as well as to meet some professional needs. It was informed by discussions during LTRC 2009 in Denver inviting suggestions and offers for consideration. A draft briefing document was developed to help future LTRC organisers and those wishing to offer a pre-conference workshop to jointly think through the logistical aspects involved. This year the Local Organising Committee was keen to repeat the very successful Welcome to LTRC Newcomers session (first done in Denver) and to offer a new Meet the Journal Editors venture for delegates. Thanks to the generous
sponsorship of various benefactors, it has proved possible to offer several of the pre-conference professional development opportunities to delegates at subsidised rates or free of charge.

I hope that what I have written above helps to explain a little of the background and decision-making that underpins this year’s LTRC. This printed conference programme should contain all you need to help you navigate your way though what I sincerely hope will be a stimulating and rewarding few days – not just professionally but also personally. I look forward to the opportunity of meeting up with many of you in face-to-face encounters throughout the week. Enjoy!

Dr Lynda Taylor  
Chair – LTRC 2010 Local Organising Committee

Acknowledgements

Programme

Cover design: University of Cambridge ESOL Examinations  
Programme production: University of Cambridge ESOL Examinations

Website

LTRC Cambridge 2010 website: University of Cambridge ESOL Examinations

Address of ILTA Business Office

ILTA Business Office  
Prime-Management, Inc  
3416 Primm Lane  
Birmingham, AL 35216 USA

Tel: +44-1-205-823-6106 (from the UK)  
Fax: +44-1-205-823-2760 (from the UK)  
ILTA@primemanagement.net (Robert Ranieri)  
www.iltaonline.com
Proposal Reviewers

Jayanti Banerjee  University of Michigan, US
Viven Berry  Independent Language Testing Consultant
Rachel Brooks  Federal Bureau of Investigation, US
Annie Brown  Ministry of Higher Education, UAE
Nathan Carr  California State University, Fullerton, US
Micheline Chalhoub-Deville  University of North Carolina, Greensboro, US
Martyn Clark  University of Oregon, US
Christine Coombe  Dubai Men’s College, Arab Emirates
Alastair Cumming  OISE, University of Toronto, Canada
Alan Davies  University of Edinburgh, UK
John H.A.L. de Jong  VU University Amsterdam, Netherlands
Craig Deville  Measurement Incorporated, US
Barbara Dobson  University of Michigan, US
Janna Fox  Carleton University, Ottawa, Canada
Evelina Galaczi  University of Cambridge ESOL Examinations, UK
Ardeshir Geranpayeh  University of Cambridge ESOL Examinations, UK
Anthony Green  University of Bedfordshire, UK
Liz Hamp-Lyons  University of Nottingham, UK
Hanan Khalifa  University of Cambridge ESOL Examinations, UK
Talia Isaacs  McGill University, Montreal, Canada
Ute Knoch  University of Melbourne, Australia
Yong-Won Lee  Seoul National University, South Korea
Lorena Llosa  New York University, US
Sari Luoma  Ballard and Tighe Publishers, US
Margaret Malone  Center for Applied Linguistics, Washington, US
Tim McNamara  University of Melbourne, Australia
Spiros Papageorgiou  University of Michigan, US
India Plough  University of Michigan, US
James Purpura  Teachers College, Columbia University, US
David Qian  Hong Kong Polytechnic University
Nick Saville  University of Cambridge ESOL Examinations, UK
Rob Schoonen  University of Amsterdam, Netherlands
Toshihiko Shiotsu  Kurume University, Japan
Elana Shohamy  Tel-Aviv University, Israel
Lynda Taylor  University of Cambridge ESOL Examinations, UK
Randy Thrasher  Okinawa Christian University, Japan
Carolyn Turner  McGill University, Montreal, Canada
Elvis Wagner  Temple University, US
Xiaoming Xi  Educational Testing Service (ETS), Princeton, US
Guoxing Yu  University of Bristol, UK

Given the unprecedented volume of proposals received for LTRC 2010, the Local Organising Committee is most grateful to all those listed above who gave their time so willingly to the process of reviewing the abstracts.
Award Winners

2008 ILTA Best Article Award

Elvis Wagner, Temple University
*Video listening tests: What are they measuring?* (Language Assessment Quarterly, 5/3)

2009 IELTS Masters Award (awarded in 2010)

Ruslan Suvorov, Department of English, Iowa State University of Science and Technology, USA
Thesis adviser: Volker Hegelheimer
Thesis title: *Context visuals in L2 listening tests: The effectiveness of photographs and video vs audio only format*

2010 Jacqueline A. Ross TOEFL Dissertation Award

Alistair Van Moere, University of Lancaster
Dissertation adviser: Professor Charles Alderson
Dissertation title: *Group oral tests: How does task affect candidate performance and test scores?*

2010 Spaan Fellowship for Studies in Second or Foreign Language Assessment

Beverly Baker, McGill University
*Use of the cloze-elide task in high-stakes English proficiency testing*

Dina Tsagari, University of Cyprus
*The influence of the examination for the Certificate of Proficiency in English (ECPE) upon test preparation materials*

Hongli Li, Pennsylvania State University
*Cognitive diagnostic analysis of the MELAB Reading Test*

Ching-Ni Hsieh, Michigan State University
*Experienced raters’ vs naïve American undergraduates’ judgments of L2 oral performance: accentedness, comprehensibility, and oral proficiency*

2010 ILTA Student Travel Award

Heike Neumann, McGill University
*What does it take to make a grade? Teachers assessing grammatical ability in L2 writing with rating scales*

Youngsoon So, UCLA
*The dimensionality of test scores on a second language reading comprehension test: Implications for accurate estimation of a test-taker’s level of L2 reading comprehension*

Weiwei Liu, The Hong Kong Polytechnic University
*A multi-dimensional approach to the exploration of rater variability: The impacts of gender and occupational knowledge on the assessment of ESP oral proficiency*
Award Committees

2008 ILTA Best Article Award
John Read (chair), *University of Auckland*
April Ginther, *Purdue University*
Constant Leung, *Kings College London*
Miyuki Sasaki, *Nagoya Gakuin University*

2010 Selection Committee for the Robert Lado Award for Best Graduate Student Paper at LTRC
Lorena Llosa (chair), *New York University*
Carol Lynn Moder, *Oklahoma State University*
Spiros Papageorgiou, *University of Michigan*
Nick Saville, *University of Cambridge ESOL Examinations*

2009 IELTS Masters Award Committee
Submissions for this award are reviewed by the IELTS Joint Research Committee which comprises members of the three partner organisations (Cambridge ESOL, British Council and IDP:IELTS Australia) and is chaired by Nick Saville. Evaluations are also sought from external academic specialists in language testing and assessment.

2010 Jacqueline A. Ross TOEFL Dissertation Award
The Jacqueline Ross TOEFL Dissertation Award recipient is selected by a panel of three independent language testing experts. A different panel is constituted for each year’s award competition, as determined by members of the TOEFL Committee of Examiners. The membership of each panel includes a previous award recipient, a current member of the TOEFL Committee of Examiners, and a well-recognized language testing professional (from outside of the TOEFL Committee of Examiners) identified by the TOEFL Committee of Examiners. ETS staff provide no input into the award selection or panel composition.

The 2010 Selection Committee for the Spaan Fellowship for Studies in Second or Foreign Language Assessment
Jeff S. Johnson (chair), *University of Michigan*
Natalie Nordby-Chen, *University of Michigan*
India C. Plough, *University of Michigan*
Stephen G. Schilling, *University of Michigan*
Eric Lagergren, *University of Michigan*

2010 ILTA Student Travel Award
James E. Purpura (chair), *Teachers College, Columbia University*
Nathan Carr, *California State University, Fullerton*
Barbara Dobson, *University of Michigan*

2009 ILTA Grant Funding for Workshops and Meetings
Carolyn Turner (chair), *McGill University*
Lia Plakans, *University of Iowa*
James Purpura, *Teachers College, Columbia University*
TIRF Awards for the Language Assessment Research Priority

TIRF, The International Research Foundation for English Language Education (TIRF) is a non-profit organization whose goal is to generate new knowledge about English language teaching and learning.

It aims:
- To promote the application of research to practical language problems
- To collect, organize, and disseminate information and research on the teaching and learning of language
- To influence the formation and implementation of appropriate language education policies, taking cognizance of the complementarity of English as an international language and indigenous languages and cultures worldwide

TIRF is committed to the development of a research agenda which is flexible and dynamic, both addressing perennial issues and evolving as critical questions present themselves. Its current focus is on promoting research and best practices that will improve the use of English in the emerging global knowledge economy of the 21st century.

Cambridge ESOL and Cambridge University Press are delighted to be able to support TIRF’s work through an annual grant to support researchers working on innovative doctoral dissertations. Four projects were supported in 2009; two of these projects relate to language assessment and the recipients will be recognized at LTRC Cambridge on 15 April 2010.

Youn-Hee Kim, An Argument-Based Validity Inquiry into Empirically-Derived Descriptor-Based Diagnostic (EDD) Assessment in ESL Academic Writing

Cecilia Guanfang Zhao, The Role of Voice in High-Stakes L2 Writing Assessment

The deadline for applications for the 2010 awards is May 15, 2010. For more details of the grants and how to apply, visit www.tirfonline.org/

Supported by

New BAAL SIG for Testing, Evaluation and Assessment (TEA)

The newly formed British Association for Applied Linguistics (BAAL) special interest group for testing, evaluation and assessment will add significantly to the future development of applied linguistics by offering an added perspective on the application of applied linguistics research in language assessment, language education, and language policies in educational assessment, and through the dynamic cross-fertilization of ideas. The SIG plans to hold regular events in conjunction with the Language Testing Forum which is held annually in a British university (University of Lancaster in November 2010) and also with other language testing conferences such as IATEFL to explore areas of mutual interest and research. Membership is open to members as well as non-members of BAAL.

For more information on the BAAL TEA SIG please contact the Convenor, Dr Vivien Berry (vberry@hkucc.hku.hk).

SAGE Online Access

In celebration of the 32nd LTRC, SAGE is pleased to offer free online access to Language Testing until 30 April 2010. To receive access, register at https://online.sagepub.com/cgi/register?registration=FTLTJ
Resources to Meet Your Language and Cultural Education Needs

The Center for Applied Linguistics is dedicated to providing a comprehensive range of research-based information, tools, and resources related to language and culture.

Online Resources

Read our latest CAL Digest now available online:

Multilingual Education Policy and Practice: Lessons from Indigenous Experience

www.cal.org/digests

Visit our Web Site
www.cal.org

Browse the CAL Store
calstore.cal.org

CAL
CENTER FOR APPLIED LINGUISTICS

Improving Communication Through Better Understanding of Language and Culture
About ILTA

ILTA Goals

1. Stimulate professional growth through workshops and conferences;
2. Promote the publication and dissemination of information related to the field of language testing;
3. Develop and provide for leadership in the field of language testing;
4. Provide professional services to its members;
5. Increase public understanding and support of language testing as a profession;
6. Build professional pride among its membership;
7. Recognize outstanding achievement among its membership;
8. Cooperate with other groups interested in language testing;
9. Cooperate with other groups interested in applied linguistics or measurement

ILTA 2009 Executive Board

President: Carolyn Turner, McGill University, Montreal, Canada
Vice-President: John Read, University of Auckland, New Zealand
Secretary: Xiaoming Xi, Educational Testing Service, USA
Treasurer: Sara Cushing Weigle, Georgia State University, USA
Immediate Past President: James Purpura, Teachers College, Columbia University, USA

Members-at-Large 2010:
Nathan Carr, California State University, Fullerton, USA
Barbara Dobson, University of Michigan, USA
Yong-Won Lee, Seoul National University, South Korea
Lia Plakans, University of Iowa, USA

ILTA Committees 2010

ILTA Nominating Committee 2009
Chair: Meg Malone, Center for Applied Linguistics, USA
Ute Knoch, University of Melbourne, Australia
David Qian, Hong Kong Polytechnic University, China
Yasuyo Sawaki, Educational Testing Service, USA

ILTA Webmaster

Erik Voss, Iowa State University

ILTA Archivists

Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Craig Deville, Measurement Inc.

ILTA Task Force on Aviation Language Assessment

J. Charles Alderson (chair), Lancaster University, UK
Dan Douglas, Iowa State University, USA
Candace Farris, McGill University, Canada
Ute Knoch, University of Melbourne, Australia
Rob Schoonen, University of Amsterdam, The Netherlands
About ILTA

ILTA Institutional Affiliates

Academic Committee for Research on Language Testing, Israel (ACROLT)
Association of Language Testers in Europe, EC (ALTE)
East Coast Organization of Language Testers, USA (ECOLT)
Japan Language Testing Association, Japan (JLTA)
Korean English Language Testing Association, Korea (KELTA)
Midwest Association of Language Testers, USA (MwALT)
National College English Testing Committee, China (NCETC)
National Institute for Testing and Evaluation, Israel (NITE)
Southern California Association for Language Assessment Research (SCALAR)
TESOL Arabia Testing & Assessment Special Interest Group (TESOL Arabia TAE SIG)

ILTA Presidents

<table>
<thead>
<tr>
<th>Year</th>
<th>President</th>
<th>Year</th>
<th>President</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992–3</td>
<td>Charlie Stansfield</td>
<td>2001</td>
<td>Caroline Clapham</td>
</tr>
<tr>
<td>1993–4</td>
<td>Charles Alderson</td>
<td>2002</td>
<td>Fred Davidson</td>
</tr>
<tr>
<td>1995</td>
<td>Bernard Spolsky</td>
<td>2004</td>
<td>Antony Kunnan</td>
</tr>
<tr>
<td>1996</td>
<td>Lyle Bachman</td>
<td>2003</td>
<td>Liz Hamp-Lyons</td>
</tr>
<tr>
<td>1997</td>
<td>John Clark</td>
<td>2005</td>
<td>Dan Douglas</td>
</tr>
<tr>
<td>1998</td>
<td>Tim McNamara</td>
<td>2006</td>
<td>Glenn Fulcher</td>
</tr>
<tr>
<td>1999</td>
<td>Elana Shohamy</td>
<td>2007–8</td>
<td>James Purpura</td>
</tr>
<tr>
<td>2000</td>
<td>Alan Davies</td>
<td>2009–10</td>
<td>Carolyn Turner</td>
</tr>
</tbody>
</table>

Message from the ILTA President

Dear Friends and Colleagues,

It is a pleasure once again to be sharing with all of you our common interest of language testing and assessment. On behalf of the ILTA Executive Board, the LTRC Organizing Committee and all the members of the International Language Testing Association, I would like to welcome you to the 32nd Annual Language Testing Research Colloquium at the University of Cambridge in Cambridge, UK.

We are hosted this year by the University of Cambridge ESOL Examinations and the conference theme is “Crossing the threshold: Investigating levels, domains and frameworks in language assessment.” The response to the LTRC Call for Papers was a record-breaking 300 submissions. In order to accommodate the number of high quality proposals, there will be some parallel sessions this year. The program is therefore full of stimulating and informative papers, posters, works-in-progress and symposia. In addition LTRC has three pre-conference workshops that correspond to the conference theme on both national and international levels: The CEFR – Its purpose, origins, development and current status; The CEFR-Alignment and standard-setting issues; and Assessing levels of academic language proficiency-Conceptualization, operationalization, alignment and validation. Another “workshop” is specifically local to our conference site: A Cambridge Assessment Tour. This provides a unique opportunity to see Europe’s largest assessment agency. As we welcome an increasing amount of graduate students and young researchers, a new type of workshop is being introduced this year: Sharing your story: developing your academic conference skills. This is a valuable professional development event. Other pre-conference events open to everyone of course, but particularly aimed at those newer to the field are Meet the Editors and LTRC Newcomers’ Session.

Each year at LTRC we find it important to acknowledge those in our field who have demonstrated distinguished scholarship. These awards will be presented for the most part at the LTRC Banquet (Gala Banquet, King’s College). Others will be presented during LTRC. We will honor those receiving the following awards: UCLES/ILTA Lifetime Achievement Award, 2008 ILTA Best Article Award, 2010 IELTS Masters Award, 2010 Jacqueline A. Ross TOEFL Dissertation Award, 2010 Spaan Fellowship for Studies in Second or Foreign Language Assessment, and 2010 Robert Lado Memorial Award.
The recipient of this final award is selected during the conference. It honors the student judged as presenting the best student paper at LTRC. So don’t miss the Banquet and the Closing events where we will personally honor and congratulate these people.

As can be seen from some of the awards above, ILTA has and continues to encourage students to get involved – to be part of LTRC, to present research, and to join ILTA. We all started as students. You are the future of our community. There is an additional award to help student presenters get to LTRC, the ILTA Student Travel Award. This year ILTA is happy to have supported several students.

I would like to take this opportunity to inform you about some of ILTA’s current activities. In January we welcomed our new Executive Board members and thanked our outgoing ones. ILTA is a growing organization and we have much to do. We continue to explore the potential of our ILTA website and other possibilities with our managing company, Prime Management. Some of our aims are to stimulate professional growth, promote dissemination of information and to develop and provide leadership in our field of language assessment and testing. By re-activating the ILTA Grant Funding for Workshops and Meetings, we were able to see the fruits of that effort in two areas of the world in 2009: the first International Language Assessment Conference in Armenia (ILACA) took place in July and the first English Proficiency Assessment Summit in the Philippines took place in December. Also, ILTA is collaborating with those in the aviation sector and struck an ILTA task force to examine the feasibility, criteria and possible procedures for a voluntary accreditation scheme for tests of aviation English. This process is ongoing. Another project we are in the midst of is translating our important documents into several languages (*ILTA Code of Ethics and ILTA Guidelines for Practice*). Internally, we continue to develop and refine our set of Standard Operating Procedures (SOPs) for ILTA. As our association grows, there is the need to formally document how we function and to keep these procedures up to date. In this same spirit of growth, ILTA is presently working on a strategic plan for the future. ILTA could not function without its membership. Many of you participate and I would like to personally thank you for all of your correspondence, ideas, committee work and devotion to the values and aims of ILTA. This of course includes the hard work of the Executive Board. For much more detailed information, come to the ILTA Annual Business Meeting at lunch time on Thursday.

It is my pleasure to now recognize those who have made it possible for LTRC 2010 to take place providing us the opportunity to reunite once again as a community. The LTRC Organizing Committee has put in endless hours. Lynda Taylor, the chair, has done a superb job and has kept the community updated in an ongoing manner as preparations were underway. Much thanks also to the members of the Organizing Committee: Michael Milanovic, Nick Saville, Sarah Brydie and Keri Ridout. Sincere gratitude goes to Sara Weigle (ILTA Treasurer) and Xiaoming Xi (ILTA Secretary) who have served as special liaisons to LTRC 2010. In addition, all the conference sponsors are much appreciated for their contribution to events and activities throughout LTRC.

Please join me in thanking all of these special people and all of the volunteers not mentioned by name who have made LTRC possible. Knowing we are in good hands, we can now go forward and appreciate the motivating days we will share together in interaction and discussion concerning language testing and assessment issues.

All the best for a stimulating LTRC in Cambridge!

Carolyn E. Turner

President of ILTA 2009, 2010
New for 2010

Announcing online submission for

ASSESSING WRITING

Editor: Liz Hamp-Lyons, University of Bedfordshire, UK

With a global audience, 50,000 full-text article downloads annually and the ability to post an accepted article online within 6 weeks of acceptance, give your work every advantage by publishing it in Assessing Writing.

Assessing Writing is a refereed international journal providing a forum for ideas, research and practice on the assessment of written language. Assessing Writing publishes articles, book reviews, conference reports, and academic exchanges concerning writing assessments of all kinds, including traditional (‘direct’ and standardised forms of) testing of writing, alternative performance assessments (such as portfolios), workplace sampling and classroom assessment. The journal focuses on all stages of the writing assessment process, including needs evaluation, assessment creation, implementation, and validation, and test development; it aims to value all perspectives on writing assessment as process, product and politics (test takers and raters; test developers and agencies; educational administrations; and political motivations). The journal is interested in review essays of key issues in the theory and practice of writing assessment.

For more information on the journal or to submit an article for publication please visit:

www.elsevier.com/locate/ASW
Sponsors and Exhibitors

The Organising Committee of LTRC Cambridge 2010 thanks all sponsors and exhibitors of this conference for their generous support.

Main sponsors

Opening Reception
Sponsored by University of Cambridge ESOL Examinations

www.CambridgeESOL.org

University of Cambridge ESOL Examinations (Cambridge ESOL) offers the world’s leading range of qualifications for learners and teachers of English. Over three million people take our exams each year in 130 countries. Worldwide over 10,000 universities, employers, government ministries and other organisations rely on our certificates as proof of English language ability. Cambridge ESOL exams are underpinned by the work of the largest dedicated research team of any UK language test provider. As experts in language assessment we advise governments internationally on language education policy and are active partners in projects delivering key language objectives for the European Union.

Gala Banquet
Sponsored by Educational Testing Service (ETS)

www.ets.org

At nonprofit ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research.

Founded in 1947, ETS develops, administers and scores more than 50 million tests annually including the TOEFL® and TOEIC® tests, the GRE® General and Subject Tests and The Praxis Series™ assessments — in more than 180 countries, and at over 9,000 locations worldwide.

In addition to assessments, we conduct educational research, analysis and policy studies, and we develop a variety of customized services and products for

- teacher certification
- English-language learning
- elementary, secondary and post-secondary education

ETS serves individual students, their parents, educational institutions and government agencies. We help teachers teach, students learn, and parents measure the educational and intellectual progress of their children by

- Listening to educators, parents and critics
- Learning what students and their institutions need
- Leading in the development of new and innovative products and services
Other sponsors

Pre-Conference Workshops 1 & 2: The CEFR: its purpose, origins, development and current status & The CEFR: alignment and standard-setting issues
Sponsored by Association of Language Testers in Europe (ALTE)
www.ALTE.org

ALTE is an association of providers of foreign language examinations, founded in 1990. It includes many of the world’s leading language assessment bodies. ALTE currently has 31 members, representing the testing of 26 languages and over 40 affiliate institutions. ALTE members provide standard examination and computer based language tests.

A principal objective of ALTE is to adhere to the common levels of proficiency established by the ALTE members. Levels are aligned to the Common European Framework of Reference (CEFR) in order to promote the transnational recognition of certification in Europe.

Pre-Conference Workshop 4: Sharing your story: developing your academic conference skills
Sponsored by ILTA

LTRC Newcomers’ Session
Sponsored by University of Michigan English Language Institute (ELI-UM)
www.lsa.umich.edu/eli

The English Language Institute at the University of Michigan (ELI-UM) is a worldwide leader in language teaching, learning, and assessment, as well as in applied linguistics and teacher education. Located in downtown Ann Arbor and part of one of the leading universities in the United States, ELI-UM is internationally known for its high professional standards in the development and administration of English as a second/foreign language examinations. For over fifty years, people who pass a University of Michigan English language proficiency examination have been able to confidently tell schools and employers that they have the language skills necessary to succeed. ELI-UM also serves as a professional resource on matters related to ESL/EFL assessment, conducts research supporting language assessment, and promotes the ethical use of tests.

Main Conference Lunches
Sponsored by STEP EIKEN
www.stepeiken.org

The Society for Testing English Proficiency (STEP) is a non-profit foundation established in Japan in 1963, and is responsible for producing and administering Japan’s national English exam, the EIKEN Test in Practical English Proficiency. Over 2 million examinees take the EIKEN tests each year at approximately 18,000 locations. The tests are offered in seven levels, which span the spectrum of commonly recognized ability levels and are designed to both measure and motivate learners. EIKEN is recognized for a broad range of educational and professional uses, including for admissions and course credit at over 2,500 Japanese high schools, colleges, and universities; by all 47 prefectural boards of education, which directly operate Japan’s public high schools; by the Japanese education ministry as benchmarks for junior high school and high school graduates; and for international admissions at a growing number of institutions in the United States, Australia, Canada, and the UK.
Main Conference Refreshments
Sponsored by Second Language Testing Inc, Michigan State University and IELTS
Second Language Testing Inc
www.2LTI.com

Second Language Testing, Inc. (SLTI), is a small business with headquarters in Rockville, MD. SLTI is essentially a test development service for hire. Nearly all of the products we develop are turned over to the client who pays for the development. The mission of SLTI is to provide language test development, translation/adaptation, and other related services in the areas of second language skills testing and the testing of nonnative English-speakers. Our clients are larger test development companies, state departments of education, and state and federal government agencies.

SLTI has a staff of 17 full-time employees, including test development managers, a psychometrician, an accountant, translation managers, and desktop publishing specialists. We maintain a worldwide network of consultants, including item writers, translators, desktop publishing specialists and language specialists for over 40 languages.

Our only products are our language aptitude tests, which include the Modern Language Aptitude Test and the Pimsleur Language Aptitude Battery.

We hope that you enjoy visiting our display and that the information provided is useful. If you have any comments please e-mail us at CStansfield@2LTI.com.

Michigan State University
www.msu.edu/

IELTS
www.ielts.org

IELTS (International English Language Testing System) is a test for people who intend to study or work where English is the primary language of communication.

Experience
IELTS is tried and tested: over the past 20 years, more than six million candidates have taken the test.

Authenticity
IELTS tests authentic skills for success. We recognise the central skill our stakeholders are looking for: the ability to communicate in real-life situations.

IELTS has always used face-to-face interaction in the speaking test because it’s the most realistic option. It generates the most natural performance from candidates.

Quality
Each IELTS test is based on an ISO-compliant development process ensuring the test material is of world-leading quality.

IELTS test scores are trusted and recognised by more than 6,000 universities and colleges around the globe. The rigorous way in which we recruit and train our examiners contributes to this.

IELTS is backed by an extensive programme of funded research into the validity and impact of the test, reflected in the publication of 10 volumes of IELTS Research Reports, as well as several volumes in the CUP “Studies in Language Testing” series which focus on IELTS.

Global reach
IELTS is the international test for global mobility. It is taken by well over a million candidates each year in over 125 countries with each test centre offering unparalleled customer service and support.
Studies in Language Testing Anniversary Celebration
Sponsored by Studies in Language Testing
www.cambridgeesol.org/what-we-do/research/silt.html

The 15th anniversary of the Studies in Language Testing is being celebrated in 2010. This series of academic volumes is published jointly by University of Cambridge ESOL Examinations and Cambridge University Press (CUP), edited by Dr Michael Milanovic and Prof Cyril J Weir. The series addresses a wide range of important issues and new developments in language testing and assessment and is an indispensable resource for test users, developers and researchers. There are currently over 25 titles available. Copies of the volumes are available from booksellers or can be ordered direct from the Cambridge University Press website.

Computer & Internet Room
Sponsored by IELTS
www.ielts.org

IELTS (International English Language Testing System) is a test for people who intend to study or work where English is the primary language of communication.

Experience
IELTS is tried and tested: over the past 20 years, more than six million candidates have taken the test.

Authenticity
IELTS tests authentic skills for success. We recognise the central skill our stakeholders are looking for: the ability to communicate in real-life situations.

IELTS has always used face-to-face interaction in the speaking test because it’s the most realistic option. It generates the most natural performance from candidates.

Quality
Each IELTS test is based on an ISO-compliant development process ensuring the test material is of world-leading quality.

IELTS test scores are trusted and recognised by more than 6,000 universities and colleges around the globe. The rigorous way in which we recruit and train our examiners contributes to this.

IELTS is backed by an extensive programme of funded research into the validity and impact of the test, reflected in the publication of 10 volumes of IELTS Research Reports, as well as several volumes in the CUP “Studies in Language Testing” series which focus on IELTS.

Global reach
IELTS is the international test for global mobility. It is taken by well over a million candidates each year in over 125 countries with each test centre offering unparalleled customer service and support.

Noticeboard
Sponsored by Studies in Language Testing
www.cambridgeesol.org/what-we-do/research/silt.html

The 15th anniversary of the Studies in Language Testing is being celebrated in 2010. This series of academic volumes is published jointly by University of Cambridge ESOL Examinations and Cambridge University Press (CUP), edited by Dr Michael Milanovic and Prof Cyril J Weir. The series addresses a wide range of important issues and new developments in language testing and assessment and is an indispensable resource for test users, developers and researchers. There are currently over 25 titles available. Copies of the volumes are available from booksellers or can be ordered direct from the Cambridge University Press website.
Sponsors and Exhibitors

Cambridge University Press Bookshop Reception
Sponsored by Cambridge University Press – English Language Teaching
www.cambridge.org

Cambridge University Press is one of the world’s leading publishers of English Language Teaching materials, supplying almost every country in the world. A number of our market-leading courses and grammar books have sold in their millions since first publication. Our authoritative Applied Linguistics series presents the latest insights from work in the field with their implications for language teaching and learning. The Studies in Language Testing series, produced in conjunction with Cambridge ESOL, focuses on developments in language testing which will be of interest to test users, language test developers and researchers. Both series are of world renown. These materials exemplify the cutting edge of English teaching methodology and scholarship, and the quality of the Cambridge ELT list is universally acknowledged by the teaching profession for both British and American English.

Exhibitors

Association of Language Testers in Europe (ALTE)
www.ALTE.org

ALTE is an association of providers of foreign language examinations, founded in 1990. It includes many of the world’s leading language assessment bodies. ALTE currently has 31 members, representing the testing of 26 languages and over 40 affiliate institutions. ALTE members provide standard examination and computer based language tests.

A principal objective of ALTE is to adhere to the common levels of proficiency established by the ALTE members. Levels are aligned to the Common European Framework of Reference (CEFR) in order to promote the transnational recognition of certification in Europe.

Cambridge University Press – English Language Teaching
www.cambridge.org

Cambridge University Press is one of the world’s leading publishers of English Language Teaching materials, supplying almost every country in the world. A number of our market-leading courses and grammar books have sold in their millions since first publication. Our authoritative Applied Linguistics series presents the latest insights from work in the field with their implications for language teaching and learning. The Studies in Language Testing series, produced in conjunction with Cambridge ESOL, focuses on developments in language testing which will be of interest to test users, language test developers and researchers. Both series are of world renown. These materials exemplify the cutting edge of English teaching methodology and scholarship, and the quality of the Cambridge ELT list is universally acknowledged by the teaching profession for both British and American English.

English Profile
www.englishprofile.org

The aim of English Profile is to create a ‘profile’ or set of Reference Level Descriptions for English linked to the Common European Framework of Reference for Languages (CEFR). These will provide detailed information about the language that learners can be expected to demonstrate at each level, offering a clear benchmark for progress that will inform curricula development as well as the development of courses and test material to support learners, teachers and other professionals involved in the teaching of English as a foreign language.
Educational Testing Service (ETS)
www.ets.org

At nonprofit ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research.

Founded in 1947, ETS develops, administers and scores more than 50 million tests annually — including the TOEFL® and TOEIC® tests, the GRE® General and Subject Tests and The Praxis Series™ assessments — in more than 180 countries, and at over 9,000 locations worldwide.

In addition to assessments, we conduct educational research, analysis and policy studies, and we develop a variety of customized services and products for

- teacher certification
- English-language learning
- elementary, secondary and post-secondary education

ETS serves individual students, their parents, educational institutions and government agencies. We help teachers teach, students learn, and parents measure the educational and intellectual progress of their children by

- Listening to educators, parents and critics
- Learning what students and their institutions need
- Leading in the development of new and innovative products and services

IELTS
www.ielts.org

IELTS (International English Language Testing System) is a test for people who intend to study or work where English is the primary language of communication.

Experience
IELTS is tried and tested: over the past 20 years, more than six million candidates have taken the test.

Authenticity
IELTS tests authentic skills for success. We recognise the central skill our stakeholders are looking for: the ability to communicate in real-life situations.

IELTS has always used face-to-face interaction in the speaking test because it’s the most realistic option. It generates the most natural performance from candidates.

Quality
Each IELTS test is based on an ISO-compliant development process ensuring the test material is of world-leading quality.

IELTS test scores are trusted and recognised by more than 6,000 universities and colleges around the globe. The rigorous way in which we recruit and train our examiners contributes to this.

IELTS is backed by an extensive programme of funded research into the validity and impact of the test, reflected in the publication of 10 volumes of IELTS Research Reports, as well as several volumes in the CUP “Studies in Language Testing” series which focus on IELTS.

Global reach
IELTS is the international test for global mobility. It is taken by well over a million candidates each year in over 125 countries with each test centre offering unparalleled customer service and support.

Oxford University Press
www.oup.com/elt

Oxford University Press is committed to producing research-based testing and exam preparation materials, and to supporting research and scholarship in the testing field. Come to the OUP stand to find out about:
• Language Assessment in Practice (Lyle Bachman and Adrian Palmer): the new edition of the bestselling *Language Testing in Practice*.
• The Oxford Online Placement Test: a computer-adaptive and automatically marked test that provides a CEFR level for any student. Pretested and validated by more than 10,000 students in over 40 countries. Available online at www.oxfordenglishtesting.com.
• Online practice tests: supported practice that helps students prepare for a wide range of international and national exams.

**Pearson Language Tests**
www.pearsonpte.com

Pearson Language Tests deliver Pearson Tests of English, a suite of English language tests designed to assess the English language usage of students who are speakers of other languages. Pearson Tests of English include Pearson Test of English Academic (PTE Academic), a new computer-based academic English language test endorsed by GMAC, Pearson Test of English General (PTE General) and Pearson Test of English Young Learners (PTE Young Learners).

**Springer**
www.springer.com/

Springer is a leading global scientific publisher, delivering quality content through innovative information products and services. In the science, technology and medicine (STM) sector, we publish around 2,000 journals and more than 6,500 new books a year, as well as the largest STM eBook Collection worldwide.

Our department of *Language Education and Linguistics* publishes a full range of journals, books, and reference works, covering many aspects of applied and theoretical linguistics. We are the publishers of the *Encyclopedia of Language and Education*, and the journal *Language Policy*, with editors Elana Shohamy and Kendall King. For information about these and other publications, please visit our website Springer.com.

**University of Cambridge ESOL Examinations**
www.CambridgeESOL.org

University of Cambridge ESOL Examinations (Cambridge ESOL) offers the world’s leading range of qualifications for learners and teachers of English. Over three million people take our exams each year in 130 countries. Worldwide over 10,000 universities, employers, government ministries and other organisations rely on our certificates as proof of English language ability. Cambridge ESOL exams are underpinned by the work of the largest dedicated research team of any UK language test provider. As experts in language assessment we advise governments internationally on language education policy and are active partners in projects delivering key language objectives for the European Union.
LTRC 2011 – University of Michigan

Please join us at the University of Michigan for

LTRC 2011
Half a Century of *Language Testing*: Back to the Basics
June 21 – June 25, 2011
[Preconference workshops June 21 – June 22]

The year 2011 marks the 50th anniversary of the publication of Robert Lado’s seminal book *Language Testing*, the first textbook in the field of language assessment. *Language Testing* was written while Lado served as director of the English Language Institute at the University of Michigan, Ann Arbor, Michigan, USA.

Founded in 1817, the University of Michigan is one of the leading universities in the United States. With a student population of approximately 40,000 from more than 113 countries, the U-M is internationally renowned for its research, education, and professional schools of Law, Business, and Medicine.

Consistently named one of the best places in the United States to live, Ann Arbor, Michigan is home to a distinguished arts environment, eclectic dining, and bookstores for all tastes. Ann Arbor is conveniently located only 20 miles from Detroit Metropolitan Airport (DTW) and 40 miles from the heart of Motown.
Half a Century of Language Testing: Back to Basics

For more information please visit our website:
www.lsa.umich.edu/eli/LTRC2011
Programme Overview

12–16 April 2010 – University of Cambridge, UK

Given the unprecedented number of proposals (300) submitted for LTRC Cambridge 2010, the Local Organising Committee has decided to program several parallel sessions at this year’s conference. As well as enabling as many people as possible to present their work, this decision ensures that we can offer a broad and stimulating conference programme for all participants, who now come to LTRC from many parts of the world and represent a growing range of professional interests. As far as possible, research papers have been grouped according to a common theme/topic.

<table>
<thead>
<tr>
<th>MONDAY 12 APRIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.15 – 4.15</td>
</tr>
<tr>
<td>9.00 – 4.00</td>
</tr>
<tr>
<td>12.30 – 1.30</td>
</tr>
<tr>
<td>9.00 – 4.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TUESDAY 13 APRIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.15 – 5.30</td>
</tr>
<tr>
<td>9.00 – 4.00</td>
</tr>
<tr>
<td>9.00 – 4.00</td>
</tr>
<tr>
<td>9.00 – 4.00</td>
</tr>
<tr>
<td>2.00 – 4.30</td>
</tr>
<tr>
<td>4.10 – 5.10</td>
</tr>
<tr>
<td>5.15 – 6.15</td>
</tr>
<tr>
<td>6.30 – 8.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WEDNESDAY 14 APRIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.15 – 6.00</td>
</tr>
<tr>
<td>8.45 – 9.00</td>
</tr>
<tr>
<td>10.00 – 6.00</td>
</tr>
<tr>
<td>9.00 – 9.10</td>
</tr>
<tr>
<td>9.10 – 9.55</td>
</tr>
<tr>
<td>9.55 – 10.00</td>
</tr>
<tr>
<td>10.00 – 10.30</td>
</tr>
</tbody>
</table>
## Programme Overview

### WEDNESDAY 14 APRIL (continued)

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Type</th>
<th>Details</th>
<th>Location</th>
</tr>
</thead>
</table>
| 10.30 – 12.00 | Paper Session 1 (plenary)   | Basic language cognition and the construct of language proficiency: theoretical and empirical perspectives (Jan Hulstijn)  
A mixed methods approach to construct definition: identifying underlying factors in L2 oral interactive tasks for nurses in a minority setting (Carolyn Turner, Michel Laurier and Talia Isaacs)  
Performance Decision Trees (Glenn Fulcher, Fred Davidson and Jenny Kemp) | Auditorium   |
| 12.00 – 1.30   | Lunch break                  | Language Assessment Quarterly Editorial Board Meeting                   | Foyer & Recital Room |
| 1.30 – 3.00   | Poster Session               |                                                                        | Foyer        |
| 3.00 – 4.30   | Paper Session 2 (plenary)   | Which aspects of fluency should be taken into account when describing oral proficiency levels? (Nivja de Jong)  
Comparing human and machine judgements of collocations and relating them to speaking proficiency (Jing Xu and Xiaoming Xi)  
Turn, turn, turn again (Michael McCarthy) | Auditorium   |
| 4.30 – 4.50   | Refreshment break            |                                                                        | Foyer & Recital Room |
| 4.50 – 5.50   | Paper Session 3 (plenary)   | Fitting the framework: how raters co-construct understandings of domains and levels from different educational and assessment contexts (Alan Urmston, Felicia Fang and Carrie Tsang)  
Can we get a little higher? Proficiency levels in aviation English (Carol Lynn Moder and Gene Halleck) | Auditorium   |
| 6.15 – 7.30   | Cambridge University Press Bookshop Reception |                                                                        | CUP Bookshop |
| 7.00 – 10.00  | ILTA Executive Board Meeting, The University Centre |                                                                        | University Centre |

### THURSDAY 15 APRIL

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Type</th>
<th>Details</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.15 – 6.00</td>
<td>Conference registration</td>
<td></td>
<td>Foyer</td>
</tr>
<tr>
<td>9.00 – 6.00</td>
<td>Exhibition</td>
<td></td>
<td>Recital Room</td>
</tr>
</tbody>
</table>
| 8.30 – 10.15 | Paper Session 4 (parallel) | Session 4a:  
Assessing young learners’ writing: using complementary validity evidence to evaluate the quality of two L2 rating scales (Ari Huhta and Mirja Tamanen)  
The interaction patterns in paired assessment among young learners and their effect on performance evaluation (Yuko Butler and Wei Zeng)  
The CEFR: its virtues and shortcomings when teaching and assessing young learners (Hanan Khalifa, Angela ffrench and Szilvia Papp)  
Session 4b:  
Multicompetence and assessment: recognizing variability in the L1 benchmark (Amanda Brown)  
Linguistic characteristics of NS and NNS responses to impromptu essay prompts (Sara Weigle and Eric Friginal)  
Validation of a benchmarking study by means of repeated measures classification consistency (Rebecca Present-Thomas and Alistair van Moere) | G-R06/07 Seminar Room, Faculty of English |
| 10.15 – 10.30 | Refreshment break           |                                                                        | Foyer        |
### THURSDAY 15 APRIL (continued)

<table>
<thead>
<tr>
<th>Time</th>
<th>Session 5a:</th>
<th>Session 5b:</th>
<th>Session 5c:</th>
<th>Location</th>
</tr>
</thead>
</table>
| 10.30 – 12.10 | Using the Common European Framework to rate test taker responses (Kirsten Ackermann)  
Getting the levels right: deriving item writer guidelines from CEFR descriptors (John de Jong)  
An alternative approach to linking a suite of examinations to the Common European Framework of Reference (CEFR) (Gergely Dávid) | Vocabulary use in writing tests tasks performance (Viphavee Vongpumivitch)  
Linking TOEFL scores to the Lexile Measure (Cathy Wendler, Fred Cline and Ellie Sandford)  
Lexical resources in legal and general English writing across second language proficiency levels (Ivana Vidakovic and Fiona Barker) | Investigating participants’ understanding of proficiency levels in different educational contexts (Spiros Papageorgiou)  
User perceptions of academic language proficiency and the TOEFL iBT (Margaret Malone, Megan Montee and Chengbin Yin)  
What can a can-do do? Developing descriptors that define levels (Anthony Green) | G-R06/07 Seminar Room, Faculty of English  
Lecture Room 2  
Auditorium |

| Time       | Lunch break  
12.10 – 2.25 | Works-in-Progress Session  
2.25 – 3.40 | Refreshment break (incl. SiLT 15th Anniversary celebration)  
3.40 – 4.10 | Symposium Session A (parallel)  
4.10 – 6.10 |
|------------|-------------|---------------------|-------------------------------|----------------------------------|
|            | Foyer & Recital Room  
12.25 ILTA Annual Business Meeting | Lecture Rooms 1, 4 and 5 | Foyer & Recital Room | Auditorium |

| Time       | Symposium 1:  
The industrialisation of language testing (Organiser: Bernard Spolsky, with contributors Jared Bernstein, Fred Davidson, Liz Hamp-Lyons, Glenn Fulcher) | Symposium 2:  
Investigating the grammatical domains of meaning conveyance in language assessments and language assessment frameworks (Organiser: James Purpura, with contributors Yoko Ameriks, Jee Wha Dakin, Kirby Grabowski and Yen-Fen Liao) |
|------------|------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|
| 7.30       | Gala Banquet, Kings College  
King’s College |  
7.30 Gala Banquet, Kings College  
King’s College |

### FRIDAY 16 APRIL

| Time       | Conference registration  
8.15 – 5.00 | Exhibition  
9.00 – 4.00 | Paper Session 6 (parallel)  
8.30 – 10.15 |
|------------|-------------------|-----------------|----------------------|
|            | Foyer | Recital Room | G-R06/07 Seminar Room, Faculty of English  
Lecture Room 2 |

| Time       | Session 6a:  
A multi-dimensional approach to the exploration of rater variability: the impacts of gender and occupational knowledge on the assessment of ESP oral proficiency (Weiwei Liu)  
What does it take to make a grade? Teachers assessing grammatical ability in L2 writing with rating scales (Heike Neumann)  
Rating L2 pragmatics: what does holistic scoring really mean? (Yuanshan Chen) | Session 6b:  
Analysis of factors that influence Korean Middle School students’ English achievement by using a latent growth model (Bokyung Cho, Woogin Jung and Kyungyoon Kim)  
The problem of continuity between primary and secondary school languages programs: what can classroom-based assessment tell us? (Kathryn Hill)  
School-based oral language assessment in Hong Kong: students’ attitudes and anxiety (Manman Gao) |
|------------|-----------------------------------|---------------------------------------------------------------|------------------------------------------|
|            | G-R06/07 Seminar Room, Faculty of English  
Lecture Room 2 | Lecture Room 2 |  
Lecture Room 2 |
<table>
<thead>
<tr>
<th>Time</th>
<th>Session/Activity</th>
<th>Location</th>
</tr>
</thead>
</table>
| 8.30 – 10.15 | **Session 6c:** Assessing health literacy: a new domain for collaboration between language testers and health professionals *(Catherine Elder)*  
Using a threshold principle to decide when someone is (or is not) at a proficiency level: examples from Hindi and other less commonly taught languages *(Pradyumnna Amatya)*  
Fairness and consistency in assigning testing accommodations in large scale testing *(Jesse Black-Allen)* | Auditorium        |
| 10.15 – 10.30| **Refreshment break**                                                                                                                                                                                              | Foyer & Recital Room |
| 10.30 – 12.30| **Symposium Session B (parallel)**                                                                                                                                                                                | Auditorium        |
|              | **Symposium 3:** Exploring interactional competence in paired and group speaking tests *(Organiser: Gillian Wigglesworth, with contributors Lyn May, Evelina Galaczi, Fumiyo Nakatsuhara, Alistair van Moere, Annie Brown)* | Auditorium        |
|              | **Symposium 4:** Testing aviation English *(Organiser: Charles Alderson, with contributors Ute Knoch, Ari Huhta, Candace Farris, John Read, Dan Douglas)* | G-R06/07 Seminar Room, Faculty of English |
| 12.30 – 2.00 | **Lunch break**  
*Language Testing Editorial Board Meeting*                                                                                                                                                                   | Foyer & Recital Room |
| 2.00 – 3.30  | **Paper Session 7 (plenary):**  
Time trials: the relationship between marker speed and quality *(Paul Jaquith)*  
A multi-round validation study of descriptors-based STEP language proficiency scales *(Eunice Jang, Jim Cummins, Maryam Wagner and Saskia Stille)*  
Theoretical basis and experiments of a computer-assisted marking system on short answer questions *(Xiangdong Gu and Fanna Meng)* | Auditorium        |
| 3.30 – 3.50  | **Refreshment break**                                                                                                                                                                                              | Foyer & Recital Room |
| 3.50 – 4.50  | **Paper Session 8 (plenary):**  
Task difficulty in L2 listening assessment: an investigation of text characteristics and test takers’ perceptions of difficulty *(Tineke Brunfaut and Andrea Revesz)*  
Test taker performance on video vs audio-only L2 listening tests *(Elvis Wagner)* | Auditorium        |
| 4.50 – 5.35  | **UCLES/ILTA Lifetime Achievement Award Lecture**  
Professor Elana Shohamy                                                                                                                                                                                           | Auditorium        |
| 5.35 – 5.45  | **Closing comments**                                                                                                                                                                                              | Auditorium        |
The Impact of Language Frameworks on Assessment, Learning and Teaching explored through the perspectives of policies, procedures and challenges

7–9 July 2011

The Association of Language Testers in Europe (ALTE) will hold its 4th International Conference in Kraków, Poland.

The plenary speakers include Professor Lyle Bachman, Professor Giuliana Grego Bolli, Dr Neil Jones, Dr Waldemar Martyniuk, Dr Michaela Perlmann-Balme and Professor Elana Shohamy.

The Call for Papers will open in April 2010.

www.alte.org/2011

Interested in attending ALTE conferences and getting more involved in ALTE’s work? Then why not become an Affiliate?

Institutional Affiliate status offers among other benefits:

- listing on the ALTE website
- opportunities to collaborate with ALTE members and affiliates in Special Interest Groups
- opportunities to build international links through the ALTE network
- reduced rates for professional development courses

Individual Affiliate status is free of charge and offers among other benefits:

- advance information of ALTE events and activities
- an invitation to attend conference days
- reduced rates for professional development courses and other events
- an invitation to join the ALTE electronic discussion forums

For further details of Institutional and Individual Affiliate status, please enquire at the ALTE exhibition or visit www.alte.org
Full Programme

Workshops
Monday 12 April

9.00 am – 4.00

Workshop 1: The CEFR: its purpose, origins, development and current status
Location: Recital Room
Neil Jones, Nick Saville and Szilvia Papp, University of Cambridge ESOL Examinations
John Trim, former Director of the Council of Europe’s Modern Languages Projects
Brian North, Eurocentres
Norman Verhelst, CITO

Workshop 3: Assessing levels of academic language proficiency: conceptualization, operationalization, alignment, and validation (Day 1)
Location: Lecture Room 3
Margo Gottlieb, Tim Boals and Carsten Wilmes, WIDA Consortium
Dorry Kenyon, Center for Applied Linguistics
Fred Davidson, Youngshin Chi and Chih-Kai (Cary) Lin, University of Illinois, Urbana-Champaign

Workshops
Tuesday 13 April

9.00 am – 4.00

Workshop 2: The CEFR: alignment and standard-setting issues
Location: Recital Room
Neil Jones and Angeliki Salamoura, University of Cambridge ESOL Examinations
Brian North, Eurocentres
Norman Verhelst, CITO
Neus Figueras, Generalitat de Catalunya
Sylvie Lepage, Centre International d’Études Pédagogiques (CIEP)

Workshop 3: Assessing levels of academic language proficiency: conceptualization, operationalization, alignment, and validation (Day 2)
Location: Lecture Room 3
Margo Gottlieb, Tim Boals and Carsten Wilmes, WIDA Consortium
Dorry Kenyon, Center for Applied Linguistics
Fred Davidson, Youngshin Chi and Chih-Kai (Cary) Lin, University of Illinois, Urbana-Champaign

Workshop 4: Sharing your story: developing your academic conference skills
Location: Lecture Room 5
Lynda Taylor, University of Cambridge ESOL Examinations
Liz Hamp-Lyons, University of Nottingham

2.00 pm – 4.00

Workshop 5: Cambridge Assessment Tour
Meeting point: Foyer, Faculty of Music

<table>
<thead>
<tr>
<th>4.10 – 5.10</th>
<th>Meet the Journal Editors</th>
<th>Auditorium</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.15 – 6.15</td>
<td>LTRC Newcomers Session</td>
<td>Auditorium</td>
</tr>
<tr>
<td>6.30 – 8.00</td>
<td>Opening Reception</td>
<td>Museum of Zoology</td>
</tr>
</tbody>
</table>
### Wednesday 14 April 2010

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.45 – 9.00</td>
<td>Welcome to LTRC</td>
</tr>
<tr>
<td>9.00 – 9.10</td>
<td>Introduction to Messick Memorial Lecture</td>
</tr>
<tr>
<td></td>
<td>Nick Saville, Cambridge ESOL</td>
</tr>
<tr>
<td>9.10 – 9.55</td>
<td>Samuel J Messick Memorial Lecture:</td>
</tr>
<tr>
<td></td>
<td>Professor Michael Kane (Sponsored by the TOEFL Board, Educational Testing Service (ETS), Princeton, New Jersey)</td>
</tr>
<tr>
<td>9.55 – 10.00</td>
<td>Presentation of Messick Award</td>
</tr>
<tr>
<td></td>
<td>Cathy Wendler, ETS</td>
</tr>
<tr>
<td>10.00 – 10.30</td>
<td>BREAK (inc. conference photograph)</td>
</tr>
<tr>
<td>10.30 – 12.00</td>
<td>Paper Session 1 (plenary)</td>
</tr>
<tr>
<td></td>
<td>Chaired by: Randy Thrasher</td>
</tr>
<tr>
<td></td>
<td>Basic language cognition and the construct of language proficiency: theoretical and empirical perspectives</td>
</tr>
<tr>
<td></td>
<td>Jan Hulstijn, University of Amsterdam</td>
</tr>
<tr>
<td></td>
<td>A mixed methods approach to construct definition: identifying underlying factors in L2 oral interactive tasks for nurses in a minority setting</td>
</tr>
<tr>
<td></td>
<td>Carolyn Turner and Talia Isaacs, McGill University</td>
</tr>
<tr>
<td></td>
<td>Michel Laurier, Université de Montreal</td>
</tr>
<tr>
<td></td>
<td>Performance decision trees</td>
</tr>
<tr>
<td></td>
<td>Glenn Fulcher and Jenny Kemp, University of Leicester</td>
</tr>
<tr>
<td></td>
<td>Fred Davidson, University of Illinois at Urbana-Champaign</td>
</tr>
<tr>
<td>12.00 – 1.30</td>
<td>LUNCH</td>
</tr>
<tr>
<td></td>
<td>Language Assessment Quarterly (LAQ) Editorial Board Meeting: Lecture Room 1</td>
</tr>
<tr>
<td>1.30 – 3.00</td>
<td>Poster Session, Faculty of Music</td>
</tr>
<tr>
<td>1. Knowing the test takers: investigating test taker background, test taker characteristics, and their test performance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ying Zheng and Kirsten Ackermann, Pearson Language Tests</td>
</tr>
<tr>
<td>2. Developing an English language proficiency test for pilots</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Henry Emery, emery-roberts</td>
</tr>
<tr>
<td>3. Proficiency assessment in less commonly taught languages</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Martyn Clark, Center for Applied Language Studies</td>
</tr>
<tr>
<td>4. Promoting assessment literacy for language educators and learners: challenges, approaches and practices</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Victoria Nier, Margaret E Malone and Chengbin Yin, Center for Applied Linguistics</td>
</tr>
<tr>
<td>5. Rating guide development for Speaking’s dimensions: delivery, language use and topic development</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joan Jamieson, Northern Arizona University</td>
</tr>
<tr>
<td></td>
<td>Kornwipa Poonpon, Kon Khaen University, Thailand</td>
</tr>
</tbody>
</table>
6. Developing authentic online test tasks for oral intercultural professional communication in CEFcult
Jan Van Maele, Group T – Leuven Engineering College
Lut Baten, Leuven Language Institute, K.U. Leuven
Lugart Dusar, K.U. Leuven/Living Stone Intercultural Entrepreneurship

7. Effects of utterance speed, timing control, and repeated exposure on elicited imitation performance in Japanese as a Second Language
Dan Dewey, Hitokazu Matsushita and Kayleen Dewey, Brigham Young University

8. Developing a new theoretical framework for cognitive diagnostic assessment of second language reading
Ah-Young Kim, Teachers’ College, Columbia University

9. Validation of rating scale descriptors for International Teaching Assistants’ Oral English
So-young Jang, Jane Alsberg, Amanda Huensch and Kyeongsuk Kim, University of Illinois at Urbana-Champaign

10. Test adaptiveness in development of an oral proficiency test for International Teaching Assistants
Kyeongsuk Kim, Jane Alsberg, Amanda Huensch and So-young Jang, University of Illinois at Urbana-Champaign

11. Language policy and language proficiency assessment in English and Putonghua: is what is proposed, possible?
Jane Lockwood and Michelle Raquel, Hong Kong Institute of Education

12. Automated rating of an ESL Writing placement test: can we trust machines?
Chih-Kai (Cary) Lin, University of Illinois, Urbana-Champaign

13. The effects of the arrangement of questions in can-do self checklists on students’ responses and their ease of use
Tomoko Fujita, Tokai University

14. The European Survey on Language Competences: an update on progress
Neil Jones, University of Cambridge ESOL Examinations

15. Graphical protocol analysis: a method to measure a writer’s language competency
Putri Alza Maria Zulkifli and Peter Cheng, University of Sussex

16. Creating a multidimensional scale of language proficiency
Michelle Raquel and Jane Lockwood, Hong Kong Institute of Education

17. Language and Integration
Martin Nuttall and Nick Beresford-Knox, Association of Language Testers in Europe

18. Evaluating the New College English Test (Band 4) with a comparative approach
Min Liu and David D Qian, The Hong Kong Polytechnic University

19. Investigating accuracy and complexity across levels: towards a valid scale for the Language Proficiency Examination for Foreign Language Teachers (EPPLE) in Brazil
Vanessa Borges-Almeida, UnB – University of Brasilia
Douglas Consolo, UNESP – State University of Sao Paulo

20. Investigating CEFR levels of Japanese University students with DIALANG: comparison of results with outcomes from the DIALANG Pilot Study
Chisato Saida, Ibaraki University, Japan
3.00 – 4.00 Paper Session 2 (plenary)
Chaired by: Barbara Dobson

Which aspects of fluency should be taken into account when describing oral proficiency levels?
Nivja de Jong, Utrecht University

Comparing human and machine judgements of collocations and relating them to speaking proficiency
Jing Xu, Iowa State University
Xiaoming Xi, Educational Testing Service (ETS)

Turn, turn, turn again
Michael McCarthy, University of Nottingham

4.30 – 4.50 BREAK

4.50 – 5.50 Paper Session 3 (plenary)
Chaired by: David Qian

Fitting the framework: how raters co-construct understandings of domains and levels from different educational and assessment contexts
Alan Urmston, Felicia Fang and Carrie Tsang, Hong Kong Polytechnic University

Can we get a little higher? Proficiency levels in aviation English
Carol Lynn Moder and Gene B Halleck, Oklahoma State University

6.15 – 7.30 Cambridge University Press Bookshop Reception

7.00 – 10.00 ILTA Executive Board Meeting, The University Centre
Thursday 15 April 2010

NOTE PARALLEL SESSIONS ACROSS MULTIPLE VENUES TODAY

8.30 – 8.35 Announcements [will be made in all 3 parallel session venues]
8.35 – 10.15 Paper Session 4 (parallel)

Paper Session 4a
Location: G-R06/07 Seminar Room, Faculty of English
Chaired by: Vivien Berry

Assessing young learners' writing: using complementary validity evidence to evaluate the quality of two L2 rating scales
Ari Huhta and Mirja Tamanen, University of Jyväskylä

The interaction patterns in paired assessment among young learners and their effect on performance evaluation
Yuko Butler and Wei Zeng, University of Pennsylvania

The CEFR: its virtues and shortcomings when teaching and assessing young learners
Hanan Khalifa, Angela ffrench and Szilvia Papp, University of Cambridge ESOL Examinations

Paper Session 4b
Location: Lecture Room 2
Chaired by: Meg Malone

Multicompetence and assessment: recognizing variability in the L1 benchmark
Amanda Brown

Linguistic characteristics of NS and NNS responses to impromptu essay prompts
Sara Weigle and Eric Friginal, Georgia State University

Validation of a benchmarking study by means of repeated measures classification consistency
Rebecca Present-Thomas and Alistair van Moere, Knowledge Technologies, Pearson

Paper Session 4c
Location: Auditorium
Chaired by: Micheline Chalhoub-Deville

Predicting difficulty in PISA reading items: the process of describing item difficulty
Tom Lumley, Juliette Mendelovits and Dara Searle, Australian Council for Educational Research (ACER)
Irwin Kirsch, Educational Testing Service
Charles Alderson, Lancaster University
John de Jong, Language Testing Services
Dominique Lafontaine, Université de Liège

The dimensionality of test scores on a second language reading comprehension test: implications for accurate estimation of a test taker's level of L2 reading comprehension
Youngsoon So, University of California, Los Angeles

Evidence of lack of processing of multiword lexical items in reading tests
Ron Martinez, University of Nottingham

10.15 – 10.30 BREAK
10.30 – 12.10  Paper Session 5 (parallel)

**Paper Session 5a**  
**Location:** G-R06/07 Seminar Room, Faculty of English  
**Chair:** Yong-Won Lee

- Using the Common European Framework to rate test taker responses  
  Kirsten Ackermann, *Pearson Language Tests*

- Getting the levels right: deriving item writer guidelines from CEFR descriptors  
  John de Jong, *Pearson Language Tests*

- An alternative approach to linking a suite of examinations to the Common European Framework of Reference (CEFR)  
  Gergely A Dávid, *Eötvös University and Euro Examinations, Budapest*

**Paper Session 5b**  
**Location:** Lecture Room 2  
**Chair:** Toshihiko Shiotoku

- Vocabulary use in writing test tasks performance  
  Viphavee Vongpumivitch, *National Tsing Hua University*

- Linking TOEFL scores to the Lexile Measure  
  Cathy Wendler and Fred Cline, *Educational Testing Service*

- Linking TOEFL scores to the Lexile Measure  
  Ellie Sandford, *MetaMetrics, Inc.*

- Lexical resources in legal and general English writing across second language proficiency levels  
  Ivana Vidakovic and Fiona Barker, *University of Cambridge ESOL Examinations*

**Paper Session 5c**  
**Location:** Auditorium  
**Chair:** Annie Brown

- Investigating participants' understanding of proficiency levels in different educational contexts  
  Spiros Papageorgiou, *University of Michigan*

- User perceptions of academic language proficiency and the TOEFL iBT  
  Margaret E Malone and Chengbin Yin, *Center for Applied Linguistics*

- User perceptions of academic language proficiency and the TOEFL iBT  
  Megan Monte, *Georgia State University*

- What can a can-do do? Developing descriptors that define levels  
  Anthony Green, *University of Bedfordshire*

12.10 – 2.25  LUNCH

| 12.25 | ILTA Annual Business Meeting | Auditorium |

2.25 – 3.40  Works-in-Progress Session, Lecture Rooms 1, 4 and 5

**Lecture Room 1**  
**Chair:** Talia Isaacs

1. 'But I had nothing to say' – building a topic validity argument for the IELTS Speaking test  
  Nahal Khabbazbashi, *University of Oxford*

2. Rater feedback on a test of oral proficiency  
  Raili Hildén

3. Automatic fluency assessment using thin slices of spontaneous speech  
  Suma Bhat, Mark Hasegawa-Johnson and Fred Davidson, *University of Illinois at Urbana-Champaign*
4. Investigation of a CAEL writing band score as diagnostic  
Christine Doe, *Queen's University, Ontario*

5. Online training of raters of university-level writing scripts  
Janet von Randow, *University of Auckland*  
Janna Fox, *Carleton University, Ottawa*

6. Linking words used on various levels of proficiency in argumentative writing  
Henna Tossavainen, *University of Jyväskylä*

**Lecture Room 4**
Chaired by: Elvis Wagner

7. An in-process meta-analytic study of language criteria across frameworks – opportunities and challenges  
Eric Surface and Kathryn Nelson, *SWA Consulting*

8. ‘Can Do’ project  
Sarah Brewer and John Slaght, *University of Reading*

9. DIF investigation on gender, academic background, and university type differences in a large-scale high-stakes test  
Xiaomei Song and Liying Cheng, *Queen’s University, Ontario*  
Lianzhen He, *Zhejiang University, PRC*

10. Comparison of students’ performance in dynamic vs static listening comprehension tests among EFL learners  
Sahbi Hidri

11. Investigating the yes/no vocabulary test: input modality, context and response time  
John Read, *University of Auckland*  
Toshihiko Shiotsu, *Kurume University*

12. Establishing standards for EFL tests in China: test construction, evaluation and use  
Jinsong Fan

**Lecture Room 5**
Chaired by: Lorena Llosa

13. Understanding teachers’ classroom assessment practices in relation to students’ linguistic and affective changes: a case study in an EFL Speaking course  
Xiaoying Wang

14. Formative assessment practice in a Bachelor of Education/Teaching English as a Foreign Language Program (BEd/TEFL) in Mainland China: a case study  
Yanming Wu

15. Developing the assessment literacy of IELTS users in higher education  
Kieran O'Loughlin, *University of Melbourne*

16. Investigating the use of English language proficiency tests for decision-making purposes in English-medium universities in Egypt  
Elizabeth Arrigoni, *American University of Cairo*

17. Describing proficiency levels for young students of Italian as L2  
Monica Barni, Sabrina Machetti and Carla Bagna, *Università per Stranieri di Siena*

18. Language aptitude and the study of less-commonly taught languages (LCTLs) by ‘average’ learners in academic settings  
Daniel Reed and Seongmee Ahn, *Michigan State University*
3.40 – 4.10 BREAK (inc. SiLT 15th Anniversary Celebration)

4.10 – 6.10 Symposium Session A (parallel)

**Symposium 1: The industrialization of language testing**

**Location:** Auditorium

**Organiser:** Bernard Spolsky

**Introduction: Industrialisation and centralisation**

Bernard Spolsky, *Bar-Ilan University (emeritus)*

The continuing high cost of industrialized testing

Jared Bernstein, *Pearson Knowledge Technologies*

The marketisation of language testing

Liz Hamp-Lyons, *University of Nottingham*

Retrofit frameworks for the testing industry

Glenn Fulcher, *University of Leicester*

Post-graduate training in language testing

Fred Davidson, *University of Illinois at Urbana-Champaign*

**Symposium 2: Investigating the grammatical domains of meaning conveyance in language assessments and language assessment frameworks**

**Location:** G-R06/07 Seminar Room, Faculty of English

**Organiser:** James E Purpura

How do we define grammatical knowledge in terms of form and meaning dimensions at six different CEFR proficiency levels

James E Purpura, Yoko Ameriks, Jee Wha Dakin and Kirby Grabowski, *Teachers College, Columbia University*

Investigating the dimensionality of grammatical knowledge across two test forms of the University of Michigan Examination for the Certificate of Proficiency in English (ECPE)

Yoko Ameriks, *Teachers College, Columbia University*

Investigating the dimensions of grammatical knowledge in relation to L2 reading and listening abilities within the General English Proficiency Test (GEPT) context

Yen-Fen Liao, *National Taiwan University*

Examining the dimensions of grammatical form, grammatical meaning, and pragmatic meaning at three levels of ability in a role-play speaking test

Kirby Grabowski, *Teachers College, Columbia University*

7.30 – 11.00 Gala Banquet, King’s College
### Friday 16 April 2010

#### NOTE PARALLEL SESSIONS ACROSS MULTIPLE VENUES TODAY

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Title</th>
<th>Location</th>
<th>Chairer</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.30</td>
<td>Announcements</td>
<td>[will be made in all 3 parallel session venues]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.35</td>
<td>Paper Session 6 (parallel)</td>
<td>A multi-dimensional approach to the exploration of rater variability: the impacts of gender and occupational knowledge on the assessment of ESP oral proficiency</td>
<td>G-R06/07 Seminar Room, Faculty of English</td>
<td>Tony Green</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weiwei Liu, <em>Hong Kong Polytechnic University</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>What does it take to make a grade? Teachers assessing grammatical ability in L2 writing with rating scales</td>
<td></td>
<td>Heike Neumann, <em>McGill University</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rating L2 pragmatics: what does holistic scoring really mean?</td>
<td></td>
<td>Yuanshan Chen</td>
</tr>
<tr>
<td>10.15</td>
<td>BREAK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.15</td>
<td>Paper Session 6a</td>
<td>Analysis of factors that influence Korean Middle School students’ English achievement by using a latent growth model</td>
<td>Lecture Room 2</td>
<td>Spiros Papageorgiou</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bokyung Cho, <em>Korea Institute for Curriculum and Evaluation</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Woojin Jung and Kyungyun Kim, <em>Korea University</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>The problem of continuity between primary and secondary school languages programs: what can classroom-based assessment tell us?</td>
<td></td>
<td>Kathryn Hill</td>
</tr>
<tr>
<td></td>
<td></td>
<td>School-based oral language assessment in Hong Kong: students’ attitudes and anxiety</td>
<td></td>
<td>Manman Gao, <em>University of Hong Kong</em></td>
</tr>
<tr>
<td>10.15</td>
<td>Paper Session 6b</td>
<td>Assessing health literacy: a new domain for collaboration between language testers and health professionals?</td>
<td>Auditorium</td>
<td>John de Jong</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Catherine Elder, <em>University of Melbourne</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Using a threshold principle to decide when someone is (or is not) at a proficiency level: examples from Hindi and other less commonly taught languages</td>
<td></td>
<td>Pradyumna Amatya</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fairness and consistency in assigning testing accommodations in large scale testing</td>
<td></td>
<td>Jesse Black-Allen, <em>OISE, University of Toronto</em></td>
</tr>
<tr>
<td>10.15</td>
<td>BREAK</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
10.30 – 12.30 Symposium Session B (parallel)

Symposium 3: Exploring interactional competence in paired and group speaking tests
Location: Auditorium
Organisers: Gillian Wigglesworth and Lyn May

Introduction
Gillian Wigglesworth, University of Melbourne

Interactional competence across proficiency levels: what (more) have we learned from analysing co-constructed paired test interactions?
Evelina Galaczi, University of Cambridge ESOL Examinations

Interactional competence from the rater’s perspective: how separable are the co-constructed performances in a paired speaking test?
Lyn May, Queensland University of Technology

Interactional competence measured in group oral tests: how do test-taker characteristics, task types and group sizes affect co-constructed discourse in groups?
Fumiyo Nakatsuhara, University of Bedfordshire

Group oral tests: what kinds of tasks and functions are optimal for eliciting and measuring functional interactions and interactional competence?
Alistair Van Moere, Pearson Knowledge Technologies

Discussant: Annie Brown, Ministry of Higher Education and Scientific Research, United Arab Emirates

Symposium 4: Testing aviation English
Location: G-R06/07 Seminar Room, Faculty of English
Organiser: J Charles Alderson

Introduction
J Charles Alderson, Lancaster University

Defining communicative effectiveness in aviation
Candace Farris, McGill University

Using pilots to verify the criteria and scoring levels of the International Civil Aviation Organisation (ICAO) rating scale
Ute Knoch, University of Melbourne

Testing English for aviation purposes: the Finnish way.
Ari Huhta, University of Jyväskylä

The politics of aviation English testing
J Charles Alderson, Lancaster University

Discussant 1: John Read, University of Auckland
Discussant 2: Dan Douglas, Iowa State University

12.30 – 2.00 LUNCH

Language Testing Editorial Board Meeting: Lecture Room 1

2.00 – 3.30 Paper Session 7 (plenary)

Chairied by: Dan Douglas

Time trials: the relationship between marker speed and quality
Paul Jaquith, Ministry of Higher Education and Scientific Research, United Arab Emirates
A multi-round validation study of descriptors-based STEP language proficiency scales
Eunice Jang, Jim Cummins, Maryam Wagner and Saskia Stille, University of Toronto

Theoretical basis and experiments of a computer-assisted marking system on short answer questions
Xiangdong Gu, Chongqing University, PRC
Fanna Meng, Henan Normal University, PRC

3.30 – 3.50 BREAK

3.50 – 4.50 Paper Session 8 (plenary)
Chaired by: Ardeshir Geranpayeh

Task difficulty in L2 listening assessment: an investigation of text characteristics and test takers’ perceptions of difficulty
Tineke Brunfaut and Andrea Revesz, Lancaster University

Test taker performance on video vs audio-only L2 listening tests
Elvis Wagner, Temple University

4.50 – 5.35 UCLES/ILTA Lifetime Achievement Award Lecture
Professor Elana Shohamy

5.35 – 5.45 Closing comments
IELTS™

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

The British Council / IELTS Australia joint-funded research program makes a significant contribution to the continuing development of the International English Language Testing System. External studies funded by these two IELTS partners complement internal validation and research studies conducted or commissioned by the third IELTS partner, Cambridge ESOL. The joint-funded studies are an integral part of the ongoing process of test monitoring, validation and evaluation within the overall quality management system for IELTS. They reflect the IELTS partners’ longstanding commitment not only to a process of continuous improvement to the test but also to an ongoing engagement with the wider academic research community for language education and assessment.

The body of published research on the IELTS test continues to grow steadily. The first volume of the IELTS Research Reports appeared in 1998, to be followed over the next 12 years by nine more volumes. Together these publications report a total of 50 empirical studies funded under the IELTS joint-funded research program. Copies can be ordered through the IELTS website: www.ielts.org. In the past four years, several other key IELTS research publications have been produced by Cambridge ESOL and Cambridge University Press in their Studies in Language Testing series. These include academic volumes authored by Roger Hawkey (2006), Tony Green (2007) and Alan Davies (2008), as well as an edited volume by Lynda Taylor and Peter Falvey (2007) of joint-funded research studies focusing on the IELTS Speaking and Writing Modules. More details of all these volumes are available from the Cambridge ESOL website: www.cambridgeesol.org

IELTS Research Reports Volume 10 brings together five studies on the theme of washback and impact – themes which continue to attract attention and provoke debate among the language testing community and beyond. The sustained focus on these issues is reflected in the fact that this is the second collection of IELTS Research Reports to focus on washback and impact as they relate to IELTS, following Volume 8 published in 2008. The contents of Volume 10 deal with contemporary themes of direct relevance not just to the IELTS test producers and other IELTS stakeholders, but to the wider language teaching and testing community.
Call for research proposals

The IELTS partners are making available grant funding to a total value of around £120,000 (AUS$225,000) for IELTS-related research projects to be conducted during 2011.

IELTS Joint-funded research program 2010/2011

Educational institutions and suitably qualified individuals are invited to apply for funding to undertake applied research projects in relation to the International English Language Testing System (IELTS). Financial support for individual projects selected will in principle be limited to a maximum of £15,000/AUS$36,000, though this upper limit may be reviewed in exceptional circumstances.

What areas of interest have been identified?

- Test development and validation issues
- Issues relating to contexts of test use
- Issues of test impact

Consideration will also be given to other issues of current interest in relation to IELTS.

Submission of applications from appropriate institutions must be by 30th June 2010.

For more information: www.ielts.org/researchers/grants_and_awards.aspx

IELTS Masters Award submissions

As part of the tenth anniversary of IELTS in 1999, the IELTS partners established an annual award of £1,000 for the Master’s-level dissertation or thesis in English which makes the most significant contribution to the field of language testing.

Submissions for this prestigious award are now being invited from research academics. Submissions for extracts must be presented by June 2010 with the full copy submission of short listed dissertation/theses by August 2010.

For more information: www.ielts.org/researchers/grants_and_awards.aspx
The Messick Award

Abstract

Validating Score Interpretations and Uses
Dr Michael Kane

The score interpretations entailed by assessment uses can be complicated. In part, these interpretations depend on the kinds of tasks included in the assessment and the domains from which these tasks are sampled, the kinds of responses called for by the test and the rules used to score them, the procedures used to administer the test, and the contexts in which the test is administered. The interpretations are also shaped by the conceptual and social frameworks within which the scores are interpreted and by the theoretical assumptions implicit in these frameworks. And fundamentally, the assessments and the interpretations are developed (or adjusted) to meet the requirements inherent in proposed test-score uses; for example, high-stakes uses generally make stronger claims and require much stronger assumptions than low-stakes uses, and the mapping of test scores to proficiency level descriptors adds a extra layer of interpretation and complexity. As a result, test-score interpretations can be very complicated. To make things worse, many of the assumptions built into score interpretations and uses may be implicit.

Validity is currently defined in term of the degree to which a proposed interpretation is justified by evidence and involves at least two stages. First, the proposed interpretations and uses have to be specified in some detail (e.g., in the form of an explicit interpretive argument, laying out the inferences and assumptions leading from the scores to the claims and decisions based on the scores). As noted above, the interpretations can be complex, and not completely spelled out, and therefore, specifying the proposed interpretation and use is not trivial. Second, the proposed interpretation/use are subjected to searching criticism, and where possible, empirical evaluation. The validity argument provides a systematic summary of this evidence.

The process is akin to the process of theory development and theory testing in applied contexts; the question is not whether the theory (interpretation/use) is true in general, but whether it works for this purpose in this context. Like scientific theory testing, validation is simple conceptually (say what you mean and justify your claims), but gets complicated as the interpretations and uses get more and more ambitious. Language assessments pose unique challenges, because they necessarily involve complex tasks embedded in social contexts and can serve ambitious goals, but they have the advantage of being associated with fairly clear performance domains. We have to pay attention to the challenges, and we should take advantage of the opportunities.

Biography

Michael T. Kane, Ph.D., has held the Samuel J. Messick Chair in Validity at the Educational Testing Service in Princeton, New Jersey since September of 2009. He was Director of Research for the National Conference of Bar Examiners from September 2001 to August 2009. From 1991 to 2001, he was a professor of kinesiology in the School of Education at the University of Wisconsin–Madison, where he taught measurement theory and practice. Before his appointment at Wisconsin, Kane was a senior research scientist at ACT, where he supervised large-scale validity studies of licensure examinations. Kane holds a B.A. in physics from Manhattan College, and an M.S. in statistics and a Ph.D. in education from Stanford University. His main research interests are in validity theory and practice, generalizability theory, and standard setting.
Only the TOEIC® test proves your employees have the English skills demanded by today’s global marketplace

Give your organization the competitive advantage it needs to succeed

English is the language of business. To achieve success worldwide, it’s critical that your company’s workforce have the English skills demanded by businesses worldwide. But how can you be sure your employees have the skills to move your business forward?

For recruitment, placement and promotion, no other English-language test is administered more often, in more places, to more employees than the TOEIC® test. That’s why more than 9,000 organizations around the world trust the TOEIC test as the one assessment they need.

Your employees are your company’s future. Make sure they have the English skills they need to help take your company to the next level with the TOEIC test.

To learn more about how the TOEIC test can help your company succeed, contact your local ETS Preferred Network office, visit www.ets.org/toeic or e-mail globalforum@ets.org.

Know English. Know Success.
Elana Shohamy—Winner of the UCLES/ILTA Lifetime Achievement Award

20% discount

Linguistic Landscape
Expanding the Scenery
Edited by Elana Shohamy, Tel Aviv University, Israel and Durk Gorter, University of the Basque

In this comprehensive and pioneering volume, language scholars from around the world examine the "linguistic landscape" from multiple perspectives – theoretical, methodological, and critical. Written by widely recognized experts, the articles in Linguistic Landscape: Expanding the Scenery analyze linguistic landscapes in a range of international contexts. Dozens of photographs illustrate the use of language in the environment – the words and images displayed and exposed in public spaces. Suitable for graduate or advanced undergraduate students in applied linguistics, sociolinguistics, and language policy studies, Linguistic Landscape is a vital contribution to a burgeoning field.

2008: 234x156: 392pp
Hb: 978-0-415-98872-8: £85.00
Pb: 978-0-415-98873-5: £30.99
Discounted price: £68.00
Discounted price: £25.00

Language Policy
Hidden Agendas and New Approaches
Elana Shohamy, Tel Aviv University, Israel

'...the book is a must in the reading list of all those who believe that we academics ought to pursue our vocation in the interests of the community at large that, after all, helps sustain the institutions that host us. It is a wake-up call to critical thinking and activism.' - Kanavillil Rajagopalan, State University at Campinas, Brazil

A critical look at language policies, how they are implemented and the hidden agendas which often lie behind them, drawing on examples from the US and UK and showing what the consequences are for the people involved.

2005: 234x156: 185pp
Hb: 978-0-415-32864-7: £70.00
Discounted price: £56.00
Discounted price: £18.00

To claim your discounted copy email laura.maisey@tandf.co.uk. Discount valid until 31st May 2010.

Visit our website for more information and online ordering:
www.routledge.com
UCLES/ILTA Lifetime Achievement Award

Elana Shohamy

On behalf of the review committee for the UCLES/ILTA Lifetime Achievement Award (consisting of Sarah Briggs, Nick Saville, Carolyn Turner and myself) it is a great pleasure to announce that after carefully considering the several highly meritorious nominations received from the field, Professor Elana Shohamy has been selected to receive the Lifetime Achievement Award to be presented at the 32nd ILTA/LTRC meeting which will be held in Cambridge, UK in April 2010.

Over a long career, Elana has maintained her programme of empirical research, her insider critical critique of the basis of language assessment and her concern for understanding the interaction between language assessment and language policy. After completing her PhD at the University of Minnesota in 1978, she soon started to raise questions about the nature and use of tests. It is particularly in this area that her influence has been most felt: her recognition of the power of tests, whether in education or in immigration, built the critical link between testing and language policy, showing language tests to be one of the most common instruments for enforcing language policy; her influence was the core of the growing emphasis on ethics.

In her career, she has built a number of significant tests and has made major contributions to the understanding of washback and other key research areas. The quality of her research and scholarship has built her a major place as a leader in the field, combining academic and professional activities. Furthermore, her example and teaching have built a generation of language testers who look to her as a leader. She is Professor of Language Education at Tel Aviv University but she has taught at a large number of universities in various parts of the world, was the research director at the National Foreign Language Center in Washington, DC, and is affiliated with CALPER (Center for Advanced Language Proficiency Education at Research) at the Pennsylvania State University.

She was one of the founders of an early national language testing group (ACROLT), chaired the AILA Commission on Language Testing and Evaluation, and served as President of ILTA in 1999. Apart from her own extensive list of publications, she has played a major role in editing, serving on the editorial boards of Language Testing, Language Assessment Quarterly and the Modern Language Journal and as founding co-editor of Language Policy.

Because of her many significant contributions over the years, the committee is pleased to announce that Professor Elana Shohamy has been selected to receive the 2010 UCLES/ILTA Lifetime Achievement Award.

J Charles Alderson

Chair, 2010 UCLES/ILTA Lifetime Achievement Award Committee
Invited Lecture Abstract

From critique to change: Building on the power of tests

Professor Elana Shohamy

The paper begins with reflections on biographical and personal episodes of encountering informal language tests as prime definers of personal and national categories, a phenomenon which is common in linguistically diverse societies where ‘one language’ serves as the marker of fixed identities and collective ideologies. In many multilingual contexts life is a series of language tests with meaningful consequences. Research in language testing in the past few years focused on various dimensions of test consequences and impacts, demonstrating the enormous power of institutionalized formal tests as devices that establish fixed categories. Indeed this work led to increased awareness of test consequences, such as washback, as well as activism in the form of codes of ethics and practice. Yet, institutionalized language testing is being expanded at an accelerated rate as tests, often with high stakes, are imposed on those who traditionally have not been included in the testing endeavors, and are backed by powerful institutions (governments, testing corporations and academic institutions). What are the implications of such policies for the language testing profession? What should the responses of language testers be when faced with these realities?

The main argument of this paper is that the very ‘power of tests’ puts an enormous amount of responsibility in the hands of those who wield those tests. A ‘responsible language testing’ approach therefore emphasizes the need to create tests that truly reflect language acquisition and use in the multiple and diverse contexts where they are used. This approach not only attempts to minimize the negative consequences of tests as tools of punishment or categorization but also constructs tests as devices for the improved functioning of society. Responsible language testing thus goes beyond research into consequences and impact — although not excluding it — expanding the language construct so as to match tests with learning processes and functionality. Such an expansion embraces multiple contextual variables, research findings on processes and products of language acquisition and uses of languages in diverse situations in today’s multilingual world. In these situations languages are far from being homogenous as they consist of bi-, multi-, hybrid and digital varieties, beyond simplistic criteria of one size fits all. Such an approach rejects the ideological role of tests as definers of what language knowledge should be but rather follows actual practices of what language proficiencies are so as to minimize the contradictions between ideology and practice. This implies negotiating with policy makers regarding unrealistic language goals and the exploitation of test makers to comply with these intentions. It calls for growing awareness and knowledge of intentions behind test driven policies as well as of research in applied linguistics regarding language learning and use in this day and age. It means taking a mediating role between governments and centralized policy makers as they interact with the schools and communities who are faced with the need to respond to these policies. These approaches will be discussed as a constructive response building on the power of tests and recommending that it becomes our future research and policy agenda.

Biodata

Dr Elana Shohamy is a professor of language education at the School of Education, Tel Aviv University. Her research and writings focus on a variety of topics related to language testing and language policy in the context of conflicts, co-existence and rights in multilingual societies. Her research and publications in language testing address the political, social, educational and ethical dimensions of tests within a critical framework. These studies point to the power and misuses of tests in education and society, and the consequences of tests for immigrants, minority and majority language learners and adults in the context of migration and citizenship. Her work in language policy introduced an expanded policy framework which showed how various mechanisms (e.g., tests, regulations, linguistic landscapes), create de facto language policies. These implicit policies lead to inequalities, lack of rights, suppression of diversity and marginalisation.

Her current work in language testing builds on an understanding of the power of tests and its consequences to create tests which are not driven by ideology with questionable goals but based instead on a valid construct that reflects the actual processes, uses and functioning of languages by diverse populations; such tests build on research into language learning processes, bi/multiple and hybrid varieties of L-1s and L-2s, variety of familiar contents, multi-modalities, the diverse literacies
of the digital world and multiple types of assessment. Her current work in language policy is situated within the framework of ‘language policy from below’, incorporating personal experiences, biographies and narratives of language users about the processes, costs and benefits of acquiring and functioning in new languages.

Elana’s publications in the past decade reflect the above interests: The languages of Israel: Policy, ideology and practice (with Bernard Spolsky, Multilingual Matters, 1999); The power of tests, (Longman, 2001); Language policy: Hidden agendas and new approaches (Routledge, 2006); Encyclopedia of Language and Education: Language Testing and Assessment, Volume 7 (ed. w/ Nancy Hornberger, Springer, 2008); Linguistic landscape: expanding the scenery (ed. with Durk Gorter, Routledge, 2009); and Linguistic landscape in the city (ed. With E. Ben Rafael & M. Barni, Multilingual Matters, in press); she is currently working on a book entitled: The cost of reviving a language. Elana is also the current editor of the journal Language Policy. She has been an active member of the Language Testing Research Colloquium since its inception in 1979.
Studies in Language Testing 15th
Anniversary Edition

33 volumes in 1 anniversary edition

Each volume focuses on important developments in language testing and assessment making them essential reading for a wide range of education and testing professionals.

The Studies in Language Testing series is a cutting-edge collaboration between Cambridge ESOL and Cambridge University Press.
Abstracts

Pre-Conference Workshops

As a major focus of language testing attention in Europe and beyond the Common European Framework of Reference for Languages (CEFR) is of interest to a large number of conference attendees, and it is well-related to this year’s LTRC theme of investigating levels, domains and frameworks in language assessment. For this reason, two one-day pre-conference workshops are offered this year focusing on the CEFR and led by colleagues who have been actively involved in the CEFR since its earliest days.

Workshop 1: The CEFR: its purpose, origins, development and current status

Neil Jones, Nick Saville and Szilvia Papp, University of Cambridge ESOL Examinations
John Trim, former Director of the Council of Europe’s Modern Languages Projects, co-author of CEFR
Brian North, Eurocentres, co-author of CEFR
Norman Verhelst, CITO, co-author of the Manual for relating examinations to the CEFR

The aim of this one-day workshop is to introduce the Common European Framework of Reference for Languages (CEFR), to familiarise people with its origins and history, with its evolution, with its aims and objectives and its work-in-progress nature. Attention will also be given to the descriptor scales, as the most visible aspect of the CEFR and the one which colours people’s interpretation of it. Workshop activities will include plenary presentations, group discussions and exercises, and the day will end with a round table question and answer session with some of the original authors of the CEFR.

Workshop 2: The CEFR: alignment and standard-setting issues

Neil Jones and Angeliki Salamoura, University of Cambridge ESOL Examinations
Brian North, Eurocentres, co-author of CEFR and of the Manual for relating exams to the CEFR
Norman Verhelst, CITO, co-author of the Manual for relating exams to the CEFR
Neus Figueras, Generalitat de Catalunya, co-author of the Manual for relating exams to the CEFR
Sylvie Lepage, Centre International d’Études Pédagogiques (CIEP)

The aim of this one-day workshop is to explore CEFR alignment issues, including the rationale for linking diverse contexts via a common point of reference, methodologies, outcomes, limitations, etc. The morning will focus on performance skills, the afternoon on objective tests. Cross-linguistic alignment issues will be of particular interest, along with standard setting and alignment as focused on the productive skills of speaking and writing. There will be discussion of the Council of Europe’s Manual for relating language exams to the CEFR, and the relative merits of different alignment and standard setting approaches. The format will involve presentation followed by workshop activities.

Workshop 3: Assessing levels of academic language proficiency: conceptualization, operationalization, alignment, and validation

Margo Gottlieb, Tim Boals and Carsten Wilmes, World-Class Instructional Design and Assessment (WIDA) Consortium
Dorry Kenyon, Center for Applied Linguistics, Washington, DC
Fred Davidson, Youngshin Chi and Chih-Kai (Cary) Lin, University of Illinois at Urbana-Champaign

In Workshop 3 (2 days) participants will become familiar with, explore issues related to, and apply within their own contexts approaches to the implementation of a comprehensive system for language assessment based on proficiency levels interpretable to a wide variety of test stakeholders. Wide-ranging experience from the World-Class Instructional Design and Assessment (WIDA) Consortium
Abstracts – Pre-Conference Workshops

(www.wida.us), a consortium of 20 states that assesses over 725,000 students in grades Kindergarten to 12 annually, will provide models and ways of approaching crucial issues in assessing language proficiency and interpreting results using proficiency levels.

The workshop will be divided into four half-day sessions: i) conceptualisation of academic language proficiency; ii) operationalisation of the construct; iii) alignment of language and content within a standards-referenced system; and iv) validation. While the specific format will vary depending on the topic, the workshop combines some lecturing with the presentation of examples and discussion (in small groups and as a whole) of ways in which the examples from the WIDA Consortium can be applied to the participants’ local contexts.

Workshop 4: Sharing your story: developing your academic conference skills

Lynda Taylor, University of Cambridge ESOL Examinations
Liz Hamp-Lyons, University of Nottingham

Sponsored by ILTA, Workshop 4 is offered as a professional development opportunity for graduate students and early career researchers. This workshop is designed to teach or improve some of the practical skills that are needed when submitting to, presenting at or simply attending academic conferences in our field. We shall look at the knowledge and skills involved in: the writing of a successful abstract; the preparation of an eye-catching conference poster; the planning and management of a productive work-in-progress session; and the delivery of a high-quality research paper or a symposium. We shall also reflect upon the experience of attending an academic conference and how to get the best out of it, both professionally and personally. This workshop is offered at a subsidised rate.

Workshop 5: Cambridge Assessment Tour

Delegates are invited to participate in a tour of Cambridge Assessment premises and to learn more about Europe’s largest assessment agency. Cambridge Assessment was established in 1858 as the University of Cambridge Local Examinations Syndicate (UCLES), and celebrated its 150th anniversary in 2008. It provides exams and tests covering a huge range of subjects and levels. More than eight million are taken every year, in over 150 countries. Cambridge ESOL is part of the Cambridge Assessment Group. The tour is offered free of charge.

'Meet the Journal Editors’ Session

This session aims to offer practical guidance on getting published in our field. A panel of journal editors in the field of language testing/assessment and related disciplines will explain the editorial policy of their respective journals and answer questions on what is involved in getting yourself published. The session is free of charge and open to all conference delegates, but may be especially helpful for early career researchers and graduate students. Editors of the following journals will be present:

Language Testing – Catherine Elder and Glenn Fulcher (Editors)
Language Assessment Quarterly – Antony Kunnan (Editor), Constant Leung, James Purpura, Nick Saville, Carolyn Turner (Associate Editors)
Assessing Writing – Liz Hamp-Lyons (Editor)
Assessment in Education – Pauline Rea-Dickins (Executive Editor)
Language Policy – Elana Shohamy (Editor)

The LTRC 2010 Local Organising Committee would like to thank the journal editors listed above who freely gave their time and energy to make this session possible.

LTRC Newcomers’ Session

This session aims to provide a welcoming introduction to anyone attending LTRC for the first time. The panel will include the LTRC 2010 organisers, LTRC ‘veterans’ and some who have joined our
LTRC community in recent years. We will briefly explain how LTRC has evolved since 1979, how it is organised today and answer any questions you may have. Though aimed at newcomers, this orientation session is open to all conference delegates and is free of charge.

The LTRC 2010 Local Organising Committee is most grateful to all those who agreed to take part in this welcoming session, as well as to the University of Michigan for its generous sponsorship support.
Abstracts – Symposia

Symposium Session A (parallel)

Time: Thursday, 15th April, 4.10–6.10

Symposium 1

The industrialization of language testing

Organiser:
Bernard Spolsky, Bar-Ilan University (emeritus), Ramat-Gan, Israel

Contributors:
Jared Bernstein, Pearson Knowledge Technologies, Palo Alto, USA
Fred Davidson, University of Illinois at Urbana-Champaign, USA
Liz Hamp-Lyons, University of Nottingham, UK/University of Hong Kong, Hong Kong
Glenn Fulcher, University of Leicester, UK

The decision of Pearson (NY Times Sept 8 2009) to enter the “commercial, competitive market” of English as a second language testing already generating “several hundred million dollars in annual revenue” confirms that this is no longer a “cottage industry” but a “consolidated enterprise” (Clark & Davidson 1993). Although the US educational testing business began in the 1920s, the industrialization of English language testing is better dated from the 1960s, when ETS acquired TOEFL. Fulcher (2009) sees this as part of growing collectivism and centrism in testing. Taylor (2009) suggests that the low level of assessment literacy among politicians, educational administrators, and the public has enabled the profitability and power of the industry to grow unhindered. The symposium will look at aspects of the phenomenon and consider what might be done about it.

Bernard Spolsky: Introduction: Industrialization and centralization

The growth of TOEFL from a language-tester directed project to a business-manager controlled profit center was traced in Spolsky (1995), which also recorded the burgeoning industrialization of UCLES and its change from a small academic test producer into a successful global competitor. The development continues apace, recently spurred by the entry into language testing of enterprises like Pearson and the German-owned Macmillan Group. English language testing is now a highly profitable commodity: more than half of The Washington Post’s revenue last quarter came from its earlier purchase of the Stanley Kaplan test preparation company. There are new government language testing agencies and businesses in Europe and Asia, and the popularity of the Common European Framework is supporting widespread bureaucratization of language testing on a single model. In this symposium, we will track some aspects of this development and consider its causes and effects, its advantages and disadvantages, and how language testers may usefully respond.

Jared Bernstein: The continuing high cost of industrialized testing

The paper will review the reasons why industrial tests have been established, widely adopted, and often applied beyond their intended domains. As with manufactured goods, industrialization of language testing has generally improved quality standards and extended function. However, unlike other industrial goods, tests may need authority, security, and/or norm reference, which can keep the price of mass-delivered standard tests high. Examples will be presented that suggest that prices also remain high because of an “open loop” among producers, adopters and payers, wherein those who choose (or authorize) tests are not burdened with the fees. Re-conception of test security and tools for mass customization of tests offer two ways to bring prices down while maintaining the advantages of industrial testing.

Liz Hamp-Lyons: The marketization of language testing

Considerable research and validation goes into the most reputable language tests. But the great majority of language testing occurs outside the vision of the agencies producing and marketing these high quality tests. Also, the tests themselves are compromised in a number of ways, and sometimes risk their own reputations due to ‘mission creep,’ allowing their use as “solutions” to wildly-
varying “problems” in contexts for which they were neither designed nor validated. Language tests are becoming, or have become, the answer to everyone’s problem, and this attitude has led to the spawning of more and worse “language tests” within key industries. For the assessment of language, Ford was wrong.

Glenn Fulcher: Retrofit frameworks for the testing industry
It should not be surprising that money is made out of any large-scale human activity. In an early volume on testing, Latham (1877: 6) says, ‘people are hardly aware of how thoroughly the educational world is governed by the ordinary economical rules.’ Language tests, like all educational assessments, are expensive to design, develop, and deliver, if it is to be done well. Testing organizations wish to recoup this investment and generate a profit. One of the easiest ways to do this is to maximize testing volume for every test created, which often requires the extension of test purpose to ever newer uses. For each new use it is incumbent upon the testing agency to provide validation evidence for the usefulness of the test for its new use. A retrofit framework provides an agenda for this validation work, and a structure for the evaluation of fair and ethical use of the test re-use.

Fred Davidson: Post-graduate training in language testing
Drawing on published sources in higher education (e.g. publications by the Carnegie Project on the Education Doctorate and Nerad and Heggelund (2008)) and on personal experience in management of a graduate student research team, the paper will argue that client-centered test development should be part of any such curriculum. Failure to do so ignores the realpolitik of the world language testing community.

Discussion: As each participant will have 15 minutes, there will be ample time for discussion from the floor. Those who would like to reserve a five-minute slot are invited to contact Bernard Spolsky bspolsky@gmail.com.

Symposium 2

Investigating the grammatical domains of meaning conveyance in language assessments and language assessment frameworks

Organiser:
James Purpura, Teachers College, Columbia University

Contributors:
Yoko Ameriks, Teachers College, Columbia University
Kirby Grabowski, Teachers College, Columbia University
Jee Wha Dakin, Teachers College, Columbia University
Yen-Fen Liao, National Taiwan University

Overview of the symposium
Despite movements to downplay the role of grammar in second and foreign language classrooms, language teachers have always known that grammar provides a fundamental resource for communication. Teachers also know that communicative effectiveness depends, to a large extent, on a learner’s grammatical knowledge and how this knowledge is used to understand or produce a range of literal and pragmatic meanings in socially and culturally-rich language use contexts (Purpura, 2004). As a result, grammar and communicative grammar instruction have never ceased to be the mainstay of most second and foreign language classrooms. The importance of grammar in language education has been further highlighted by the extensive body of research in SLA (e.g., Doughty & Williams, 1998; Ellis, 2008; Pienemann, 1998; VanPatten, 2004; VanPatten et. al., 2004), providing evidence that language learning is enhanced when grammar instruction emphasizes the development of explicit and implicit form-meaning connections. These connections begin with initial links between phonological, lexical and morphosyntactic forms and their meanings to the use of forms to convey a range of meanings in situated communication. While empirical research in SLA on the development of grammatical knowledge of L2 learners has evolved considerably since the 1990s, research in language assessment on the measurement of grammatical knowledge has lagged behind. Reasons for this lack of research might arguably be attributed to concerns that explicit grammar assessments might lead: (1) to practices in which grammar and grammatical accuracy are privileged as the predominant criterion for communicative effectiveness; (2) to practices where grammar assessment is based solely on a syntactocentric view of grammar rooted in structural linguistics.
and discrete-point measurement of the 1960s rather than on current views of grammar conceived in terms of linguistic pragmatics and functional grammar, measured through performance; or finally (3) to opinions that learners who have achieved “functional” levels of communicative effectiveness should no longer be assessed in grammatical knowledge. While these arguments might have some traction, it is evident that grammatical knowledge remains one of the fundamental resources for being able to express a wide range of meanings in communication, and merits, therefore, further empirical inquiry. In fact, compelling questions still remain concerning: (1) the dimensions of grammatical knowledge represented in assessment frameworks at different proficiency levels; (2) the measurement of grammatical knowledge in the domain of form-meaning construction; (3) the validity of using dichotomous scoring methods with items designed to measure partial knowledge; (4) the role of grammatical knowledge, defined in terms of form-meaning connections, in situated comprehension and production; and (5) the potential that grammar assessment has to provide teachers and learners with concrete information for closing learning gaps, thereby furthering SLA. In an effort to reopen a discussion of grammar in the context of language assessment, the purpose of the current symposium is to bring together empirical studies that have: (1) investigated the measurement of grammatical knowledge in a number of assessment contexts with reference to differing frameworks; and (2) empirically explored how the development of form-meaning connections may contribute to language use. The first talk explores how different levels of grammatical knowledge (i.e., knowledge of form and meaning) in the Common European Framework of Reference (CEFR) can be theoretically conceptualized and empirically derived through content analysis and Rasch measurement. The second paper investigates the comparability of two forms of the grammar section of the University of Michigan Examination for the Certificate of Proficiency in English (ECPE). After comparing the domains of grammatical knowledge across both test forms, the dimensionality of grammar sections of each test was examined using multi-group structural equation modeling. The third paper investigates the extent to which grammatical knowledge measured in General English Proficiency Test (GEPT), a high-stakes test of communicative language ability in Taiwan, serves as a predictor of students’ L2 reading and listening ability. The last study examines the extent to which examinees’ performance on a role-play speaking test at three different levels of ability was influenced by the students’ ability to use grammatical forms to express literal, intended and a range of pragmatic meanings (i.e., sociolinguistic, sociocultural, and psychological) in socially and culturally-rich situations as opposed to selected test method facets. The symposium will conclude with a critique of these papers by a discussant.

Paper 1: How do we define grammatical knowledge in terms of form and meaning dimensions at six different CEFR Proficiency Levels? – James E Purpura, Yoko Ameriks, Jee Wha Dakin and Kirby Grabowski, Teachers College, Columbia University

Over the years, the Common European Framework of Reference (CEFR) (Council of Europe, 2001) has been increasingly used for the development of language curricula, textbooks, and assessments. This framework claims to provide a comprehensive description of “what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively” (COE, 2001, p.1). The CEFR also claims to provide language proficiency descriptors that can be used to portray language growth on a vertical scale across different levels of achievement (COE, 2001, p. 17-18); however, with respect to the grammatical resources underlying language ability, the CEFR is notoriously vague. In fact, a close examination of the CEFR scales shows that the CEFR does not actually specify what knowledge components learners need to have acquired at different proficiency levels. Rather, it describes what learners might be expected to do with the language at these levels. Also, while the CEFR attempts to account in some way for the grammatical resources involved in language use, it makes no systematic attempt to define and scale the ability to convey meaning at each of the six CEFR levels. The goal of the current paper was to examine the nature of grammatical knowledge, defined in terms or grammatical form and meaning (Jaszczolt, 2002; Larsen-Freeman, 1991, 2002; Leech, 1983; Rea-Dickins, 1991; Purpura, 2004, 2007; VanPatten et al., 2004), at six CEFR levels of proficiency, so that scaled descriptors for each domain could be generated in the context of the Oxford Online Placement Test (OOPT). In the study, 1,000 grammatical forms and 715 meaning items were created and coded in accordance with Purpura’s (2004) model of grammatical knowledge, and then administered to approximately 900 learners spanning six CEFR levels. A vertical scale was constructed via a counterbalanced single-group design, where each CEFR group was given items appropriate to its level along with items immediately below and above the level. The construction of vertically-scaled descriptors for grammatical form and meaning drew on the following analyses. First, student
responses were analyzed by means of Winsteps (Linacre, 2009) and a vertical scale of item difficulties was created. The item CEFR levels, produced by Winsteps (Linacre, 2009), were then compared with the levels hypothesized by the item coders in a comprehensive content analysis. The results from both analyses served as a basis for depicting grammatical form and meaning domains expected of learners at each of the six CEFR proficiency levels. The paper will describe the process for constructing these scales.

Paper 2: Investigating the dimensionality of grammatical knowledge across two test forms of the University of Michigan Examination for the Certificate of Proficiency in English (ECPE) – Yoko Ameriks, Teachers College, Columbia University

When multiple test forms are used, test developers commonly use test equating procedures to confirm that the scores across forms are comparable. Since the scores from these tests are used to make high-stakes decisions about test takers, examining the comparability of the scores and that of their underlying dimensions are critical evidence in the validation argument. The present study examined the comparable dimensionality of the grammar, cloze, and vocabulary section of two different forms of the Examination for the Certificate of Proficiency in English (ECPE), developed by the University of Michigan. The ECPE was administered at over 125 test centers in 20 countries over a span of two years. Between the two test administrations, participants included a total of 66,135 EFL test-takers. The comparability of the two grammar tests was examined by means of multi-group structural equation modeling. The results suggested that the items in the grammar, vocabulary and cloze section measured grammatical knowledge, specifically grammatical form and meaning. Furthermore, the two forms of the same test had virtually identical underlying constructs.

Paper 3: Investigating the dimensions of grammatical knowledge in relation to L2 reading and listening abilities within the General English Proficiency Test (GEPT) context – Yen-Fen Liao, National Taiwan University

The current study investigated the predictive validity of the dimensions of grammatical knowledge in relation to L2 reading and listening abilities in the context of the GEPT, a high-stakes, criterion-referenced test commonly used in Taiwan. One form of the Intermediate GEPT reading and listening sections was administered to 609 Taiwanese university freshmen. The grammar items were then coded for knowledge of grammatical form and meaning, while the reading and listening items were coded for comprehension of literal and pragmatic meanings. The data were then analyzed using structural equation modeling and discriminant analysis in order to investigate the contribution of grammatical knowledge on reading and listening abilities. The results showed that knowledge of grammatical form and meaning seemed to be weaker predictors of listening than of reading ability. Finally, grammatical meaning, rather than form, had a strong effect on reading ability, and an even stronger effect on listening ability.

Paper 4: Examining the dimensions of grammatical form, grammatical meaning, and pragmatic meaning at three levels of ability in a role-play speaking test – Kirby Grabowski, Teachers College, Columbia University

The purpose of the study was to investigate the extent to which test-takers’ performance on a role-play speaking test at three different ability levels (i.e., intermediate, advanced, and expert) was affected by the dimensions of language ability operationalized in this test as opposed to method factors (i.e., raters, tasks, and their interactions). In this study, 102 participants were administered a role-play speaking test and scored in terms of five dimensions of language ability: grammatical accuracy, grammatical meaningfulness, and sociolinguistic, sociocultural, and psychological appropriateness. The data were analyzed using many-facet Rasch measurement and multivariate generalizability theory. The results suggested that all three dimensions of pragmatic knowledge should be explicitly tested at all levels of communicative language ability, and at the expert level, in particular. The results also showed that knowledge of grammatical form and meaning should be assessed separately at all levels of proficiency.
Symposium Session B (parallel)

Time: Friday, 16 April, 10.30–12.30

Symposium 3

Exploring interactional competence in paired and group speaking tests

Organisers:
Gillian Wigglesworth, University of Melbourne
Lyn May, Queensland University of Technology

Contributors:
Evalina Galaczi, Cambridge ESOL
Fumiyo Nakatsuhara, University of Bedfordshire
Alistair vanMoere, Pearson Knowledge Technologies

Discussant:
Annie Brown, Higher Colleges of Technology

Paired and group speaking tests are now commonly used in both low and high stakes assessment contexts in order to elicit a wider range of interactional skills. While these test formats provide opportunities for candidates to co-construct discourse and demonstrate conversational management skills, the importance of defining the construct of interactional competence that underpins these tests cannot be underestimated. Conversation Analysis (CA) has proved an invaluable tool for exploring the complexity of interactional patterns that are co-constructed by candidates in paired and group test interactions. Other research tools, such as Stimulated Verbal Recall (SVR), have been used in rater studies of paired speaking tests to analyse the features of a co-constructed performance that are salient to raters. Issues of concern have included the impact of variables such as proficiency levels, gender, personality, and task types on the co-constructed interaction. The papers in this symposium explore the construct of interactional competence in paired and group speaking tests from different, and yet complementary, perspectives.

The first two papers focus on the construct of interactional competence in the context of paired tests at different proficiency levels, and explore it through the lens of both the elicited discourse and the raters’ perceptions of the performances. The third and fourth papers shift the focus to the interaction in group oral tests and explore the impact of test-taker characteristics, task types and task implementation conditions on the interaction and scores as well as continue to discuss the issue of scoring co-constructed performance. Each paper concludes by summarising both its contribution to our understanding of the construct of interactional competence, and the ways in which various variables may impact upon this construct.

The 120 minute symposium is designed to allow ample time for both panel and audience discussion. The timing is: Introduction: 10 minutes; Papers 1–4: 20 minutes each; Discussant: 10 minutes; Discussion: 20 minutes.

Individual abstracts

1. Interactional Competence across proficiency levels: What (more) have we learned from analysing co-constructed paired test interactions? Our understanding of interaction in paired speaking tests has developed greatly in the last decade, as a growing body of literature has focused – both empirically and theoretically – on this alternative speaking test format. Little empirical work, however, has been carried out on interactional competence within the context of varying proficiency levels. The current study aims to fill that niche and offer an investigation of the talk which is co-constructed by learners at CEFR levels B1 to C2 who are engaged in a general English speaking test. The participants in the study (3 pairs at each proficiency level) were selected from video-recorded benchmarked speaking performances used in rater training. Interactional competence was analysed in terms of the degree of mutuality and equality in topic development. The analysis drew on quantitative methodologies (e.g., number of turns and words), as well as qualitative ones (e.g., micro-analytic features of interactional competence, such as turn-taking, topic initiation and expansion, listener support, and cohesion between turns). Features of the performances salient at different proficiency levels will be overviewed in the presentation and an argument will be made for lower- and higher-level interactional competence. The findings will be related to the definition and operationalisation of...
interactional competence, as well as to the development of more sophisticated assessment criteria and rating scales.

2. Interactional competence from the rater’s perspective: How separable are the co-constructed performances in a paired speaking test? Recent discourse studies on paired speaking tests have provided valuable insight into the co-construction of the discussion, and in particular the impact of variables, including proficiency levels, on the interactional patterns that candidates orient toward. However, very little is known about the features of the jointly negotiated performance that are salient to raters when they award scores for interactional competence. This presentation, therefore, examines features of interactional competence that were salient to four trained raters of twelve paired EAP speaking tests through the analysis of rater notes, stimulated verbal recalls, rater discussions and excerpts of candidate discourse. Features of the performance that raters noted when awarding higher and lower scores for interactional competence will be identified. Raters clearly perceived certain features to be mutual achievements, and the implications of this – including the extent to which it is possible to assess individual contributions to the co-constructed performance – will be discussed. The findings will add to our understanding of the complexity of the construct of interactional competence in tests requiring candidate-to-candidate interaction, and be linked to the the development of rating scales.

3. Interactional competence measured in group oral tests: How do test-taker characteristics, task types and group sizes affect co-constructed discourse in groups? This study explored the nature of co-constructed interaction in group oral tests by examining the impact of a test-taker’s own and his/her group members’ extraversion- and oral proficiency-levels on interactions in groups. The impact was examined across three task types and two group sizes, to obtain a comprehensive picture of joint performance in groups. Data were collected from 269 test-takers, who performed information-gap, ranking and free-discussion tasks either in groups of three or four. The candidates’ extraversion-levels were measured by Eysenck Personality Questionnaire, and their oral proficiency-levels were measured by classroom teachers’ judgements. All group oral sessions were video-taped and analysed using CA methodology to discover 1) how test-takers with different extraversion- and oral proficiency-levels interact in group oral tests in general (regardless of task types and group sizes), 2) how they interact in different task types, and 3) how they interact in different group sizes. The findings provide insights to aid in understanding interactional competence measured in group oral tests, and to help test-designers select task types and group sizes according to the aspects of interactional competence that they would like to assess. Based on several CA examples, different suggestions from Paper 2 will be made in regard to how to rate co-constructed discourse in groups.

4. Group oral tests: What kinds of tasks and functions are optimal for eliciting and measuring interactional competence? Although it is logical to assume that one of the best ways to evaluate candidates in ‘natural’ interaction is to observe them in peer discussion, not all group speaking tests have been arranged so as to allow such interactions to take place (Hilsden, 1991; He and Dai, 2006). This paper reports a study on group speaking tests in which task characteristics were manipulated to optimally elicit a wide a range of functional interactions on which to judge interactional competence. 60 candidates each took three tasks – discussion, picture-difference, and consensus-reaching – in a counter-balanced design, and their performances were recorded and transcribed. Conversation Analysis and Interactive Function Analysis revealed that the discussion task did not consistently elicit authentic and co-constructed conversations, but that the desired interactions were more likely to be elicited when the tasks contained an information gap or specific goal. On the other hand, analysis of ratings data revealed that not all interactions are suitable for the assessment of interactional competence: clarifying, confirming and communication repair were not useful predictors of proficiency as compared to more cognitive functions such as persuading, suggesting or arguing. The findings complement the previous three papers by focusing on elicited performance to help define the construct and usefulness of interactional competence.
Symposium 4

Testing aviation English

Organiser:
Charles Alderson, Lancaster University, UK

Contributors:
Ute Knoch, University of Melbourne
Ari Huhta, University of Jyvaskyla, Finland
Candace Farris, McGill University, Canada

Discussants:
John Read, University of Auckland, New Zealand
Dan Douglas, Iowa State University

This symposium addresses challenges raised by the high stakes nature of the testing of proficiency in the language of aviation. The International Civil Aviation Organisation (ICAO) has developed a rating scale which seeks to define proficiency in the language needed for aviation purposes, at six different levels. Pilots, air traffic controllers and aeronautical station operators are required to achieve at least Level 4 on this scale in order to be licensed to fly aeroplanes or control air traffic on international (cross-border) flights, originally by 5 March 2008. However, by 2007 it became apparent that many Contracting States of ICAO were not going to be in a position to comply with this requirement by the deadline. It was agreed, therefore, that the March 2008 deadline could be extended to March 2011. Previous research (Alderson, 2008, Alderson and Horak, 2009) concluded that this lack of compliance gives cause for concern, that ongoing monitoring of implementation plans and compliance is essential, that there is reason to suspect that the ICAO’s recommendations with respect to evidence for test quality are not being taken seriously, and that we can have little confidence in the quality of several of the aviation language tests and assessment currently available for flight crew and air traffic controller licensure. This symposium will report on studies that investigate the quality of aviation English tests, problems in the development of such tests, the validity of the ICAO Language Proficiency Requirements (LPRs) and the need for better informed definitions of the construct of effective communication in non-routine situations. Paper One, by Alderson, reports on a study which indicates the state of readiness for the implementation of the ICAO requirements one year before the deadline of March 2011. Paper Two, by Huhta, describes the development of a new test of aviation English by the Finnish Civil Aviation Authority (FCAA) and discusses the potential threats to the validity of the testing inherent in both the overall system and the new FCAA test. Paper Three, by Knoch, presents a study which aimed both to establish the ‘indigenous’ assessment criteria (e.g. Erdosy, 2007) pilots use when assessing the language ability of peers, and to establish whether Level 4 is, in the pilots’ opinion, sufficient as the operational level. Paper Four, by Farris, argues that aviation language should be assessed using tools that identify the factors that are critical to effective communications in non-routine situations, and applies these factors to the analysis of a major aviation accident for which miscommunication due to insufficient proficiency in English was cited by accident investigators to be a contributing factor.
Abstracts – Research Papers

Paper Session 1 (plenary)
Time: Wednesday, 14 April, 10.30–12.00

Basic language cognition and the construct of language proficiency: Theoretical and empirical perspectives
Jan Hulstijn, University of Amsterdam

In line with the theme of LTRC 2010, concerned with levels, domains and frameworks of language proficiency (LP) for language assessment, the proposed paper addresses the question of what LP is both theoretically and empirically.

In the first part of the presentation, I offer a definition of adult native speakers’ Basic Language Cognition (BLC). In contrast to recent and current views on LP (Canale & Swain, 1980; Bachman, 1990, 1996; Common European Framework of Reference, Council of Europe, 2001), my proposal places linguistic knowledge and skills at the center of LP, assigning peripheral status to sociolinguistic, pragmatic, strategic and metalinguistic competences.

Empirical support for the claim that LP of adult native and nonnative language users differs first and foremost in linguistic knowledge and skills comes from several studies that have been and are being conducted at the University of Amsterdam. These studies concern the L1 and L2 reading and writing skills of 300 high school students in the Netherlands, who were tested in three consecutive school years (study 1); the L1 speaking skills of 150 adult native speakers of Dutch, differing in age and level of education/profession and the L2 speaking skills of 200 adult nonnative speakers of Dutch (studies 2a and 2b); and the L1 listening skills of 250 adult native speakers of Dutch, differing in age and level of education/profession, and the L2 listening skills of 150 adult nonnative speakers of Dutch (study 3). In all three studies, a number of linguistic subskills were tested. The test batteries included not only traditional paper-and-pencil tests of vocabulary and grammar knowledge but also computer-administered reaction-time tests requiring oral and written word recognition, word recall, recognition of words in connected speech, verification of sentence meaning, and placement of sentence constituents in the right order. Performance in tasks of reading and writing (study 1), speaking (studies 2a and 2b), and listening (study 3) was regressed on performance on the linguistic knowledge and skill tests.

One of the most salient findings of these studies is that performance on paper-and-pencil knowledge tests explained large amounts of variance in the reading, writing and speaking tests; although most speed measures were substantially associated to the reading, writing and speaking skills, they explained little or no unique variance beyond the variance explained by the knowledge tests. [At the time of submission of this paper proposal, data collection in study 3, concerning listening, is still in progress.]

In the third part of the paper, I argue that, although the notion of “level of language proficiency” cannot be upheld on theoretical or empirical grounds, it can be used for purposes of L2 instruction and testing. It should be possible to reach consensus on a sufficiently precise description of the linguistic knowledge and skills required at lower proficiency levels (A1, A2, and B1 in the CEFR). At higher levels, however, it will only be possible to characterize required knowledge and skills by means of typical examples, taking into account test takers’ education, occupation and interests. “The” native speaker does not exist and hence there is no single highest level of L2 proficiency for L2 learners to attain or approximate. For language assessment, I argue in favor of testing L2 users’ linguistic subskills (discrete-point testing) both separately from and prior to testing functional communicative skills in quasi-authentic tasks (integrative-skill testing), in the spirit of Carroll (1961).
A mixed methods approach to construct definition: Identifying underlying factors in L2 oral interactive tasks for nurses in a minority setting

Carolyn Turner, McGill University
Michel Laurier, Université de Montreal
Talia Isaacs, McGill University

One of the most demanding situations for members of linguistic minorities is a conversation between a health professional and a patient (Bender et al., 2004), a situation that frequently arises for linguistic minority groups in Canada. 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. There is a growing desire among health practitioners to accommodate their patients’ language of preference, and therefore, to improve their own L2 proficiency. – Situated in a language for specific purposes context in Quebec, this study draws on input from nurses on their workplace communication needs (i.e., indigenous assessment), on the path toward developing a nursing-specific oral interaction tool for formative assessment purposes (Douglas, 2000). The focus of this methodological paper, however, is to explicitly examine the research design that was used to define the construct(s) underlying nurse-patient communication and to relate the uniqueness and novelty of the mixing that occurred in both the data collection and analysis phases of the study. The goal of the study was threefold: (1) to identify the nature of the language that nurses need in their L2 oral interactions with patients, (2) to define the traits or constructs underlying the nurse-patient interactive tasks, and (3) to calibrate those tasks onto a well-known, previously validated rating instrument commonly used for workplace purposes, the Canadian Language Benchmarks (CLB). – In order to adequately address the research problem, reliance on a single method would have been too restrictive. Rather, in our need to draw on multiple sources of evidence to address our research questions, we embraced a pragmatic approach (Tashakorri & Creswell, 2007). This is to say that our research questions, the exploratory nature of our inquiry, and our need to triangulate our data ultimately dictated the methods we used. – We employed a sequential, exploratory mixed methods design (Creswell & Plano-Clark’s typology, 2007), where earlier stages of the study informed all subsequent stages. The first stage was qualitative, and resulted in the identification of 19 relevant nurse-specific interactive tasks. This included: an extensive literature review; the development of an initial list of relevant tasks; and validation of this list with a focus group of in-service nurses. The retained tasks were then developed into a questionnaire and piloted through the use of verbal protocol. This resulted in the development of a final bilingual questionnaire. – The second and final stage of the study, the quantitative stage, was informed by the qualitative evidence that had been collected up to this point. First, the questionnaire was administered to 127 Quebec nurses, who rated each task for language difficulty. Next, Rasch analyses and descriptive statistics were employed for the purpose of item analysis and data cleaning. Finally, the underlying factor structure was identified in exploratory followed by confirmatory factor analyses. All 19 speech tasks were retained. The manner in which the items clustered was readily interpretable when aided by the previous qualitative results in stage 1. The overarching factor was labeled ‘health-specific verbal interaction with patients.’ This construct loaded heavily on two subconstructs, labeled ‘emotional aspects of caregiving’ and ‘factual aspects related to the patient’s medical condition,’ but was weakly negatively correlated with a third subconstruct that dealt with ‘routine aspects of communication’ but was not healthcare specific. Finally, we examined the Rasch indices and item distributions to calibrate each interactive task onto the appropriate CLB level. – While much language testing research makes use of multiple methods (Taylor, 2005), the methodological framework that underpins the actual mixing is rarely emphasized in published research. L2 researchers often describe the method in the detail required for replicability; however, the methodOLOGY (theoretical underpinning) is usually not described in much depth. Bachman states that "empirical research in applied linguistics has... much to learn from the literature on mixed methods research" (2006, p. 201). Based on our exercise in examining the novel ways we mixed methods in this study, we agree that language testers would benefit from being exposed to research within this “third paradigm” (Teddlie & Tashakkori, 2009).
Performance decision trees

Glenn Fulcher, University of Leicester
Fred Davidson, University of Illinois at Urbana-Champaign
Jenny Kemp, University of Leicester

The scales and descriptors of the Common European Framework (CEF) were created using a measurement-driven approach that prioritizes the ordering of descriptor fragments onto a single scale. Meaning is derived from the scaling methodology and the agreement of trained judges to the place of any descriptor on the scale. This paper analyses the service-encounter scale in the CEF to show that it does not reflect what happens in actual service encounters, because the design methodology does not take actual communication into account. The argument is constructed by comparing the CEF descriptors with an analysis of new empirical data of service encounter discourse, as well as existing literature on the service encounter as a verbal genre. We contend that measurement-driven approaches generate impoverished descriptions of communication, while performance data-driven approaches have the potential to provide richer descriptions that offer sounder inferences from score meaning to performance in specified domains. We demonstrate how the analysis of performance data from travel agency encounters can be used to construct a new scoring instrument that we call a Performance Decision Tree (PDT). This approach to scoring avoids reified, linear, hierarchical models that purport to describe typical performances at arbitrary levels. Instead, PDTs offer a boundary choice approach on a range of observational discourse and pragmatic variables. The resulting score is more meaningful because of its origin in performance data, and provides rich diagnostic information for learning and teaching.
Paper Session 2 (plenary)

Time: Wednesday, 14 April, 3.00–4.30

Which aspects of fluency should be taken into account when describing oral proficiency levels?

Nivja De Jong, Utrecht University

In language testing practice, measures of oral fluency are often used to determine levels of (oral) proficiency. However, from both research on native and second language speech we know that measures of fluency are not only related to linguistic proficiency, but also related to individual characteristics (e.g., extraversion). In this study, we will disentangle the share of variance that can be explained by individual characteristics from the share of variance that can be explained by second language proficiency for the construct of fluency. Tavakoli and Skehan (2005) have noted that fluency is multifaceted in nature and distinguish three main aspects of fluency, with several measures for each aspect. The three aspects are breakdown fluency (e.g., number and length of pauses), speed fluency (e.g., speech rate), and repair fluency (e.g., false starts and repetitions). In our study, twenty-nine intermediate to advanced speakers of Dutch, with native language English, performed four speaking tasks in their second language and four similar speaking tasks in their native language. The tasks in L1 and L2 were matched on topic, formality, complexity, and discourse type. The tasks elicited long turns (up to two minutes) in discourse, where participants were directing their contribution at the computer. We calculated several measures for each of the three aspects of fluency (breakdown fluency, speed fluency, and repair fluency) using both transcripts and automated methods. For each measure of fluency, we compared the amount of variance that could be explained by proficiency with the amount of variance that could be explained by individual characteristics. We took the difference between L1 and L2 fluency to reflect proficiency. For example, if participants have longer pauses in their L2 compared to their L1, we can hypothesise this difference must be due to lack of L2 proficiency, because we used matched tasks in L1 and L2. At the same time, we gauged the amount of variance that could be explained by individual characteristics by correlating L1 with L2 measures. The results showed that the contribution to explained variance differed for each measure of fluency, where some measures were mainly related to the difference with L1 measures, and others showed high correlations between L1 and L2. We will argue that when deciding whether someone is (or is not) at a certain proficiency level, only those aspects of fluency that are mainly related to proficiency should be taken into account and should be included in descriptions of proficiency levels.

Comparing human and machine judgments of collocations and relating them to speaking proficiency

Jing Xu, Iowa State University
Xiaoming Xi, Educational Testing Service

SpeechRater is an automated scoring system used to score the TOEFL® iBT Practice Online (TPO) Speaking Test. As the current system uses a subset of the criteria evaluated by human raters to predict human scores, it is important to develop new features to expand its construct coverage and improve its score prediction. The new speech features investigated in this study were non-native English speakers’ use of collocations in spontaneous speech. Being able to use collocations effectively is a defining aspect of native-like fluency (Nesselhauf, 2004) and can increase speech automaticity (Martyska, 2004). Effective use of collocations is also perceived as a marker of advanced speaking skills (Towell, et al. 1996; Luoma, 2004). In addition, empirical research has supported a strong correlation between L2 knowledge of collocations and L2 speaking proficiency (Sung, 2003; Chiu, 2006; Hsu & Chiu, 2008). However, it remains an open question as to how actual use of collocations in L2 speech is related to overall speaking proficiency. The goals of this research study are twofold. First, we aim to evaluate the robustness of a collocation tool, a computer system developed to detect collocation errors in L2 writing, for use with L2 spoken data. A total of 555 human transcriptions of TPO Speaking responses were processed by the collocation tool and annotated by 6 trained linguists for acceptability of collocations. The performance of the tool was evaluated based on adjudicated human judgments in 1) extracting valid collocation strings of 7 syntactic patterns...
from the TPO Speaking responses, and 2) distinguishing between good and erroneous collocations. Human-machine agreement was compared. The second goal is to determine the relationship between the collocation features we have developed and human holistic scores. Such features include, for example, good or erroneous collocations per 100 words and the ratio of erroneous to good collocations. The performance of the collocation tool will be discussed along with the relationships between human holistic scores and collocation features computed based on human and machine judgments respectively. The suitability of the collocation features for SpeechRater will also be addressed. 


Turn, turn, turn again
Michael McCarthy, University of Nottingham

In this paper I will briefly summarise the literature on turn-taking in everyday conversation, both in the native-speaker and non-native speaker domains. I will propose that turn construction is the locus of interactional competence as well as a key contributor to fluency. Using several spoken corpora, I will examine in particular the linguistic patterns that occur around turn openings in the everyday speech of native English speakers, as well as in spoken academic contexts and in learner oral examination contexts (e.g. IELTS, FCE). Recent work on turn construction (Tao, Evison) has underlined the importance of turn-openers in performing a variety of functions, including organisational and interactional ones. I look at the distribution of turn-openers in the data, comparing casual conversation with academic seminar and small-group talk and examination data. There is an overwhelming preference at the turn opening for items which link with the previous speaker’s turn, suggesting that turn-openers not only establish and maintain interaction but also contribute to creating the ‘flow’ that is the mark of successful multi-party talk. In the examination data, I look at the effects of task on turn construction and how task affects the choice of turn-openers. I conclude that evidence of turn-construction supports the notion of ‘confluence’ in talk and is a key feature of collaborative learning. I also conclude that awareness of turn construction phenomena should feed into task construction in testing and assessment.
Fitting the framework: How raters co-construct understandings of domains and levels from different educational and assessment contexts

Alan Urston, Hong Kong Polytechnic University
Carrie Tsang, Hong Kong Polytechnic University
Felicia Fang, Hong Kong Polytechnic University

The processes for relating language assessments to external frameworks have been reported to an increasing degree since the publication of the preliminary pilot version of the Manual Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF) (Council of Europe, 2003). The Manual outlines the four phases of relating language assessments to the CEFR of Familiarisation, Specification, Standardisation and Empirical Validation. The fourth phase is covered in quite extensive detail in the Manual and its accompanying Reference Supplement with a focus on methods of standard setting for item-based tests. For performance tests of spoken and written language, the Manual is less clear as to how empirical validation should be carried out and so the third phase, Standardisation, in which a process of benchmarking of spoken and written performances from the target assessment to CEFR levels is carried out, is as far as the process goes. Beginning in January 2008, researchers at the Hong Kong Polytechnic University conducted an eighteen-month project to explore the potential of linking the Graduating Students’ Language Proficiency Assessment (GSLPA) to the CEFR by following the linking procedures outlined in the Manual. Working with a relatively small team including seven experienced GSLPA raters, the researchers adopted a ‘negotiation-discussion’ approach in which each step of the process was fully discussed so that decisions could be made on which of the procedures outlined in the Manual could be followed directly, which might need to be adapted, and which would and could be omitted. In this way, each member of the group (rater or researcher) was able to be fully engaged in the process so in a sense validating the process itself, given the experience and expertise of each individual. The group met twenty times during the project for approximately three hours each time. Each meeting was video recorded and transcribed. Notes were taken of each meeting and disseminated to the participants prior to the following meeting to allow for reflection and re-evaluation and participants were required to complete feedback questionnaires after each meeting. This paper focuses on the standardisation of spoken performances so as to claim linkage of the spoken language module of the GSLPA to the CEFR. The presenters describe raters’ rating processes as they verbalised them, aiming to provide insights into the cognitive processes that raters engage in while rating spoken performances, an area that has been under-researched to date. The presenters also discuss the raters’ interactions amongst themselves and their perceptions of the usefulness of CEFR rating scales for rating GSLPA performances and GSLPA tasks. The paper highlights the particular difficulties encountered when trying to relate performances on tasks specific to certain domains to descriptions of language use and proficiency from quite different domains, a practice becoming increasingly prevalent as the field of language learning, teaching and assessment moves inexorably towards a common framework of reference. Reference Council of Europe (2003). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF): Manual. Language Policy Division: Strasbourg.

Can we get a little higher?: Proficiency levels in aviation English

Carol Lynn Moder, Oklahoma State University
Gene B. Halleck, Oklahoma State University

Performance-based testing of English for Specific Purposes focuses on the identification of situation appropriate language functions and tasks that allow generalizations to real world contexts (McNamara 1996, Douglas 2000. Appropriate ESP assessment practices also require that cutoffs and proficiency levels be determined with reference to the target language work domain. In the assessment of Aviation English, general testing criteria have been mandated by the International Civil Aviation Authority (ICAO Document 9835) with proficiency levels specified by the ICAO Proficiency rating scale. Although the ICAO rating scale references the ability to communicate effectively on “work related topics,” there is little research available to aviation agencies and testers on the actual language
used in authentic Air Traffic Control (ATC). Some studies have surveyed the relative importance of prescribed ATC tasks using standard ICAO phraseology (Teasdale 1996, Ramos, Chatham, Henning, Thomas, & Mogilks 1999) while others have examined specific aspects of authentic ATC discourse (Barshi 1997, Linde 1988, Nevile 2004, Wyss-Buhlimann 2005). Only one small study has investigated the grammatical structures needed in authentic ATC communication (Mell 2004). This lack of information is particularly problematic in the assessment of structure, where the scale indicates that “relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task.” (ICAO 9835, p. A-8). This paper reports on a corpus analysis of authentic ATC communication which identifies common structures necessary for the successful performance of work-related aviation tasks. The study shows that prescribed ICAO phraseology identifies a very limited range of grammatical structures and lexical domains that would be insufficient for many critical contexts. On the other hand, basing ratings on non-situation specific language structures would result in assessment of many structures not critical to work performance. This paper provides key information to language assessors on the structures relevant for the determination of situation-appropriate language proficiency levels.
Paper Session 4 (parallel)

Time: Thursday, 15 April, 8.35–10.15

Session 4A

Assessing young learners’ writing – using complementary validity evidence to evaluate the quality of two L2 rating scales
Ari Huhta, University of Jyväskylä
Mirja Tarnanen, University of Jyväskylä

This paper is a part of a project called CEFLING (Linguistic Basis of the Common European Framework for L2 English and L2 Finnish) which is a three-year Finnish Academy funded project. The main aim of CEFLING is to find out how L2 English and L2 Finnish learners’ proficiency develops from one CEF level to the next. The data of the study is based on learner texts written by pupils in grades 7–9 (ages 13–16). Each participant (English n = 250; Finnish n = 230) completed four writing tasks (e.g. e-mail, narrative, argumentation); in total, the learners wrote 1,920 texts. Student performances were rated with two scales, a CEFR scale and a Finnish application of the CEFR scale. The CEFR scale was in fact a compilation of several writing scales that appear in the Framework. The Finnish scale is an adaptation of the 6-point CEFR scale and is part of the national foreign and second language curriculum. Its content is slightly modified and each CEFR level is divided into sublevels (e.g. A2 into A2.1 and A2.2) to provide learners and teachers with more quickly attainable targets. The English performances were rated by four raters and the Finnish by three; the ratings were analysed with Facets. In addition, the raters completed a feedback form consisting of Likert-scale statements covering e.g. their previous experience in rating, perception of their leniency / severity, and usability of the scales. Six of the raters were also interviewed and the interview data were analysed in terms of content. The paper makes use of the results of the statistical analyses, rater feedback and interviews to answer the following questions. How did the raters comment on the differences in the contents of the scales? Were the same performances rated at the same (major) CEFR level with both scales, that is, were the level boundaries in the same place in both scales? How did the raters’ comments correspond with the results of the analyses? Did the raters’ behaviour differ in general in the two scales? Finally, the paper discusses the overall usability of the two scales and, in particular, the implications of the study on the validity of the Finnish application of the CEFR scale, which is intended to be used as a rating tool in the schools throughout the country.

The interaction patterns in paired-assessment among young learners and their effect on performance evaluation
Yuko Butler, University of Pennsylvania
Wei Zeng, University of Pennsylvania

Foreign language education at primary schools has been popular in recent years across the globe. However, it is not yet totally clear what accounts for foreign language proficiency among young learners (Johnstone, 2000), and in turn, we still do not know how best to assess young learners’ performance in foreign language. As an attempt to incorporate assessment as an instructional tool as well as a measurement tool, the present study examines the interaction patterns during paired-assessment among young learners and focuses on the role of scaffolding in the interaction. Drawing on sociocultural theory, the study aims to describe how scaffolding was employed in teacher-student dyads as well as student-student dyads, and how such scaffolding may influence the students’ performance in the paired-assessment (i.e., the diversity of speech content, the variety of language use, the use of L1, and the task completion). The study also examines how such interaction was evaluated by the participating teachers as well as the students. Twenty-four 6th grade students who learn English as their foreign language in China and their teachers participated in the study. The students were first asked to fill in a survey concerning various affective factors related to English learning. The students then engaged in two sets of collaborative assessment tasks; one set was with their teachers and the other set was with their fellow students. Each set was composed of two different tasks; one is a simple information gap task and the other is an open-ended problem-solving task. The interaction during the tasks was video-taped and transcribed. After each task, the students
self-evaluated their performance. The teachers were also asked to evaluate the students’ performance in each of the assessment tasks. It was found that the types of scaffolding and the amount of scaffolding were different depending on the types of dyads as well as the types of tasks. The diversity of speech content and the variety of language use observed as a result of such scaffolding were also different accordingly. The students’ general proficiency level as well as their affective factors also influence the way that they interacted with each other in complicated ways. The study also found that both teachers and students reached different evaluation depending on the different pairing patterns in the paired-assessment. The study concludes with various implications for using paired-assessment for young learners, including a possibility of incorporating the notion of ZDP in assessment.

The CEFR: Its virtues and shortcomings when teaching and assessing young learners
Hanan Khalifa, Cambridge ESOL
Szilvia Papp, Cambridge ESOL
Angela ffrench, Cambridge ESOL

The Common European Framework of Reference (CEFR) is currently gaining momentum and political stance in language education and assessment not only in Europe but worldwide as well. Some stakeholders within the educational field and the assessment community regard curriculum and examination alignment to the CEFR as an essential quality mark. Although the CEFR was not intended to be a frame of reference for describing and assessing young learners, this paper explores the extent to which a demonstrable link can be established between examinations developed for young learners (aged 7–12) and the CEFR. The examination considered is Young Learners of English Suite of Examination (YLE) which assesses proficiency levels corresponding to CEFR levels Breakthrough (A1) and Waystage (A2) as well as a level assessing proficiency at pre-breakthrough.

A mixed method approach was used to investigate the alignment of YLE to the CEFR. The approach involved the used of anchoring procedures, carrying out a content mapping exercise and conducting a standard setting study using the analytical judgement method for the productive skills and modified Angoff for receptive skills. The presenters will start by providing a brief literature review of similar alignment studies conducted and their findings. They will then provide a short description of the examinations used in the study and elaborate on the abovementioned methodology. They will move on to report on the study findings and limitations. In doing so, they will share one of the outcomes of the study which is the development of a set of Can-Do statements that describe children’s language knowledge, skills and strategies. These statements are perceived as contributing to the student centered assessment practices. The presenters will end up by reflecting on the role of the CEFR within a young learner teaching and assessment context.

Session 4B

Multi-competence and assessment: Recognizing variability in the L1 benchmark
Amanda Brown

“Multi-competence” (Cook, 1992) refers to the multiple competencies exhibited by multilinguals which differ from the single competencies exhibited by monolinguals. This construct is relevant for assessment of a second language, where the benchmark against which L2 performance is measured is often the monolingual “standard”. This paper questions the efficacy of such practices by highlighting variability in native speaker performance as a function of second language experience. The study examines construal of motion, specifically the expression of Manner and Path, where substantial crosslinguistic differences lie (Slobin, 2004; Talmy, 2000). Narrative descriptions were elicited from monolingual Japanese speakers (n=15), monolingual English speakers (n=13), and native Japanese speakers with intermediate knowledge of English as a second language (n=28) in their L2 but also in their L1. Results showed that monolingual Japanese and English speakers patterned similarly with respect to clausal packaging of motion descriptions, that is, information about Manner and Path of motion generally appeared within a single clause as shown in (1) and (2). (1) [sakamichi-o koroge-ochite] hill-ACC roll-descend ‘he rolls down the hill’ (2) [he rolls down the street] However, L2 speakers patterned differently in their L2, but also in their L1, tending to distribute Manner and Path information across separate, but grammatical clauses as shown in (3) and (4). (3) [he rolled with a ball
in his stomach] [and went straight to the bowling] (4) [korokoro korogatte] round.round roll [bouringu jou ni itte] bowling alley to go ‘(he) rolls and goes to the bowling alley’ Insofar as L2 speakers differ from monolingual speakers of the target language, this is typically interpreted as non-target-like performance. However, when the same speakers show parallel differences in their L1, we begin to question the notion of what constitutes target-like performance. Findings will be discussed in relation to assessment and the pedagogical norm (Gass et al., 2003).


Linguistic characteristics of NS and NNS responses to impromptu essay prompts

Sara Weigle, Georgia State University
Eric Friginal, Georgia State University

In this paper we present the results of a corpus-based analysis of the linguistic features of native and non-native speaker responses to two TOEFL iBT Independent writing tasks (30-minute impromptu essays). The study addresses two important issues in second language writing assessment: first, an analysis of text features of essays at different score points helps to identify the features of responses that are most relevant in discriminating among proficiency levels. Second, the study addresses the problematic issue of the ‘idealized native speaker’ as a target for language learning and assessment by exploring the degree to which native and non-native speakers’ essays share similar features. (See, e.g., Davies, 2003, for a full discussion of this issue.) Responses to two different writing tasks were collected from 386 NNS and 150 NS university students and scored by two trained raters using the TOEFL rubric. Using a variety of corpus tools, the following linguistic features were analyzed:

- Lexico/Syntactic Complexity: Prepositions; coordinators/conjunctions; complement clause constructions; quantifiers; transitions (adverbial connectors, metadiscourse markers)
- Expression of Stance: Explicit expressions of assessments, evaluations, synthesis; use of markers of intensity and affect; use of discourse markers
- Vocabulary Use: Vocabulary size (type/token ratio); average word length; content word classes (e.g., nouns, verbs, adjectives, adverbs); nominalizations; hedges/vague references/keywords
- Comparison of Informational Content: Discourse markers of elaboration and informational content; distribution of semantic classifications of nouns, adjectives, and adverbs
- Formal vs. Informal Stylistic Features: Pronouns (especially first and second person pronouns); contractions; tense/aspect shifts; that deletion; private verbs (e.g., believe, think)

Results of these analyses are compared across language groups (native vs. non-native speakers), proficiency levels (ratings) and prompt. Implications for rater training, task design, and test preparation will be discussed.

Validation of a benchmarking study by means of repeated measures classification consistency

Rebecca Present-Thomas, Knowledge Technologies, Pearson
Alistair Van Moere, Knowledge Technologies, Pearson

Increasingly, tests are linked to frameworks or descriptors to show how they relate to a criterion. In the United States, one such framework is the National Reporting System’s (NRS) Educational Functioning Levels. The NRS is the accountability system for the U.S. federally funded adult education program. Approximately half-a-million English language learners are enrolled in adult education programs every year. All students take an approved test at the start and end of their course which classifies them according to one of six Educational Functioning Levels. By this method, student learning gains are measured and reported to the central government (NRS, 2007). Several tests have been recognized by the NRS as suitable pre- and post-test measures for showing learning gains according to the Functioning Levels. This paper reports two analyses of the NRS speaking and listening scales. The
first was a classification consistency study. An NRS approved test, the BEST Plus, was compared with the Versant English Test (VET). Participants of 26 different language backgrounds representing the target population (n=151) completed four counter-balanced tests consecutively: two BEST Plus interviews and two Versant English Tests. Both tests were internally reliable (r = .90 to .94) and fair predictors of each other (r = .74 to .78). However, it was difficult for either test to consistently classify individuals according to the established NRS levels. For example, less than half of the students who were placed in NRS levels 2 through 6 on their first BEST Plus test were classified into the same level on their second test. The NRS bands were so fine-grained as to overlap with measurement error even on these two reliable and standardized tests. Next, a qualitative analysis of the NRS scales was undertaken. The descriptors reveal that Level 6, “Advanced ESL”, describes a basic level of proficiency: “Can understand the main points of simple discussions and informational communication in familiar contexts. Shows some ability to go beyond learned phrases patterns and construct new sentences.” Five linguists who were shown NRS and CEFR descriptors judged that NRS Level 6 was no higher than B1 on the CEFR. In other words, the entire NRS framework fits within just two or three CEFR bands. The study raises questions about how many levels (and dimensions) of language proficiency can be defined such that multiple tests can be reliably benchmarked to the levels. Further, a method for independent validation of existing benchmarks is proposed.

Session 4C

Predicting difficulty in PISA reading items: The process of describing item difficulty

Tom Lumley, Australian Council for Educational Research
Juliette Mendelovits, Australian Council for Educational Research
Dara Searle, Australian Council for Educational Research
Irwin Kirsch, ETS
Charles Alderson, Lancaster University
John de Jong, Language Testing Services
Dominique Lafontaine, Université de Liège

The Programme for International Student Assessment (PISA) surveys 15-year-olds in several domains, including reading, every three years. In the current cycle, a single PISA instrument will be administered in 74 countries and 45 languages. PISA is administered in the language of education, normally but not necessarily the student’s first language. In PISA 2009 reading literacy is defined as “understanding, using, reflecting on and engaging with written texts, in order to achieve one’s goals, to develop one’s knowledge and potential, and to participate in society.” Results are reported against a reading scale with five levels, each of which occupies a score range based on converted logit values. The scale includes a description of the kinds of tasks students at a given level are likely to be able to perform. Subscales also allow reporting against two other features of the reading framework: aspect (the cognitive processes associated with individual reading items) and text format (continuous text, or prose; and non-continuous, including diagrams and graphs). The PISA reading framework describes features of reading that drive difficulty, derived from empirical and theoretical work conducted by Kirsch and colleagues (e.g. Kirsch & Mosenthal, 1990; Kirsch, deJong, Lafontaine, McQueen, Mendelovits & Monseur 2002), but it has been unclear how useful these features are in accurately predicting the difficulty of individual reading items in PISA. The PISA reading scale provides broad descriptions of skill levels associated with reading items, intended to communicate to policy makers and teachers about the abilities of students at different levels. However, the described scales are not explicitly tied to features that predict difficulty. For PISA 2009, an empirical study attempted to provide a systematic description of features that explain difficulty of reading items. In a pilot study, four judges rated 100 print reading items from the 2009 Field Trial (FT09) according to two sets of variables, relating to aspect and text format. The ratings were compared to item difficulty values obtained from FT09. This approach was limited in its ability to predict difficulty, and analysis suggested possibilities for improvement. This paper describes the steps taken to extend the range of variables used in this process, including clarifying each variable, and to ensure commonality in the process judges use in their ratings. Additional rounds of data were collected and analysed. The usability of this model is examined, including analysis of the levels of agreement shown by raters applying it to PISA items;
the amount of variance in difficulty predicted by raters' judgements; and identification of variables that appear to be more and less successful predictors of difficulty.


The dimensionality of test scores on a second language reading comprehension test: Implications for accurate estimation of a test-taker's level of L2 reading comprehension
Youngsoon So, University of California, Los Angeles

This study is to check the dimensionality of scores on a passage-based second language (L2) reading comprehension assessment, where items are nested within a passage, and consequently, two fundamental assumptions of current measurement models – the local independence and unidimensionality assumptions – are likely to be violated. Four confirmatory factor analytic models – one-factor, correlated-factor, bi-factor, and correlated-uniqueness models – were compared to investigate the best factor structure of scores on the reading comprehension section of the Certificate of Advanced English (CAE). The statistical results, when interpreted in conjunction with substantive considerations, indicated that the bi-factor model is the most appropriate model to explain the data, with the loadings of the passage factor being significant on most of the items analyzed. Furthermore, the size of the passage factor was found to be either similar to or greater than the size of the main construct (i.e., reading comprehension) on many of the items. This result implies that current measurement models may not be effective to accurately estimate test takers' ability levels in L2 reading comprehension on the basis of sets of items scores that violate the mathematical assumptions of these models. This study then conducts content analyses of the test items and the passages to better understand the nature of the passage effect on test takers' performance on the test. This presentation will discuss both theoretical and practical implications of the study, focusing on suggestions for (a) the ways in which to design measures of English as a Second language reading comprehension and (b) how to interpret and use scores from such tests in order to make an informed and valid inference about test takers' ability levels.

Evidence of lack of processing of multiword lexical Items in reading tests
Ron Martinez, University of Nottingham

To what extent do students notice multiword formulaic sequences in reading tests, and to what degree do such items affect comprehension? The preceding questions formed the basis of a research undertaking that aimed to better understand what effect – if any – the presence of idiomaticity in text has on L2 reading comprehension. Using a within-groups, repeated measures design, a two-part test of reading comprehension revealed not only a significant difference between levels of comprehension when texts contained multiword expressions (even when the texts themselves in each test part were written using the exact same pool of high-frequency words), but also that the candidates consistently had an inflated view of how good their comprehension was of the texts that contained idiomaticity, via a self-reported measure of comprehension. In brief, on the basis of the results of the study, it can be asserted that even learners at higher proficiency levels tend to not notice many multiword expressions in text, and when they do are generally not good at guessing what they mean in context. Moreover, there is a tendency for candidates to overestimate how much they understand as a function of those unnoticed and/or misunderstood expressions. Therefore, drawing on the results of this study, the presenter will posit that 1) multiword expressions are prevalent in naturally-occurring texts, and those expressions are mainly comprised of the most common words in English – though the meaning of those individual words does not always match the meaning of the words in combination – and therefore can be ‘deceptively transparent’ (Laufer, 1989); 2) as the results of this study provide evidence that multiword expressions can negatively affect comprehension, a principled methodology for their inclusion in L2 reading tests can positively contribute to the validity of those tests; 3) a reliance on the frequency information of individual words alone to inform the design of a reading test and/or determine the ability of a text to discriminate between thresholds of proficiency can result in a test of limited construct validity; and 4) particularly in determining the thresholds of higher levels
of reading proficiency, a demonstrated recognition and/or understanding of the most common non-compositional multiword expressions can be one way in which we decide when someone is (or is not) at a level. Limitations of the study will be presented, as well as possible directions for future research and applications.
Paper Session 5 (parallel)
Time: Thursday, 15 April, 10.30–12.10

Session 5A

**Using the Common European Framework to rate test taker responses**
Kirsten Ackermann, *Pearson, plc*

The Common European Framework (CEF) describes language proficiency in reading, writing, listening and speaking on a scale comprising six levels from A1 or Basic User to C2 or Proficient User. The individual levels provide guidance to educators to identify existing levels of language learners’ proficiency and to develop course syllabi and teaching materials to advance learners’ level of language competency. The rationale for mapping marking criteria as well as the marking scales onto the CEF is to achieve a robust alignment with the framework from the very beginning of test development. This goes beyond standard setting (mapping of final test scores onto the CEF). Thus the purpose of this research paper is to show how marking criteria based on the level descriptors of the CEF help distinguish one level from another and allow us to decide whether a test taker is at the level he or she was entered for in a general English exam. This paper will describe the approach to using the framework and reflect on its advantages and drawbacks. It will also show how the use of additional language level descriptors that specify the CEF criteria in relation to the test specifications ensures alignment to the CEF. Thissen and Wainer (2001), amongst others, showed that any scoring issue can be related to the accurate representation of candidate performance with respect to the measured constructs or the domain of knowledge, skills, or ability. For this reason, FACETS analysis was conducted. FACETS provides estimates of examinee ability, rater harshness and scale difficulty and maps these facets on a common log-linear logit scale. The results of this analysis show whether the marking criteria and marking scale successfully discriminate between CEF proficiency levels using field test data from the aforementioned general English exam. Preliminary results show that raters using marking criteria based on the CEF can successfully discriminate between CEF levels. The paper will also report on qualitative findings based on post-marking questionnaires and de-briefing sessions with trainee raters. Findings further revealed that improvement of the marking scale is required as well as greater more rigorous quality control during rater training programs and live rating to ensure validity and credibility of the exam. This study aims to enhance the alignment of language tests with the CEF and will thus be of considerable significance for researchers and practitioners in the field of language teaching, assessment and development.

**Getting the levels right: deriving item writer guidelines from CEFR descriptors**
John de Jong, *Pearson Language Tests*

The presentation will describe a set of guidelines developed to help item writers to construct test items whose difficulty levels match the intended CEFR levels of the test. The guidelines were formulated in the course of the revision of a test of general English at six levels corresponding to the six CEFR levels. The guidelines take the form of a series of statements that define various aspects of test items – such as the degree of complexity of a reading passage or the type of information required to answer a multiple choice question. The process of developing the guidelines for a given item type can be viewed as a linked series of questions.

- What language competence(s) is the item designed to assess?
- Which CEFR scale(s) relate(s) most closely to those competences?
- In the wording of the descriptor(s) in the said CEFR scale(s) at the target level, which are the key terms that effectively define the expected level of competence?
- For each of those terms, what is its practical implication for the make-up of the various elements of the test item (the comprehension passage, item stem, multiple choice options etc.)?
The presentation will:

- explain the development process with practical examples
- consider how this process relates to those recommended in the Council of Europe’s Manual on relating language tests to the CEFR (Council of Europe, 2009) and to similar initiatives aimed at determining the CEFR level of test items, such as the Dutch Grid.
- discuss some of the problems of interpretation that arise in the process of formulating the guidelines and propose some tentative solutions
- outline possible avenues of future research findings aimed at validating the guidelines.

An alternative approach to linking a suite of examinations to the Common European Framework of Reference (CEFR)

Gergely A. Dávid, Eötvös University and Euro Examinations, Budapest

This paper discusses a comprehensive approach to linking examinations to the CEFR in the context of Euro Examinations, Budapest. It addresses the issue of how minimum CEFR proficiency levels (standards) may be determined for examinations, i.e. how it can be decided whether someone is at a level or not. The approach relies on the original calibrations (North, 2000) that formed the basis of the level classifications in the illustrative CEFR scales. It has grown out of an attempt to avoid using previously published CEFR-calibrated items and performance samples, as it was rather difficult to find these in sufficient numbers. Examinations may be linked to the CEFR directly, without doing the same through calibrated items and performances from somewhere else. The approach was based on a large number of level judgements by examiners of i) illustrative descriptors, ii) local, “own” items iii) selected local test performances taken from different levels. These judgements were processed with Many-Facet Rasch Measurement (Linacre 2006). As a result, a CEFR-linked measurement scale with bands A1-C2 was created, using common “descriptor” items and main-stream item anchoring methodology. Local items and performances were then directly linked to this scale via step-anchoring, a lesser known technique specifically designed to anchor rating scales. Thus, step anchoring allowed the researcher to determine the CEFR level of local performances. The procedures by which the research was carried out will also be discussed. This approach also provides for the inclusion of languages other than English in contexts where a number of different languages are tested on the basis of the same set of specifications. This project included German and a less commonly taught language in terms of examiners and descriptors. German, for example, was included because the CEFR descriptors for reading had originally been calibrated in German. The approach can also relate different proficiency levels to one another on the A1-C2 scale. Once the judgements about the selected local performances have been anchored to the CEFR-linked scale, they are treated as reference results and are compared with (even interpolated into) tabled test results. They help interpret the levels tested and the level of candidates passed, i.e. help validate minimum passing proficiency levels (passing scores). For example, a B2 reference result of a low scoring C1 candidate would indicate a similarly low level (range of scores) if it is compared with equated results from later examinations, whereas a higher reference score computed to be a borderline can usefully indicate similar borderlines in later tests.

Session 5B

Vocabulary use in writing test tasks performance

Viphavee Vongpumivitch, National Tsing Hua University

This study aims to investigate vocabulary use in standardized writing tests taken from the practice versions of ETS TOEFL iBT and Taiwan’s LTTC General English Proficiency Test (GEPT) high-intermediate level. The writing test tasks include two argumentative writing tasks and two integrated writing tasks. The integrated writing tasks are based on reading and listening materials. It is hypothesized that different task types will require the test takers to use their vocabulary knowledge to different degrees. Participants in this study are Taiwanese English language learners aged between 18–30 who either have taken the TOEFL iBT test or are planning to take the test within one year after data collection. The study asks whether the participants who perform better on the writing tests would
show a more advanced level of vocabulary knowledge, as indicated by their higher vocabulary tests’ scores, and use more academic words in their free production. The study also investigates whether the learners’ ability to use their vocabulary knowledge is influenced by the different writing task types. The participants’ essays are analyzed using the Lexical Frequency Profile (Laufer & Nation, 1995). In addition, the participants also took different vocabulary tests, namely the Vocabulary Levels Test (Schmitt, Schmitt, and Clapham, 2001), the Controlled Productive Vocabulary Test (Laure & Nation, 1999), and Meara and Fitzpatrick’s (2000) Lex30 word association test. The participants also highlighted unknown words on the transcripts of the listening materials used in the integrated writing tasks. Preliminary results based on the TOEFL independent writing task show that the students whose essays were rated as “high”, i.e., receiving the rating of 4–5 on the TOEFL iBT independent writing scale, wrote longer essays (have more tokens) and use more types than the students whose essays were rated as “low”, i.e., receiving the rating of less than 3 on the TOEFL iBT independent writing scale. The essays written by the “high” group also include more word families and more “advanced” words (academic words plus words that are beyond the 2000 level). The “high” group also performed significantly better on each section of Laufer & Nation’s (1999) Controlled Productive Vocabulary test. Similar analyses are being conducted on the other writing tasks to examine whether the results would vary across task types. If time permitted, the quantitative results will be presented together with the information obtained from the interviews conducted with the participants who wrote the essays. The results from this study will shed light on the role of vocabulary knowledge in writing test tasks performance.

Linking TOEFL Scores to the Lexile Measure
Cathy Wendler, Educational Testing Service
Fred Cline, Educational Testing Service
Ellie Sanford, MetaMetrics, Inc.

The TOEFL® iBT measures the ability to use and combine listening, reading, speaking, and writing skills to communicate in an academic environment (Chapelle, Enright, & Jamieson, 2008). The reading measure is one of four sections on the test; examinees receive section scores as well as a total test score. Examinees also receive performance feedback that describes tasks that examinees within a particular score range are able to do. Beginning in fall 2009, examinees will receive a Lexile® measure as part of their TOEFL reading performance feedback. A Lexile measure represents two dimensions: (1) a reader-level measure that indicates a person’s reading ability on the Lexile scale and (2) a text-level measure that represents the difficulty of the text that is read. The same scale is used for both reader ability and test difficulty. When used together, they provide an estimate of the person’s reading level and help determine the appropriate difficulty level of reading material. The higher the Lexile measure, the higher the level of reading ability. In order to provide examinees with meaningful Lexiles, the TOEFL reading scale was linked to the Lexile scale. This presentation will describe the method used to link the scores from the two scales. All examinees registered for several spring TOEFL administrations were invited to participate in the study by completing a Lexile test online prior to their TOEFL administration. The linking study design used four Lexile linking test forms. Each linking test was comprised of 45 items (a short reading selection followed by a multiple-choice sentence completion) with 22 of the items common across the forms. The linking tests were randomly assigned to participants. Data were screened and a participant omitted if lack of effort was apparent (e.g., if the participant spent less than 15 minutes on the entire linking test, etc.). This resulted in a sample of over 1,700 participants with complete Lexile and TOEFL data. The sample reflected the country distribution of TOEFL examinees but appeared to be slightly more able than the total test-taking population. A linear equating approach was used to link the scales. In linear equating, two score scales can be linked when the underlying item response models used to develop assessments are different. The linear equating method is most appropriate when (1) sample sizes are small, (2) test forms have similar difficulties, and (3) simplicity in conversion tables or equations are desired (Kolen & Brennan, 1995). A transformation is chosen so that scores on two tests are considered to be linked if they correspond to the same number of standard deviations above (or below) the mean in some group of examinees. In addition to describing the linking approach, the final linking table will be presented and discussed along with implications for the use of the Lexile measure.
Lexical resources in legal and general English writing across second language proficiency levels

Ivana Vidakovic, University of Cambridge ESOL Examinations
Fiona Barker, University of Cambridge ESOL Examinations

The growing importance of English as a lingua franca in the work place has seen a corresponding demand for English-for-Specific Purposes (ESP) tests. For language testers this growth in ESP assessment has brought about challenges for test design in terms of construct definition and the level of specificity of task input and expected candidate output. This paper focuses on one of the key features of ESP assessment, namely the specificity of lexical resources, and offers an investigation of lexical resources within two contexts – that of domains (general vs. legal) and proficiency levels. The paper also presents a way of exploring lexical progression in L2 writing which moves beyond the limitations of some of the current methods. Within applied linguistics, single-word vocabulary frequency measures have been found to be insufficiently robust for capturing the differences between proficiency levels (Read 2005, Shaw & Weir 2007: 104), and the investigation of multi-word units (e.g. collocations and lexical bundles) has been called for. Studies have explored lexical bundles across disciplines (e.g. Hyland 2008), but not extensively in L2 contexts or across the proficiency continuum. The present study addresses these issues by exploring both individual words and lexical bundles and by identifying the key lexical features that distinguish two domains of English language use at adjacent proficiency levels. The two domains are a legal domain of English language use and a more general one, representing a variety of work and study contexts. This study investigated lexical resources in three exams: an ESP exam (Cambridge ESOL’s International Legal English Certificate), and two corresponding general English exams (First Certificate in English and Certificate in Advanced English), which are taken for a variety of work and study purposes. The exams chosen are at comparable CEFR levels (B2 and C1). The data in the study comprised the written output of 80 candidates (40 from ILEC and 20 each from FCE and CAE). Lexical resources were analysed both quantitatively in terms of corpus-based analyses of individual words and lexical bundles, and also qualitatively in terms of the functional roles of lexical bundles in L2 writing. Features of the candidate output salient in the two domains and proficiency levels will be discussed in the presentation and compared to the existing CEFR descriptions of lexical competence. The presentation will end with a discussion of the implications of the study for building a test validity argument, as well as for ESP teaching and L2 acquisition research.

Session 5C

Investigating participants’ understanding of proficiency levels in different educational contexts

Spiros Papageorgiou, University of Michigan

Proficiency levels such as those in the Common European Framework of Reference (CEFR-Council of Europe, 2001) are increasingly used by examination providers to report and explain test scores to their users. Because the CEFR proficiency levels are widely known, it is assumed that they are understood similarly in different local educational contexts. This paper questions this assumption and explores whether professionals from two different educational contexts share a common understanding of the CEFR levels. 22 standard setting panelists from two groups, one in Latin America and one in Central Europe, sorted 169 CEFR descriptor statements for listening, reading, grammar and vocabulary into the six main CEFR levels (A1-C2). This resulted in more than 3500 level judgments. After they completed the task, the correct level was revealed and a discussion followed, focusing in particular on the misplaced descriptors. The level judgments were analyzed using Classical Test Theory and many-facet Rasch analysis to investigate:

• whether the two groups differed in the way they scaled the descriptors into levels
• whether the two groups were similar in terms of consistency

Notes taken during the group discussions were used to offer qualitative insights into the understanding of the levels by the two groups. The findings of the study challenge the assumption that widely known
User perceptions of academic language proficiency and the TOEFL iBT
Margaret E. Malone, Center for Applied Linguistics
Megan Montee, Georgia State University
Chengbin Yin, Center for Applied Linguistics

The Test of English as a Foreign Language Internet-based Test (TOEFL iBT) assesses the domain of academic language proficiency. Developed from an effort to synthesize theories of communicative language proficiency and academic tasks encountered in university settings, the TOEFL iBT presents students with integrated tasks which require the use of multiple skills. Previous research (Chapelle, Enright & Jamieson, 2007) discusses the validity argument for the test and establishes evidence for score interpretation and the use of the test for university admissions and placement. Now that the test has been operational for several years, additional validity evidence is needed regarding the extent to which various stakeholder groups understand and agree with the construct of academic language underlying TOEFL iBT tasks. This multi-stage, mixed-method study explores user beliefs about the constructs assessed by the TOEFL iBT and the relationship between test performance and academic language use. Participants include three stakeholder groups: students (at both the undergraduate and graduate levels), English language instructors, and university administrators. Participants included students from three cultural contexts: Germany, Korea and Saudi Arabia. The mixed-method study included three phases of data collection. First, university administrators, instructors and students participated in focus groups (N=46). Based on this data, online surveys were developed and administered to all three groups (N=1250). Finally, researchers conducted stimulated recalls with individual students (N=12) from each cultural context. The stimulated recalls provide an in-depth view of students’ perceptions of integrated tasks. The study examines and synthesizes data across groups differentiated on the basis of participants’ country of origin (Germany, Korea, and Saudi Arabia) and the type of language study (English language class participation versus student matriculation in a U.S. university). The presentation will discuss the results of the study with a focus on stakeholder beliefs about the alignment between integrated tasks and participants’ understanding of academic language proficiency. The results will shed light on the extent to which users understand and agree with the construct of academic language that underlies the TOEFL iBT tasks across different contexts. This comparison has implications for the evolving definition of proficiency within different user groups, including university administrators, a user group without extensive research. In this way, the paper will provide additional perspectives on the validity of the TOEFL iBT based on user knowledge and perceptions. Chapelle, C., Enright, M., & Jamieson, J. (2007). Building a validity argument for the Test of English as a Foreign Language. New York: Routledge.

What can a can-do do? Developing descriptors that define levels
Anthony Green, University of Bedfordshire

Definitions of levels provided in the Common European Framework of Reference for Languages (CEFR) are intended to support three assessment-related purposes: 1) specification of the content of tests; 2) criteria to determine the attainment of a learning objective; 3) comparisons and alignments across different systems of qualifications. However, critics have pointed out that the CEFR alone is not sufficient to support these purposes. In response, the Council of Europe has proposed that for specific languages, more detailed specifications or ‘Reference Level Descriptions’ (RLD) are needed. This paper reports on the processes of drafting, refining, elaborating, exemplifying and calibrating descriptors in the development of a RLD for English at the highest (C) levels of the CEFR. Approaches to the definition of learning objectives and proficiency scale descriptors were reviewed and a variety of schemes currently in use in language education in various regions of the world were analyzed to identify ways in which levels are defined. A database of over 1,000 descriptors was built from these schemes and a keywords analysis was employed to identify the tasks, criteria for judging performance and relevant conditions that might constitute consensus level definitions. The approach adopted to the validation of these definitions builds on the content and methods of the CEFR – judgements of descriptor difficulty made by language educators – supplementing this with evidence from learner performance: corpora of written and spoken learner production and of input texts employed in educational materials targeted at the different levels of the CEFR. This evidence indicates
how learners perform the tasks associated with the levels, further elaborating the level descriptions. This process also serves to identify aspects of language use where further evidence is required and where observed learner performance raises questions about the validity of the framework descriptions. Issues and limitations encountered in carrying out this process are described and implications suggested for the use of frameworks in assessment.
Paper Session 6 (parallel)
Time: Friday, 16 April, 8.35–10.15

Session 6A

A multi-dimensional approach to the exploration of rater variability: The impacts of gender and occupational knowledge on the assessment of ESP oral proficiency
Weiwei Liu, The Hong Kong Polytechnic University

The past several decades witnessed the enthusiasm of both the business community and the pedagogical community in ESP assessment. However, as with other types of assessment, care must be taken to minimize the distorting effects of construct-irrelevant variance. A number of sources of error have been discussed in the literature on ESP assessment. Prominent among these sources is the variance associated with raters. A considerable number of studies have thus attempted to approach the impact of rater variability on the rating or scoring from different angles. These studies have produced different or even conflicting findings concerning the impacts of certain individual rater variables. Furthermore, these studies can only be viewed as useful starting points in the process of exploring the impact of rater variability as they have all focused on one certain individual rater variable. The inconsistent results imply that there might be other rater-related factors that may exert an impact on the scoring. In view of the preoccupation with individual rater variables in previous studies that took a one-dimensional approach to the exploration of rater variability, we conducted the present study. The focus was on the possible interaction between two rater variables – gender and occupational knowledge. 180 raters participated in the study. Based on their levels of occupational knowledge, these raters were divided into three groups, good mastery group (GG), basic understanding group (BG), and no understanding group (NG). In terms of gender, raters within each of the three groups were further divided into two subgroups. Two-way ANOVA was performed to compare the ratings provided by the six groups of raters. Results showed that both gender and occupational knowledge significantly influenced the holistic rating and some aspects of the analytic rating. An interaction between raters' gender and level of occupational knowledge was detected in the analytical rating for pronunciation. The study showed that it is necessary to investigate the potential interactions between different rater variables, rather than focus on individual rater variables. Thus the study cast doubt on the conventional one-dimensional approach to the exploration of rater variability. The study also has potential implications for the appropriate use and interpretation of ESP assessment results.

What does it take to make a grade? Teachers assessing grammatical ability in L2 writing with rating scales
Heike Neumann, McGill University

In multi-level university ESL writing programs, promotion from one level to the next rests upon the teacher's assessment of student progress towards the English language proficiency deemed necessary for academic study. An important component of this proficiency is the students' ability to apply their grammatical knowledge to academic writing tasks, in other words what Purpura (2004) calls grammatical ability. When teachers assign grades for this grammatical ability using rating scales, which indicators do they attend to? Do their students know how to satisfy teachers' expectations for grammatical ability? This paper reports on a mixed methods case study designed to investigate the process by which two teachers assess the grammatical ability of 32 students in their ESL writing class. Previous research on differences in proficiency levels and grade structure in L2 academic writing has focused on measuring grammatical ability using only teacher-independent instruments: 1) rating scales with general descriptors or 2) sentence-level accuracy and complexity measures (e.g. Ortega, 2003; Polio, 1997, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). In this study, sentence-level measures are combined with a qualitative analysis of student texts. Thus, the role grammatical ability plays in the construction of meaningful messages and texts and how text construction relates to teacher assessment of this ability can be considered. This has not been done in previous research. In addition, teacher expectations are examined directly through interviews and stimulated recall in order to see which indicators teachers attend to. To analyze which indicators teachers attend to during the assessment process, a text-based analysis of academic ESL essay exams previously assessed by
the course instructors was conducted. First, the level of accuracy and complexity was determined
using established measures. Multiple regression was employed to determine which of the measures
best predicts the teacher’s grammar grade. Second, an information structure analysis at the sentence,
paragraph and text levels using a theme-rheme distinction (Halliday & Matthiessen, 2004) was carried
out to analyze whether the L2 writers’ ability to use grammar effectively in written communication
also influenced the teacher’s assigned grammar grade. To determine the teachers’ expectations
and criteria, teacher interviews and delayed stimulated recalls were conducted. The students’
understanding of the assessment process was explored through a questionnaire administered to all
student participants and through semi-structured follow-up interviews conducted with eight student
participants. The combined analysis of results indicates that teachers rely primarily on indicators of
accuracy when assigning grades and assessing proficiency levels, which students aware of. This
negatively affects the students’ writing and learning process because avoidance of error becomes a
prime objective to make the grade. The teachers’ criteria, therefore, creates negative washback in
these writing courses.

References:
relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics,
An overview of what we investigate and how. In B. Kroll (Ed.), Exploring the dynamics of
Assessing grammar. Cambridge: Cambridge University Press. Wolfe-Quintero, K., Inagaki, S., & Kim,
(Technical Report No. 17). Honolulu, Hi: University of Hawai’i, Second Language Teaching and
Curriculum Center.

Rating L2 pragmatics: What does holistic scoring really mean?
Yuanshan Chen

For the past decades, L2 pragmatics has received increasing attention in SLA research since
applied linguists and language teachers began to realize that linguistic development does not always
guarantee pragmatic proficiency (Bardovi-Harlig, et al., 1991). To evaluate learners’ speech act
performances, interlanguage pragmatics (ILP) researchers in general employ holistic scoring method,
which involves two or three readers giving a quick, impressionistic marking for DCT or roleplay
productions. However, such a scoring instrument often obscures much valuable information since
a lot of times learner productions are so internally complex that they require more than a single
rating to capture strengths and weaknesses. Therefore, this study aims to highlight the limitation
of holistic scoring through raters’ eyes. Five compliment exchanges produced by Taiwanese
university students in response to a DCT situation were sampled. Following the scoring criteria
adapted from Sasaki’s (1998) study, two native speakers of English were invited to rate holistically
in terms of appropriateness and grammar. Each exchange received two separate ratings—one for
appropriateness and one for grammar—ranging from Excellent (level 4) to Poor (level 1) from each
rater. After they finished scoring, the raters reported the rating process in retrospect. The protocols
showed that although appropriateness and grammar were rated separately, there was interplay
between these two categories in at least two ways. First, when it comes to assigning levels to both
categories, it was difficult for them to decide if an error was appropriateness or grammar failure.
Second, in some cases, the learners may have covertly intended to give compliments, but their intent
was blocked or misled by grammatical deficiency. These findings suggest that pragmatic competence
involves different components, which interact closely with each other and that a learner may succeed
in one aspect of pragmatic competence (e.g. illocutionary force), but fail in another (e.g. sociolinguistic
ability). This study closes by advocating the development of multiple-trait assessment in L2 pragmatics
since componential scales can provide raters with different windows from which a learner’s pragmatic
competence can be more accurately judged. In addition, profile scores can provide teachers with
useful diagnostic information to work on learner problems and help to improve their pragmatic
proficiency in a more effective way.
Session 6B

Analysis of factors that influence Korean middle school students’ English achievement by using a latent growth model
Bokyung Cho, Korea Institute for Curriculum and Evaluation
Hyeonae Sim, Korea University
Woojin Jung, Korea University
Kyungyun Kim, Korea University

Learning English is currently a key issue in Korean society. Many Koreans are infatuated in learning it because English can function as a survival kit in the globalized world. For the international trades and relationships, English proficiency is essential demand. In the domestic context, English is extremely important requirement to receive an admission to an advanced educational system, get a job, achieve success in a socio-cultural area, or obtain even political power. To meet all of these needs, the government has invested tremendous money, established an educational infrastructure, and altered the educational policy drastically. For example, they hired native English-speaking teachers for the every school, designated English as an official language in some locations, increased the hours of English class in school, encouraged teachers to teach English in English, etc. To increase effectiveness of teaching and learning English, the experts of English education have introduced the current teaching methods, modified traditional teaching theories for the EFL situation, and designed a new testing system. Compared to previous years, students are exposed to an improved environment of learning English. However, it is necessary to analyze which factors can influence students’ English proficiency development in a long term because acquiring proficiency of a foreign language does not seem to occur immediately after encountering a new learning environment. To investigate the factors influencing Korean students’ English proficiency development, this study analyzes data which Korea Educational Development Institute has collected since 2005. Over 6,000 students have participated in this data collection from their 1st grade of middle school and responded to identical items every year. This accumulated data includes students’ attitudes toward their English classes, motivations for learning English, learning styles and strategies, and English test scores, etc. This study analyzes the first three years’ data because collecting data related to middle school was finished. This research investigates factors that can impact Korean middle school students’ English proficiency development for three years by employing a Latent Growth Model (LGM). The reason why this study uses LGM is that this model can specifically demonstrate the magnitude of observed and latent factors’ to the changes of students’ English achievement scores based on the time differences. The result of analysis reveals that variables related to emotion, learning behavior and learning environment can significantly impact on students’ English achievement.

The problem of continuity between primary and secondary school languages programs: what can classroom-based assessment tell us?
Kathryn Hill

What appears to be a local event also can be understood as a far-flung tendril in a much more elaborate vine. Brandt & Clinton, 2002 p.347. This paper uses an ethnographic study of classroom-based assessment (CBA) practices to investigate the impact of education and languages education policy on the issue of continuity between primary and secondary school languages programs. The context for the study is the trend in Australia, as elsewhere, towards commencing languages at primary school level. While there are a number of additional benefits cited for this policy, there is a clear assumption that it will result in higher proficiency outcomes in the longer term (e.g., Lo Bianco 2001). However, research findings suggest that this is not necessarily the case (e.g., Johnstone, 1999). A number of researchers have identified ‘discontinuity’ as one of the key factors responsible for undermining any proficiency advantages gained from an early start to language learning (e.g., Kabanek-German et al., 1998). Yet this area remains relatively under-researched (Rea-Dickins, 2009). A distinctive feature of the current study is that it uses classroom-based assessment (CBA) practices as a lens for investigating the continuity issue. The research, which took place in Victoria, Australia, involved Indonesian language classrooms at two successive levels of schooling; the last year of primary (Year 6) and the first year of secondary school (Year 7) respectively. Data collection took place in the final term of Year 6 and Terms 1 and 2 of Year 7, comprising a total of ten weeks in each
The study was longitudinal in that the Year 7 cohort included a group of students who had also participated in Year 6. Data comprised classroom observation and field notes, audio recordings (including classroom interactions, teacher and student focus group interviews, and meetings) and documents (including policy documents, reports, course-books, and work samples). Key differences were identified in how learning was assessed as well as what was assessed in the respective classrooms with consequences for continuity between the two levels. The paper will explore how some of these differences can be related to aspects of education and languages education policy, including the introduction of a new assessment and reporting framework. The findings have relevance for languages education policy, planning and practice.

School-based oral language assessment in Hong Kong: Students’ attitudes and anxiety
Manman Gao, The University of Hong Kong

A large number of studies have provided support for the claim that conventional testing, especially high-stake exams, has a negative impact on students’ motivation for learning that militates against preparation for lifelong learning (Harlen & Deakin Crick, 2003). As one important component of motivation, it is argued that anxiety is “quite possibly the affective factor that most pervasively obstructs the learning process” (Arnold & Brown, 1999, p. 8). In line with the shift from Assessment of Learning to Assessment for Learning, an innovative school-based assessment (SBA) has been incorporated into the English subject of Hong Kong Certificate Exam of Education. According to the Hong Kong Examinations and Assessment Authority, the key rationale for SBA was to reduce students’ pressure exerted by public exam and alleviate their anxiety. Adopting mixed methods, this study focused on Hong Kong students’ attitudes and anxiety about this assessment innovation, and examined whether and in what way SBA has effect on students’ affective factors. In the quantitative part, a questionnaire of attitudes and anxiety in SBA was administered to 530 students from seven Hong Kong secondary schools. Results from statistical analysis indicated that Hong Kong students generally held four interrelated attitudes about SBA (i.e., SBA improves the quality of learning; SBA makes students accountable; SBA is irrelevant; and SBA is enjoyable). They experienced a moderate level of anxiety in SBA, and a majority of students agreed that they felt more relaxed in SBA than external oral exams. Students’ attitudes were found to significantly predict their anxiety in SBA. Complementary to the large-scale survey, case study was conducted in two schools so as to penetrate situations in ways that are not susceptible to numerical analysis. Drawing on data from students group interview, teacher interview and observation, the case study explored students’ interpretation and affective experience of SBA in a specific context. On the one hand, the qualitative data provided concrete evidence for the survey results; on the other hand, it examined specifically in what way SBA had effect on students’ affective factors. Among the contextual conditions of SBA, the crucial one was teachers’ mediation, which intimately interacted with students’ attitudes and anxiety in SBA. Peer influence, both peer pressure and peer cooperation, had remarkable effect on a single student’s experience in SBA. Synthesizing findings from quantitative and qualitative data, a tentative model was proposed that illustrated how students’ anxiety about SBA was mediated by their attitudes of SBA and by the interaction with contextual conditions. This research captured a humanistic view about young learners’ affective status in the assessment innovation, which could shed light on the impact of assessment on their motivation for learning. It has theoretical and practical implication for further improvement of SBA and Assessment for Learning.
Session 6C

Assessing health literacy: a new domain for collaboration between language testers and health professionals?

Catherine Elder, University of Melbourne

Health literacy, defined as an individual’s capacity to obtain, process and understand basic health information and services needed to make appropriate health decisions (Simonds 1974) has recently become the focus of increasing attention in medical circles as a result of the growing awareness that sub-optimal health literacy is associated with inferior health outcomes and reduced quality of life. In the interests of identifying the extent of the problem, a number of tools have been developed to measure health literacy in clinical settings. While the internal consistency of such instruments is always reported, and their scores tend to correlate highly with performance on general literacy tests, their validity as measures of the target construct is seldom explored using sources other than correlational evidence. This is disturbing given the discrepant outcomes yielded by the different measures (Barber et al. 2009). The current study, involving collaboration between health professionals and a language tester, set out to assess the validity of the Rapid Estimate of Adult Literacy in Medicine (REALM), widely used as a measure of health literacy, but perceived to be problematic from a linguistic perspective. The REALM describes itself as a “reading recognition” test that measures a patient’s ability to pronounce common medical and lay terms. To test its validity as a measure of understanding of the relevant terms, 155 participants, including both native and non native speakers of English, who had participated in a cross-sectional survey of health literacy in a random sample of the Victoria (Australia) population, were asked to explain the meaning of a subset of 13 words selected from the REALM. Definitions were scored by 3 assessors (including a medical expert, a science graduate and a linguist) guided by a pre-established marking key indicating what was deemed to be a reasonable lay understanding of each word. Findings revealed that although 89% of the participants score in the top REALM band, participants, on average, defined only 7.3 (±2.8) of the targeted terms correctly. Understanding of particular terms varied widely, independently of how accurately they were pronounced. This discrepancy between the REALM score and actual understanding suggests a language-related method effect with the consequence that the REALM tends to overestimate health literacy skills in the majority of participants (and may well underestimate it for particular groups such as non-native speakers of English). These findings point to an expanded role for language testers, working in collaboration with experts from other disciplines, in developing and scrutinizing measures used in domains that are outside their traditional preserve.


Using a threshold principle to decide when someone is (or is not) at a proficiency level: Examples from Hindi and other less commonly taught languages.

Pradyumna Amatya

The Oral Proficiency Interview (OPI) is a highly valid and reliable means of assessing speaking ability in a given language administered in many of the U.S. federal government agencies and academia alike. The OPI has two major components – elicit a ratable speech sample and apply rating according to the Interagency Language Roundtable (ILR) criteria. Once the testers elicit a sample (must follow standard structure, have all required tasks at the level, and use variety of topics to get a ratable sample), they use The Rating Factor Grid (RFG) to rate the sample. RFG groups excerpts from the ILR Speaking Level Descriptions for the six Base Levels (0+, 1, 2, 3, 4, and 5) under each of the six Rating Factors. The six Rating Factors are – Global Tasks and Functions, Lexical Control, Structural Control, Sociolinguistic Competence, Delivery, and Text Types Produced. In order to assign any base level rating, a tester must apply a threshold principle to assess whether an examinee proves during the OPI that he or she can minimally meet all the criteria for the six Rating Factors in a given task and that can sustain all the language tasks required at a level. This paper will generally introduce the criteria for the six Base Levels under each of the six Rating Factors and will specifically discuss
those for the Levels 1, 2, and 3, recognizing the fact that majority of the foreign language learners’ speaking proficiency lies within these levels. The discussion will be supplemented by Hindi and Urdu OPI speech samples (audios, text renditions) to illustrate the distinction among the criteria under each factor.

Fairness and consistency in assigning testing accommodations in large scale testing

Jesse Black-Allen, Ontario Institute for Studies in Education, University of Toronto

In assessing students’ proficiency level in content areas (e.g. science, mathematics, literacy), it must be acknowledged that the language of the test will always have some impact, regardless of whether language itself is the intended content (AERA, APA, & NCME, 1999; Wolf, Kao, Griffin et al., 2008). Testing accommodations for English language learners (ELL) and students with disabilities (SWD) are often intended to reduce construct-irrelevant barriers of low English proficiency or specific impairments, allowing students to demonstrate their full proficiency in the content area being tested (Abedi, 2008). In large scale testing, test administrators must assign accommodations that are appropriate and effective, without compromising the reliability and validity of test results (Sireci, Han, & Wells, 2008). Such decisions are complex because ELL and SWD students have diverse and uneven individual proficiencies in reading, writing, speaking, and listening (Rhodes, Ochoa, and Ortiz, 2005), but these students are often classified within broad categorical frameworks and assigned to language learning levels that preclude a detailed view of their diversity (Solano-Flores, 2008). Researchers have suggested that decisions about assigning accommodations are not driven by research, theory, or even the individual needs of the student, but rather by overarching policies and the interpretation of these policies in local contexts (Fairbairn & Fox, 2009). The current study, conducted in a Canadian context, examines practices of assigning accommodations to ELLs and SWDs on the Ontario Secondary School Literacy Test (OSSLT), a high-stakes test that serves both as a graduation requirement for students and a measure of quality and accountability for schools. Data for approximately 200,000 students in each of four successive years of test administration (2006 to 2009) is analyzed to examine whether accommodations are assigned consistently to the intended groups of ELLs and SWDs across levels of study, subgroups of students, schools, school districts, and years of test administration. Rates of eligible SWDs and ELLs who receive accommodations are examined using crosstabs analysis with chi-square significance testing. To account for the hierarchical and nested structure of the data, multilevel modelling (Goldstein, 1998) is employed to examine relationships at levels of student, school, and school district. Analysis suggests that Ontario schools have widely varying practices of assigning accommodations to ELLs and SWDs and that receiving accommodation is associated with higher success rates on the test for ELLs. Results also suggest that the subgroup of students who speak a first language other than the official English and French, are less likely to be tested for learning disabilities and are less likely to receive SWD accommodations on the OSSLT. This study has implications for policy makers and researchers who seek to promote consistent, fair, and valid accommodation strategies in large scale testing contexts.
Paper Session 7 (plenary)

Time: Friday, 16 April, 2.00–3.30

Time Trials: The relationship between marker speed and quality
Paul Jaquith, Ministry of Education, United Arab Emirates

There is a considerable body of research, both quantitative and qualitative, on raters and the rating process. For example, some studies have investigated rater quality, and the impact of rater background and training. Others have examined the cognitive processes involved in rating essays. However, one area that has until now been practically impossible to investigate on a large scale is the relationship between rating quality and the amount of time spent on marking individual scripts. Intuitively, one would think that very fast markers would be less accurate and that slower, more methodical markers would be more accurate. However, anecdotal evidence suggests that the reality is more complex. The advent of online marking of scripts, which allows rater response times to be accurately tracked, makes it possible to investigate the effects of speed on rater accuracy. This study engages with this issue through an analysis of data gathered from one administration of the high school final English exam in the United Arab Emirates. The data consists of essays produced by nearly 30,000 grade-12 students. All the scripts were scanned into a central database and distributed to 224 trained and accredited markers across the country using an innovative online marking system. After first participating in an online training and accreditation session, the markers were allowed to log in and out at will during the 5-day marking period, during which time they marked their allotted scripts and recorded scores into the database. The system attached a time stamp to each recorded score. All the scripts were double or triple marked, and each marker marked between 100 and 2000 scripts over the 5-day marking period. A FACETS analysis was performed on the score data to assess differences in rater quality as determined by severity and fit. The study examines intra-rater stability and inter-rater diversity in response times across sessions, and the relationship between response time and rating quality. The results of the study have implications for rater training and the design of online rater programs. In addition, they provide important information for ongoing research into rater cognition and rater processes, and present a challenge to research approaches using think aloud protocols.

A multi-round validation study of descriptors-based STEP language proficiency scales

Eunice Jang, University of Toronto
Jim Cummins, University of Toronto
Maryam Wagner, University of Toronto
Saskia Stille, University of Toronto

The wide use of descriptors-based proficiency scales reflects a shift and expansion in the scope and purpose of assessment from a testing culture toward an assessment culture that promotes authentic assessment in a specific language-in-use context (Brindley, 2001; Chalhoub-Deville, 2003; Shepard, 2002). The fields of language testing and assessment have been proactive in researching broader validity and measurement issues related to proficiency scaling in the past two decades. Descriptors-based proficiency scales are often used in making inferences about students’ underlying language abilities based on observable linguistic behaviors in natural environments. This application presents unique challenges to warranting validity claims and psychometric quality, partly because of different philosophical grounds underlying the two cultures (Brindley, 2001; Gipps, 1994; North, 1995; North & Schneider, 1998). The present paper reports an empirical study that examines validity issues arising from the development of a six-step language proficiency scale called Steps to English Proficiency (STEP) used by classroom teachers to assess and track K–12 ESL students’ English development. An initial set of STEP descriptor pool for reading, writing, and speaking was created based on the survey of existing language proficiency scales, then edited further through consultation with teachers. Multiple rounds of field studies provided valuable input from teachers and educators. In 2009, a group of ESL teachers (n=10) rated a total of 477 individual descriptors for three modalities (reading, listening, and oral) across three grade clusters on construct relevance, grade appropriateness and clarity at extended workshops. While the development of the STEP descriptors was led by educators at the Ministry of Education in Ontario, the research team was commissioned to lead a series of validation research that involved educators (N=194) from ten school systems. In the first round of validation research which was largely qualitative, we focused on the following aspects of validity:
(1) the construct representation; (2) consistency in the interpretations of the descriptors; (3) equity for ESL students; and (4) the impact of STEP. We conducted interviews and focus groups with ESL teachers (n=35) who used STEP to assess and track their students over a year. We gathered student achievement data as well as teachers’ STEP assessment results. The study results supported the positive impact of STEP through increase in meaningful collaboration and communications among various constituencies. Clearly, STEP became a common framework of reference that enhanced teachers’ sensitivity and empathy with ESL students while granting legitimacy for teachers’ assessment. At the same time, various challenges to validity repeatedly emerged, including a lack of distinguishability between steps, especially the upper half steps, 4, 5, and 6, partly because of scale shrinkage (North, 1995) and a lack of sufficient observations available for high levels of proficiency. Further, ESL students’ English development was clearly multidimensional, and uneven development of English proficiency across three modalities reflected the idiosyncratic nature of English proficiency development. Such discrepant developmental patterns were observed more frequently with older ESL students. This result discouraged the development of a unidimensional scale for all three modalities and had important implications for making policy and program decisions, such as placement of students in ESL programs and resource allocations. The research team is currently conducting a second-round validity study that examines the scaling issues of the STEP descriptors through multidimensional proficiency scaling, the socio-linguistic sensitivity of the STEP descriptors through multidimensional differential descriptor functioning, and a large-scale survey of teachers’ ratings of the STEP descriptors on clarity, relevance and usefulness. Each of these foci presents psychometric measurement challenges due to the potential violation of various psychometric assumptions and a lack of control over measurement variables. In this paper, we discuss validity issues and challenges unique to developing and validating descriptors-based proficiency scales such as STEP and call for a need to revisit the standard notions of validity and reliability in performance-based assessment contexts.

Theoretical basis and experiments of a computer-assisted marking system on short answer questions

Xiangdong Gu, Chongqing University, P.R.C.
Fanna Meng, Henan Normal University, P.R.C.

The present study aims at how to reduce the influence of human raters and save human resources while keeping scoring reliability in marking Short Answer Questions (SAQs) by way of making use of computer technology. It introduces the Theory of Single Similar Degree in Fuzzy Mathematics—the theoretical basis for the design of a computer-assisted marking system on SAQs and reports the results of three experiments, in which the system is used to mark 220 pieces of SAQs done by undergraduates in an imitated test using CET-4 test paper (a national English test for undergraduates at tertiary level in China) of June, 2005. After each experiment, the marking performance is analyzed and the possible causes leading to the marking errors are explored so that the system is gradually modified and improved. The findings of the three experiments show that “single similar degree” plays a crucial role in measuring the similarity between the students’ responses and reference answers, but it does not function quite well in judging the semantic meaning of students’ responses; thus “key word ratio”, an important factor affecting the semantic meaning of students’ responses is added as a marking index in the system, which remarkably increases the marking correctness ratio of the system. The experiments prove that compared to those for human raters’ marking, the reference answers for computer-assisted system marking should be as specific as possible, which is of great help to improve the marking accuracy of the computer-assisted system. The system in the study only achieves 91.07% agreement with the score assigned by human raters. Nevertheless the study explores some factors affecting the marking accuracy of this computer-assisted marking system. This may offer some implications for future studies. Besides, the present study combines the theory of fuzzy mathematics and linguistics, which confirms the great significance of the interaction and application of different disciplines in language assessment. This paper is part of the research project “A Longitudinal Study of the CET Washback” supported by the National Philosophy and Social Science Foundation of China (07BYY030) and by the National Research Centre for Foreign Language Education (MOE Key Research Institute of Humanities and Social Sciences at Universities), Beijing Foreign Studies University.
Paper Session 8 (plenary)

Time: Friday, 16 April, 3.50–4.50

Task difficulty in L2 listening assessment: an investigation of text characteristics and test takers' perceptions of difficulty

Tineke Brunfaut, Lancaster University
Andrea Revesz, Lancaster University

The effects of task variables on L2 performance have been the subject of much attention in both the second language testing and task-based literature in recent years. So far, however, the research focus has been mainly on writing and speaking skills. Few empirical studies (Brindley & Slatyer, 2002) exist that have examined task variables in relation to receptive skills. The present study intends to help fill this gap by investigating the effects of a group of task input factors on L2 listening difficulty. In particular, we examined how linguistic complexity (i.e., the phonological, morphosyntactic, lexical, and discourse complexity of the text), speech rate (i.e., the speed with which the text is delivered), and explicitness (i.e., the extent to which the ideas are explicitly expressed in the text) affect the difficulty of L2 listening comprehension tasks used for assessment purposes. The participants were 90 ESP students at a UK University. Out of the 90 students, 80 were randomly assigned to four groups. Each group performed the same 18 versions of a listening test task. Participants were presented with the tasks in a split-block design to avoid sequence effects. Immediately after completing a version of the task, students also completed a brief perception questionnaire, which assessed their perception of overall task difficulty and of the linguistic complexity, speed, and explicitness of the text. The remaining 10 students were asked, through a process of stimulated recall, to describe their thought processes during task performance. The texts were analysed in terms of speed and a range of linguistic complexity measures. The participants’ responses were also examined for linguistic complexity as a function of task version. Rasch analysis was used to gauge the comparability of the various versions of the test task, and regression analyses to examine the impact of text characteristics on the relative difficulty of the test versions. The data obtained via the stimulated recalls were subjected to qualitative analysis. Drawing on previous second language testing and acquisition research, the results of the study will be discussed in terms of models of task difficulty (Bachmann & Palmer, 1996; Buck, 2001; Skehan, 1998; Robinson, 2001), with a particular emphasis on the role of text characteristics in L2 listening difficulty. In addition, the implications of the study for test and pedagogic task design will be considered.

Test-taker performance on video versus audio-only L2 listening tests

Elvis Wagner, Temple University

As technology becomes ever more prevalent in society and in education, the use of technology in L2 assessment deserves further exploration. The research (e.g., Baltova, 1994; Gruba, 1997; Progosh, 1996; Shin, 1998; Sueyoshi & Hardison, 2005; Wagner, in press) investigating how the use of video in the assessment of L2 listening proficiency affects test-taker performance has provided conflicting results. While many L2 researchers have claimed that the non-verbal components of videotexts are useful for listening test-takers, and numerous studies have suggested that the use of videotexts might lead to increased test performance (in comparison to audio-only texts), the limited amount of research on this topic has had conflicting results. In addition, there is a dearth of research investigating what L2 listening test-takers actually do while taking a listening test. Finally, an issue that has been debated in the literature but not investigated empirically, is whether test-takers should have the test questions in front of them while the text is being played, or if they should receive the test questions only after the text has been delivered. Consequently, the study had three main goals: (a) to examine how the channel of delivery of the input (audio-only versus video) and test condition (i.e., having access to the test questions while the text is played) affects test-taker performance; (b) to examine the extent to which test-takers make eye contact with the video monitor while taking a video listening test; and (c) to examine if the rate of eye contact with the video monitor is affected by test condition (i.e., if test-takers have access to the test questions while the text is played). 150 English language learners in an intensive English language program at a university in the Northeastern United States participated in the study. There were three conditions: video with access to the test questions while the text was played; audio-only with access to the test questions while the text was played; and video without...
access to the test questions while the text was played. Quantitative analyses showed that the type of input (video versus audio-only) led to varying performance on the test. In addition, having access to the test questions greatly affected the extent to which test-takers made eye contact with the video monitor while the text was played. The implications of the results for testing will be discussed.
Abstracts – Posters

Time: Wednesday, 14 April, 1.30–3.00

Knowing the test takers: Investigating test taker background, test taker characteristics and their test performance
Ying Zheng, Pearson Language Tests
Kirsten Ackermann, Pearson Language Tests

This research investigates the influence of test taker variables on performance in an Academic English proficiency test by exploring the relationships between test taker background, test taker characteristics, and their test performance. The purpose of the research is to identify possible construct-irrelevant factors in language tests that might prevent test takers with different backgrounds from responding to the tests in ways that allow appropriate inferences about their abilities (Zikey, 2007). Meanwhile, ETS (2002) maintained that test fairness requires that construct-irrelevant personal characteristics of test takers have no appreciable effect on test results of their interpretation. Individual test performance and data collected from a test taker questionnaire will be analysed for this study. The researchers will investigate the relationships among test taker backgrounds, including demographic factors, such as gender, country of origin, educational background, and years of studying English, as well as affective factors, such as motivation to learn and perception of the test. These data will be analyzed using structural equation modelling to establish the correlational and casual relationships. Test taker performance at different CEF levels on the four language skills will be compared and correlated with their demographic information. This type of comparison can provide diagnostic information by relating test takers’ specific language competences to their different backgrounds, thus providing useful pedagogical suggestions to language learners of different levels as well as their language educators. To increase test fairness and lower test bias as much as possible, further analysis will be carried out by comparing the sensitivity issues raised by test takers collected from open-ended questions in the questionnaire and those raised by a sensitivity review panel concerning specific test items. Analysis of this nature can also channel test takers’ perspectives on the sensitivity/fairness issues and compare them to the expert judgement. Comparing perspectives from test taker at different proficiency levels can also help categorise the issue, that is, is it a sensitivity/fairness issue or is it a proficiency issue? This research aims to increase our understanding of the individual differences amongst test takers that influence their test performance. As such, this research will be of considerable relevance to academic researchers and foreign language practitioners in the field of language testing and assessment.

Developing an English language proficiency test for pilots
Henry Emery, emery-roberts

Oxford Aviation Academy is the largest independent provider of flight training and aviation crew resourcing in the world. The introduction of the ICAO Language Proficiency Requirements in 2003 created a new demand from the airline industry for aviation English language education and assessment services which OAA was uniquely positioned to serve. Emery-roberts (ER), an aviation English consultancy, began working in partnership with OAA in 2007, supporting OAA with the provision of such services. In August 2008, Oxford Aviation Academy (OAA) and emery-roberts (ER) began an English language proficiency test development project for two reasons:
• There was clearly a market demand for a high quality English language test for pilots
• Both parties doubted the commercial viability of acting as a 3rd party provider of the very few tests available with documented evidence of test quality A test development team was formed comprised of 3 subject matter experts (SMEs) at OAA (flight instructors and examiners) and 3 English Language Experts (ELEs) at ER (teachers and authors of English language training materials for aviation professionals with experience of language testing). Based on ICAO guidance
documentation, needs analysis, experience, and literature from the field of language testing, the team established the following broad test specifications:

The test shall:

• measure English language proficiency for VFR and IFR radiotelephony communications;
• assesses the skills of listening comprehension and speaking only;
• measure proficiency in plain English in an aviation context;
• include, but not overtly measure, standard ICAO phraseology;
• be designed for pilots operating aircraft on international flights;
• measure language proficiency according to the ICAO rating scale and holistic descriptors at levels 3 to 5;
• not be linked to and can be used independently of a particular course of study.

In the development of task specifications and preliminary task trials, and development of administration and assessment procedures, the team employed the services of a UK based language testing expert familiar with the aviation language testing context. At the time of writing, the test is due to be trialled at the OAA Type Rating Training Organisation at London Gatwick airport. The trials will be conducted on a number of professional pilots from a range of first language backgrounds, and video recorded. The resulting performances will be rated by a team of 8 examiners who are both SMEs and ELEs. The data generated from these trials will be subject to a Many Facet Rasch Analysis to investigate how the test tasks are performing and the extent of rater agreement.

Proficiency assessment in less commonly taught languages
Martyn Clark, Center for Applied Language Studies

Efforts to incorporate proficiency-oriented teaching into US classrooms are often hampered by the lack of appropriate assessment instruments, especially in Less Commonly Taught Languages (LCTLs). The Center for Applied Second Language Studies (CASLS) has been developing reading and listening tests in Arabic, Chinese, Hindi, Japanese, Persian, Swahili, and Yoruba as part of a comprehensive Web-based assessment to measure student proficiency in a second language. Test items are intended to be consistent with the ACTFL language proficiency descriptions, which are widely referenced in state foreign language standards in the US. The poster will discuss test design decisions and present results from pilot testing and standard-setting sessions for several languages. Issues of congruence and divergence between the proficiency descriptions and the empirical results will be highlighted, as well as the challenges involved in working with Less Commonly Taught Languages. This project is being funded in part by a Fund for the Improvement of Postsecondary Education (FIPSE) grant.

Promoting assessment literacy for language educators and learners: Challenges, approaches and practices
Victoria Nier, Center for Applied Linguistics
Margaret E. Malone, Center for Applied Linguistics
Chengbin Yin, Center for Applied Linguistics

Assessment literacy typically refers to familiarity with those measurement basics related directly to what occurs in classrooms; language assessment literacy extends this definition to the specific issues of language assessment (Stiggins, 2001; Inbar-Lourie, 2008; Taylor, 2009). While it is widely agreed that classroom teachers need to assess students’ progress (Schaefer, 1993; NEA, 1983), many teachers have a limited understanding of assessment fundamentals (Popham, 2009). This poster focuses on data from three assessment literacy projects and explores the challenges that arise when language testers’ beliefs about measurement basics are contrasted with those of language educators. The first project describes the development of an online tutorial designed to accompany a national directory of foreign language assessments. The poster shows findings from focus groups and reviews by stakeholders (N=50) who informed the development of modules focusing on what language educators need to know in order to select an appropriate test for their purposes. The second project outlines the results of an online introduction to oral proficiency assessment. Data from course participants (N=100) show the extent to which online courses can promote assessment literacy through the results of weekly course requirements. The final project shows the results of a blended
learning course that promoted assessment literacy in teachers of less-commonly-taught-languages. Participants (N=80) in this program gained knowledge of fundamental language assessment concepts both online and in a face-to-face workshop. Stakeholders in all projects included language testers and language educators, and data highlighted the challenge of considering the perspectives and needs of both groups. By examining the data gathered from three approaches to assessment literacy, this poster highlights areas for growth in informing language teachers about what they need to know about testing, and language testers about the real and practical needs of teachers. It also emphasizes the dilemma inherent in language assessment literacy: to deliver responsible content without either patronizing or overwhelming the target audience.

References:

Rating guide development for speaking’s dimensions: Delivery, language use, and topic development
Joan Jamieson, Northern Arizona University, USA
Kornwipa Poonpon, Khon Kaen University, Thailand

This poster will show rating guides that resulted from a test development research project in 2008–2009. The project focused on the speaking section of TOEFL iBT, seen by some as the most noticeable change in the new version of the test, and as an area that would have positive washback effects on the English language teaching community (Butler et al., 2000; Wall & Horak, 2006, 2008). In addition to its traditional use for admissions decisions, the TOEFL iBT score is claimed to be useful in guiding English-language instruction (Chapelle et al., 2008). What can be done to make the criteria and standards for good performance clear to all English language learners and their teachers? How can stakeholders be guided in the steps necessary to improve performance? One possible answer to these questions is to modify the scoring rubric. For the dimensions of Delivery, Language Use, and Topic Development, the project applied both quantitative and qualitative approaches to scale expansion. The quantitative approach included analysis of linguistic features that were reflected in the TOEFL iBT rubric or that had been important in previous research (e.g., Brown et al., 2005). It also included analysis of raters’ scores. For Delivery, Praat (Boersma & Weenink, 2009) was used to measure syllables per second, mean length of run, silent pauses, and pitch (Kang, 2008). For Language Use, transcripts of responses were tagged for lexico-grammatical features (Biber, 1991); the counts of the tagged features were imported to SPSS 17.0 and analyzed. The transcripts were also used to analyze vocabulary richness. Analysis for Topic Development was conducted through an investigation of cohesive devices (Halliday and Hasan, 1976; Biber et al., 1999; Tanskanen, 2004). Additionally, introduction and number of key ideas were analyzed. Data derived from think-aloud protocols were analyzed, based on Miles and Huberman’s (1994) framework, by dimensions and score levels. Previous research, quantitative results of linguistic features present in spoken responses, and qualitative themes of raters’ judgments were synthesized to identify salient features associated with different score levels. These features were used to develop analytic scoring guides based on a series of yes/no questions modeled on the work of Turner and Upshur (1995, 2002). The poster will present the rating guides for each dimension, illustrating the key criteria used to distinguish one level from another. Positioned near each rating guide will be evidence from related literature, quantitative results, and qualitative results supporting the decisions.
Developing authentic online test tasks for oral intercultural professional communication in CEFcult

Jan Van Maele, Group T – Leuven Engineering College
Lut Baten, Leuven Language Institute, K.U. Leuven
Lutgart Dusar, K.U. Leuven/Living Stone Intercultural Entrepreneurship

The authors represent a consortium uniting twelve partners from eight European countries that joined hands to carry out the CEFcult project (2009–11). With the support of the E.U. Lifelong Learning Program, the CEFcult partners set out to develop an online environment for the assessment of oral proficiency for intercultural professional purposes, making optimal use of web 2.0 possibilities. Target groups include employees, those about to (re-)enter the workplace, language teachers, and in-company trainers. The project envisages that learners enter a modular, customizable web-environment and engage in activities that mobilize their oral skills for intercultural professional purposes in selected foreign languages. CEFcult allows for self-, peer-, and expert assessment against the CEFR complemented with other user-relevant scales for language and intercultural competence. The poster will document and discuss the development of the serialized language tasks (‘scenarios’) used in CEFcult. Scenarios will be presented for spoken language use in intercultural professional settings with respectively, a formal, a non-formal, and an informal register. The discussion will focus on the authenticity of the test tasks in view of the domain of generalization (‘authenticity’ in the strict sense) as well as in view of the engagement of the speaker’s competences (‘interactiveness’), two of the qualities that contribute to ‘test usefulness’ (Bachman & Palmer 1996). In the process the authors will also draw on the lessons learned from the earlier WebCEF project (2006–09, www.webcef.eu), which resulted in a web environment for the collaborative evaluation of oral language skills.

Effects of utterance speed, timing control, and repeated exposure on elicited imitation performance in Japanese as a second language

Dan Dewey, Brigham Young University
Hitokazu Matsushita, Brigham Young University
Kayleen Dewey, Brigham Young University

Elicited Imitation (EI) has been used to measure second language (L2) competence (Bley-Vroman & Chaudron, 1994), implicit knowledge (Ellis, 2005; Erlam, 1996), syntactic development (Hamayan et al., 1977; Smith, 1973; van Boxtel et al., 2003, 2005; Roebuck et al., 1999), pronunciation (Burger & Chrétien, 2001), and discourse marker use (Burger & Chrétien, 2001). For EI, test takers typically hear an utterance and then repeat back that utterance aloud. There is considerable variation in terms of the length and types of utterances participants are asked to reproduce, time and tasks placed between hearing and reproduction of utterances, and evaluation of the language students reproduce. Some of this variation is intentional. For example, Jessop and her colleagues (Jessop et al., 2007, 2009) recommend using sentences that prevent learners from using rote repetition (i.e., memory) alone. Some suggestions include placing the structure of interest in the middle of a sentence, using sentences of sufficient length, and carefully controlling for linguistic complexity or difficulty. In this poster, we will present a study addressing methodological variables related to EI that have not, to our knowledge, been intentionally manipulated for research purposes. These include rate of model speech, learner control over utterance timing, and familiarity with utterances. Our data comes from a Japanese EI test presently being developed for use in research on second language acquisition. We are currently collecting data from sixty-five learners of Japanese as a second language, using a computer-delivered sixty-item EI test. We are controlling rate of speech by having learners listen to equal numbers of slow, moderate, and fast utterances (same utterances at three different speeds). All participants are taking two different versions of the EI, one where they have control over when the next sentence will come up (using the mouse on the computer to click a button on the screen and bring up the next sentence) and one where they have no control (sentences come one after the other, with only enough time to repeat the sentences in between). Half of the students are receiving the controlled version first and the other half the non-controlled version. This research is largely in response to pilot testing of our Japanese EI instrument. Learners in that pilot claimed speed and control were two issues that heightened anxiety level, and therefore influenced their performance. In our poster we will present analyses of the possible effects of speed, control, and repeated exposure.
Developing a new theoretical framework for cognitive diagnostic assessment of second language reading

Ah-Young Kim, Teachers College, Columbia University

In L2 reading assessment, there has been a new interest to implement cognitive diagnostic assessment (CDA) due to its capability to provide fine-grained information regarding learners’ strengths and weaknesses (e.g., Buck, Tatsuoka, & Kostin, 1997; Jang, 2005). CDA uses cognitive diagnostic models (CDMs) in identifying the various attributes—“procedures, skills, or knowledge a student must possess in order to successfully complete the target task” (Birenbaum, Kelly, & Tatsuoka, 1993, p.443)—necessary for successfully completing each item on a test. However, previous CDA studies in the field of L2 reading were not based on a theoretical framework of L2 reading ability, making it difficult to interpret the findings in relation to theory. The purpose of the current study is to develop a new theoretical framework for L2 reading ability and examine the various L2 reading attributes necessary for completing an L2 reading test. In addition, the nature of item difficulty will be explored in terms of the number and types of attribute involved in completing the item. For example, a difficult item may require the mastery of more attributes whereas an easy item may require less. The study will begin with the development of a new theoretical framework for L2 reading ability based on the communicative language ability model by Bachman and Palmer (1996). Based on the new framework, five experts in language testing will construct a Q-matrix, which contains information on the attributes required for successfully completing each item on an English as a second language (ESL) reading placement test. The test contains four texts and 30 items in total. Also, reading test scores will be collected from 1000 participants, who are new incoming students to an ESL program. Afterwards, the Q-matrix and reading test data will be analyzed, using the Fusion Model, a type of CDM. Study results will identify the L2 reading attributes involved in the successful completion of the reading test in concern. In detail, the number and the type of attributes will be explained in relation to the new L2 reading ability framework. Also, results will indicate how items interact with each other. For example, some attributes may tend to co-occur due to their interactive nature. Moreover, the findings will shed light on the nature of item difficulty.

Validation of rating scale descriptors for International Teaching Assistants' oral English

So-young Jang, University of Illinois at Urbana-Champaign
Jane Alsberg, University of Illinois at Urbana-Champaign
Amanda Huensch, University of Illinois at Urbana-Champaign
Kyeongsuk Kim, University of Illinois at Urbana-Champaign

The Oral Proficiency Interview for International Teaching Assistants (ITAs), assessing Oral English proficiency, has been developed in order to replace the current SPEAK test at a Midwest state university in U.S. Most typical rating scale descriptors which operationalized language tests currently use are intended to be used for general language proficiency, so they do not provide guidance for academic situations, particularly for testing ITAs. The purpose of this study is to develop empirically grounded rating descriptors, intended to be used for academic purposes on campus. Some efforts were made to develop a more valid rating scale and its descriptors. Rating scale descriptors were initiated by adopting the basic framework of the ACTFL OPI. However, a revision process based on empirical evidence was undertaken to create valid rating scale descriptors for this testing context. The data were collected from both pilot and field-trial tests implemented in fall 2009. Six raters and 40 examinees participated in this study. A variety of measures were employed to validate the rating scale for ITAs. Feedback based on multifaceted Rasch model (North, 2000; Wigglesworth, 1993), analysis of prototype samples for each level, and expert raters’ commentaries on the rating scale using retrospective written reports were reflected in the scale revision process (Furneaux & Rignall, 2000). Based on the results of quantitative and qualitative analyses, rating descriptors were validated, and findings contributed to enhancing the accuracy and appropriateness of rating scale descriptors. Educational implications for both individual examinees and educational organizations will be discussed in terms of washback impacts and rater training.
Test adaptiveness in development of oral proficiency test for International Teaching Assistants
Kyeongsuk Kim, University of Illinois at Urbana-Champaign
Jane Alsberg, University of Illinois at Urbana-Champaign
Amanda Huensch, University of Illinois at Urbana-Champaign
So-young Jang, University of Illinois at Urbana-Champaign

SPEAK (Speaking proficiency English Assessment Kit), known as a computer-mediated test, has been employing for assessing oral proficiency of international TA candidates. However, in spite of high practicality in its implementation, SPEAK has been criticized in terms of face validity, authenticity and technical problems. The purpose of the present study was to investigate the effects of an adaptive test when assessing oral proficiency by administering an interview. The test development process was done in two major stages: development of the test specification, and a validity study of the test (Fulcher & Davidson, 2007). The test specification, a blue print of the test, was developed by reflecting on the nature of the test adaptiveness by acquiring high interactivity and authenticity (Davidson & Lynch, 2002). One of the features that distinguishes a face-to-face interview from a computer-mediated test in assessing oral proficiency is its adaptiveness in questioning. The interview test is structured with a fixed format of questioning phases and both general and field-specific questions are given based on the interviewees’ field of study and their proficiency level in order to obtain high degree of authenticity. The test was validated based on multiple sources of findings from the pilot test and field trial test, which were conducted fall 2009. The test was evaluated by multiple stakeholders; examinees and raters. Surveys and focus group interviews were employed to identify the interviewees’ perception in terms of test adaptiveness and it was compared with SPEAK. In addition, interviewers (or raters)’ perspectives were considered during the test revision process. Results suggest that the interview test will function better as a screening test for international TAs in an academic setting. Findings were meaningful in terms of giving feedback on the revision of test specification. Also, findings provided some implications for the screening process of interviewers (or raters) and the development of effective rater training models for the interview test.

Language policy and language proficiency assessment in English and Putonghua: Is what is proposed, possible?
Jane Lockwood, Hong Kong Institute of Education
Michelle Raquel, Hong Kong Institute of Education

Hong Kong prides itself on being trilingual and biliterate, but how proficient is its population across the three languages of Cantonese, Putonghua and English? The Hong Kong Institute of Education, the largest provider of teacher education in Hong Kong has, until recently, conducted most of its undergraduate degrees in Cantonese. Whilst there is modest language enhancement support in English and Putonghua provided by HKIEd, there is little impact on language proficiency gain in these two important languages given the language of instruction. This is about to change. A new language policy, still under consultation and yet to approved, will be piloted next year. This policy will promote English medium instruction and will impose, for the first time, mandatory exit benchmark levels for both English and Putonghua. For English the mandatory exit test is the International English Language Testing System (IELTS). This poster will first exhibit the salient features of the proposed new language policy and consider its implications for both languages. In particular, this poster will provide information about the newly developed Tertiary English Language Test (TELT) and specifically how this test is being developed by the Centre for Languages in Education (CLE) for the multiple purposes, within the tertiary context, for placement, diagnosis and proficiency progress reporting. A big challenge for CLE will be to ‘track’ language proficiency gain over the length of the undergraduate degree and make regular reports on proficiency gain (or the lack thereof) of the students. Visitors to this poster session will be able to view the TELT placement test (including data analysis on the items), the speaking and writing proficiency scales and descriptors and also talk to the authors of the test. We will be interested in discussing the issue of calibrating the multiple-choice placement test with the oral and written test. This poster will also report on the preliminary work being done on the Putonghua test to mirror the purposes, frameworks and approaches outlined in the TELT.
Automated rating of an ESL Writing placement test: can we trust machines?
Chih-Kai (Cary) Lin, University of Illinois, Urbana-Campaign

The effects of the arrangement of questions in can-do self checklists on students’ responses and their ease of use
Tomoko Fujita, Tokai University

The focus of this research is on how the arrangement of questions affects students’ responses and their ease of use regarding a listening can-do self-checklist form. Because the new can-do based curriculum in a Japanese university English language program requires 11,000 students to self-evaluate using a can-do self checklist, creating valid checklists is essential for the success of the program. Previous studies indicated that early questions become the context for later questions, and placing related questions together increases their contextual influence because the set of questions together provides a context for one another (Knowles et al., 1996, Goldberg, 1992). In order to examine the focus of this research, three different forms with 30 questions each for the can-do self checklists were created: (a) form R has questions arranged randomly, (b) form L has them arranged according to the three difficulty levels, and (c) form C has them arranged by the ten categories of listening, for example, listening speed or context. These checklists were written in Japanese and form the basis for this original can-do framework. Each of the ten categories has 3 different difficulty levels, adapted to the A2, B1, and B2 levels of the Common European Framework. At the end of the listening course, 630 students randomly chosen from all three proficiency levels evaluated their listening skills using one of the 3 forms. All items, including the form’s ease of use, were evaluated on a 4-point scale. Firstly, the means of each item on the 3 forms were calculated and found to be very similar to each other. Secondly, one-way ANOVA among the 3 forms was conducted. Only 5 questions showed significant differences. Thirdly, the reliability of the 3 forms is similar as well (.92 << .93). Finally, form C has the highest ease of use (2.09) and form R the lowest (1.95). The same analyses were conducted for the three different student levels (advanced, intermediate, and basic), but the results did not show significant differences by form. The results indicate that each question was answered individually regardless of its arrangement on the form. This supports the assumption of local independence of IRT models. The lack of significant differences among the 3 forms in terms of proficiency level could refute the common teacher’s claim that students in basic level classes do not answer questions seriously. Moreover, the highest ease of use toward form C might suggest that employing form C increases students’ motivation in completing the checklists.

The European Survey on Language Competences: an update on progress
Neil Jones, University of Cambridge ESOL Examinations

A European Indicator of Language Competences, to report against the CEFR, was first mooted in 2002 as one measure of progress towards the Lisbon Strategy goals. SurveyLang, a consortium led by Cambridge ESOL, is developing the ESLC, scheduled for 2011. It will test English, French, German, Italian and Spanish. This update on progress reviews the language test development, other major work areas, and s have been defined, and approaches to cross-language alignment and standard setting against the CEFR

Graphical protocol analysis: A method to measure a writer’s language competency
Putri Afzan Maria Zulkifli, University of Sussex
Peter Cheng, University of Sussex

Present language tests such as IELTS and TOEFL measure a person’s English ability across all four language skills – listening, reading, writing and speaking – by means of exam-oriented approaches. Recent advances in Cognitive Science now means that it may be feasible to assess language competence by examining lower level cognitive processes that are common to and underpin all four language skills. Graphical Protocol Analysis (GPA) is a new technique that records and analyses information about the cognitive process of writing using data captured on a graphics tablet. By interpreting patterns of pauses between writing actions it can provide an insight into what happens at the cognitive level of the mind. It has been well established in Cognitive Science that durations of pauses between actions reflects the amount of mental processing that is needed to prepare the actions. During writing pauses can be captured at various levels; within a letter, between letters,
Abstracts – Posters

between words, between phrases or sentences. We are investigating whether the analysis of pauses during the production of freehand writing within copy-writing tasks may be a viable method to measure the degree of language competency of writers. A pilot study has been conducted to explore suitable designs of copy-writing tasks. Fifteen types of task were examined, including: sentences on technical topics; sentences with jumbled words; sentences containing words with jumbled letters; letters of the alphabet in correct and reserve orders; people’s own names. Five participants with different levels of English ranging from an expert (native speaker) to a novice (newly arrived overseas student) took part in the study. The trials within the tasks were repeated either with an initial reading period of 30s or with no initial reading, in which the participant starts writing immediately after stimulus is shown. For each of the trials various statistical measures of the durations of pauses were computed, including the mean, median, quartile ratio, and others. It was found that in the no initial reading condition that the rank order of the English competency of the participants was correctly and reliably persevered by some of the measures across the majority of tasks. Thus, it appears that there may be a temporal signal in the cognitive process of writing that might be used as a novel method to measure language competence.

Creating a multidimensional scale of language proficiency
Michelle Raquel, Hong Kong Institute of Education
Jane Lockwood, Hong Kong Institute of Education

The Tertiary English Language Test (TELT) was developed at the Hong Kong Institute of Education (HKIEd) for placement and screening purposes. It was developed with an intention of having a placement test that is linked to the curriculum and so addresses the particular needs of the English enhancement programmes of the Institute. The test has three components: a multiple-choice test that measures students’ receptive skills (listening, reading, vocabulary and grammar), an oral test, and a written test. The test aims to divide students into three proficiency levels based on the results of these three components of the TELT. Rasch analysis was used to analyse each component of the test. The results revealed that while each test component is able to divide students into three groups, the researcher was faced with two challenges: (1) how to correlate the results of the three components to create a multidimensional scale of language proficiency; and (2) how to determine cut-off points that reflect the curriculum. This presentation will discuss the process involved in creating a multidimensional scale of language proficiency, and the process of identifying cut-off points. It will outline stages of student/item calibration, stages of test component correlation, and evidence gathered to justify cut-off points. The presentation will also report on stakeholders’ response to the new tool and immediate challenges for further development such as redesign to serve as a diagnostic and exit test.

Language and Integration
Martin Nuttall, Association of Language Testers in Europe
Nick Beresford-Knox, Association of Language Testers in Europe

The focus on language assessment for integration has grown in recent years with a number of European countries having introduced or formalised linguistic requirements for the purposes of migration, residency and citizenship. Within this context, on behalf of the Association of Language Testers in Europe (ALTE) Special Interest Group for Language Assessment for Migration and Integration (LAMI), we would like to propose a poster focusing on language assessment for integration in a number of our members’ countries. ALTE was founded in 1990 and has grown to 31 members, representing 26 European languages. The LAMI group within ALTE, which investigates issues in language and migration, was established in 2002. The group has been active in organising and taking part in symposia over the last 7 years. The latest event was in September 2009 when the group organised a one-day conference on ‘Language and Integration’ in Berlin within the framework of a three-day event, ‘Sprachen ohne Grenzen’, being organised by the Goethe-Institut, Germany. The proposed poster will combine recent work by the LAMI group in preparation for the conference mentioned above. It will illustrate the background to the policy making in individual countries as well as information regarding the teaching, learning and language testing processes. Key discussion points of the poster will include: the policies, course systems, course providers, course participants, testing systems, test providers, test participants and consequences of testing in the countries represented in the sample. We propose to focus on European countries including Bulgaria, Czech Republic,
Denmark, Italy, Germany, Luxemburg, Norway, The Netherlands and United Kingdom. The poster will be an opportunity to compare and contrast the key issues within the countries involved.

Evaluating the new College English Test (Band 4) with a comparative approach
Min Liu, The Hong Kong Polytechnic University
David D Qian, The Hong Kong Polytechnic University

The College English Test Band 4 (CET-4), as a high-stakes national test, has been implemented in China for about two decades and has become the largest EFL test in the world (Jin, 2008). In 2006 the new CET-4, following a substantial reform, debuted nationwide. It is intended to serve as an accurate measure of students' overall English proficiency, and maximize its beneficial washback on college English teaching and learning. Given the huge impact of the test on the university population in China and the scant empirical research on the revised version, the present study was conducted to investigate the following aspects: 1. What has caused the reform of the old CET-4? 2. What new developments have been introduced into the new CET-4? 3. Statistically speaking, to what extent does the new CET-4 serve as an improved measure of students' overall English proficiency? The study involved 192 test takers, who sat for both the old and the new versions of the CET-4, and 232 questionnaire respondents. Both qualitative and quantitative methods were employed. A comprehensive review was conducted on the history of CET-4 and its reforms in the past two decades to synthesize the previous validation efforts and highlight the newly introduced components. An analysis of the test contents and item types in the old and new versions of the CET-4 were made to compare the similarities and differences between various aspects of the two versions of the test. Statistical procedures for analyzing the two sets of test scores included descriptive analysis, item-total correlations within each test, EFA, and intercorrelations and shared variance between the two tests. The questionnaire explored students' motivations, general perceptions of the CET-4, respondents' perceptions of the structure, content, and test difficulty of different subtests, and their test preparation strategies. A needs analysis of university students' English language use in their daily, academic and prospective workplace contexts was also conducted using factor analysis in order to determine whether the new version of the CET-4 is better aligned with the language needs of the university students. Since the examination of content relevance and coverage is a necessary part of the validation process (Bachman, 1990), content analysis was conducted on the basis of task characteristics proposed by Bachman and Palmer (1999). The input characteristics of the listening and reading passages were mainly analyzed from the aspects of length, topics, genres and readability. The analysis of expected response was carried out in accordance with the listening and reading skills listed in the CET-4 test syllabus. The paper will discuss the preliminary findings as related to the validity and effectiveness of the new CET-4.

Investigating accuracy and complexity across levels: towards a valid scale for the Language Proficiency Examination for Foreign Language Teachers (EPPLE) in Brazil
Vanessa Borges-Almeida, UNESP – State University of Sao Paulo
Douglas Consolo, UNESP – State University of Sao Paulo

This poster reports on an investigation into the validity of the grammar descriptors for the Test of Oral Proficiency in English (TEPOLI; Consolo, 2004; Consolo & Teixeira da Silva, 2007), an oral test specifically designed for teachers of English as a foreign language in Brazil. The research study, of a hybrid nature, is socially justified by the results of several studies in Applied Linguistics which attest the low proficiency level (particularly in oral skills) attained by many students graduating from teacher education courses and by in-service teachers working in regular schools. Those teachers' level of English proficiency has been accounted as one of the reasons for unsuccessful teaching and learning in such settings. At present a complete examination named EPPLE (Language Proficiency Examination for Foreign Language Teachers; Consolo et al, 2009), covering the four skills in an integrated fashion, has been developed and needs empirical information in order to build on valid scale descriptors and levels. Such is the contribution this present low-scale study has to offer, aligned with other studies investigating the validity of the other descriptors in the same scale. This poster presents part of a PhD dissertation conducted with graduating students of a foreign language (English) teacher education course in a state university in Brazil. The research question that guides this study is ‘how is grammar characterized across the TEPOLI’s proficiency levels in the participants’ performance
in terms of (a) accuracy and (b) complexity in the oral test and in a seminar? Data were collected by means of recordings of paired administrations of the TEPOLI and of recordings of seminars given by students during their last term at university. Accuracy was analyzed in terms of an errors-per-unit (AS-unit; Foster et al., 2000) index, an error-free unit index, and a self-correction index. For complexity, two indexes were taken into consideration: a clauses-per-unit index and a tokens-per-unit index. A closer look was given to the construction and communicative use of relative clauses and if clauses across levels. Results point to quantitative and qualitative differences between the proficiency levels for both accuracy and complexity. They also indicate that participants’ performance at the TEPOLI has predictive power over their performance at the seminar. These results can inform important future decisions regarding the calibration of descriptors in the proficiency scale for the EPPEL examination.

Investigating CEFR levels of Japanese University students with DIALANG: comparison of results with outcomes from the DIALANG Pilot Study
Chisato Saida, Ibaraki University, Japan
Abstracts – Works in Progress

Time: Thursday, 15 April, 2.25–3.40

Lecture Room 1
Chaired by: Talia Isaacs

‘But I had nothing to say’-building a topic validity argument for the IELTS Speaking Test
Nahal Khabbazbashi, University of Oxford

Variability in spoken language performance is an area of considerable interest within the context of second language assessment. The aim of this research project is to examine the role and influence of test task topics on variability in oral test performance in the IELTS speaking test. It is hypothesized that there is an interaction between an individual’s background knowledge of topic and the test task topics with the potential to influence performance (both in terms of discourse and test scores). The results of this project can help to increase our understanding of whether a factor such as topical knowledge or topic schemata (an arguably construct-irrelevant factor) can help or hinder crossing a given proficiency threshold within the context of the IELTS speaking test. Within Weir’s (2005) socio-cognitive framework for test validation, the study will adopt a mixed methods approach to gathering empirical evidence for building a comprehensive topic validity argument for the IELTS speaking test. A parallel forms reliability method (using retired IELTS material) will allow samples of speech to be gathered from the same participants on two sets of IELTS tests, each with different topics. Questionnaires and short interviews will elicit a) the participant’s background knowledge of each topic, b) the perceived difficulty of the topics and c) the cognitive processing associated with each speaking test and task topics. The speaking tests will then be rated by a minimum of two IELTS examiners in order to investigate whether performance in speech varied significantly enough to be translated into differences in speaking scores. A sample of performances will also be explored from a discourse analytic perspective which can reveal nuances in performance which might have gone unnoticed during overall ratings. Additionally, the raters are asked to provide a retrospective account of their rating process so as to determine whether the choice of topics exerted any influence on their scoring decisions. By bringing forth these different types of evidence, a solid basis is provided for constructing a topic validity argument for the IELTS speaking test.


Rater feedback on a test of oral proficiency
Raili Hildén

The background of this study is a 3-year research project funded by the University of Helsinki, Finland, with a focus on the assessment of oral proficiency in foreign languages. The project aim is to build a validation argument regarding relevance, utility, and sufficiency for the illustrative scales of oral proficiency included in the new National Core Curricula (NCC). This scale with subscales for overall task completion, fluency, pronunciation, range and accuracy is empirically aligned to relevant scales of the CEFR. Three sets of test tasks have been designed, targeted to CEFR levels A2, B1 and B2 as an average. The test addresses three linking points of language study: at the end of year 6 (about 13 year-old pupils), at the end of compulsory basic education (at the age of 16), and at the end of upper secondary education (at the age of 18). The data (about 160 speaking samples) are collected from schools and university institutions and rated by 5–10 experienced judges. The task content, i.e. themes, functions and situations, are derived from the NCC. The task prompts being in Finnish, the test is meant to be the same in all tested languages. Research problems address issues of inter-rater and intra-rater reliability of expert members of the Swedish, English, German and French juries, each 5–10 in number. Inter-rater reliability is scrutinized in relation to overall and task specific ratings on the subscales. Rater comments and interaction patterns emerging at the sessions are also taken into consideration.
into account. The findings are compared across members of each jury and also across languages. Intra-rater reliability is deduced from the correspondence between thematic commentaries of a rater and his/her ratings attached to congruent categories. The data consist of level ratings assigned by the raters, feedback sheets and video recordings from rating sessions. A multi-method approach is adopted implying statistical analysis of numeric matrices and content analysis for qualitative data. Some preliminary results from the German jury already support appropriateness of the selected techniques for the purpose. For instance, the themes focused by the judges derived from the results of the qualitative analysis tend to match the statistical findings regarding inter-rater reliability and a range of task features. The analyses will be replicated for the three other languages. Envisaged results have impact on design of an oral course to be introduced at upper secondary level in 2010.

Automatic fluency assessment using thin slices of spontaneous speech
Suma Bhat, University of Illinois, Urbana-Champaign, USA
Mark Hasegawa-Johnson, University of Illinois, Urbana-Champaign, USA
Richard Sproat, Oregon Health and Science University, Beaverton, USA
Fred Davidson, University of Illinois, Urbana-Champaign, USA

Whether human-rated or automated, language proficiency assessment is typically done on spoken language segments several minutes in length. The segments are task-specific utterances and the proficiency assessment is done based on criteria designed to reflect components of language proficiency such as fluency, lexical accuracy and topic coherence. In this context, two natural questions arise, one, can we assess oral fluency by using portions of utterances, and two, what measures obtained from a short slice of the utterance can be used towards fluency assessment. This study seeks to answer the two questions with focus on the assessment of second language fluency automatically. Answering the first question, we show that fluency scores based on a random twenty-second snippet of the utterance approximate human judgments based on the entire utterance reasonably well. Addressing the second question we show that a set of measures obtained from direct signal-level measurements quantify the automatic fluency score. By way of this work our goal is to understand the contribution of automated fluency assessment to defining a level in second language testing while conforming to a particular framework. The novelty of our work lies in the automatic assessment of language fluency based on a “thin-slice” of the original utterance using signal level measurements. In choosing the length of the snippet, we draw insights from social psychology where studies have shown that behavioral judgments based on “thin-slices” of the observation stream are accurate in a wide range of settings. The measures quantifying oral fluency are the temporal measures of speech drawn from existing literature on automated methods of language assessment. The quantifiers that we study include, among others, phonation time ratio, syllable rate and mean length of silences obtained using standard signal processing techniques. Experiments were carried out on a rated dataset of spontaneous speech produced by 28 second language learners of English.

Investigation of a CAEL writing band score as diagnostic
Christine Doe, Queen’s University, Ontario

This work in progress investigates the test construct of a large-scale test used as a diagnostic assessment. By drawing on raters’ knowledge of marking and teaching, this study specifically will explore the distinguishing diagnostic features about test takers’ writing at a particular level (i.e. band score) and the feedback that can be offered to those test takers. Diagnostic assessments identify test taker strengths and weaknesses resulting in feedback that informs student placement, programming and learning outcomes (Alderson, 2007; Fox, 2009; Huhta, 2008; Read, 2008). The construct of a diagnostic test has been examined by mapping test items onto a theoretical model of language proficiency (Alderson, 2005; Jang, 2008). While others have explored the construct by identifying skills needed for academic success (Fox, 2009; Llosa, Beck & Zhao, 2009). Previous research has also looked at the characteristic features of essays across proficiency levels through discourse analysis (Knoch, 2007). However, missing from the diagnostic assessment literature is an in-depth look at what features of writing are characteristic of a level from the rater’s perspective, by drawing on their knowledge of scoring essays and teaching. Context. This study looks at the Canadian Academic English Language (CAEL) Assessment, a language proficiency test used for university admissions purposes. The test consists of nine band scores from 10 to 90. Focus will be placed on scores of 50 since a score of 60 may be high-stakes for test takers wanting to enter a Canadian university (Fox,
2004). The borderline essays for the 50 band score will be extremely meaningful as they are “on the border between the two categories, possessing not quite enough knowledge or skill (in the [rater’s] judgment)” (Cizek & Bunch, 2007, p. 112). Drawing on raters’ expert judgment this study will explore the diagnostic test construct for the 50 band score by looking at essays that are borderline, but also characteristic of the level. Methods. Twelve previously scored CAEL essays will be collected for the 50 band score, with borderline essays represented by a plus or minus sign. These 12 essays will form three groups (40+/50-) (50) (50+/60-), with four essays in each group. Two experienced raters will complete verbal protocols (Lazaraton & Taylor, 2007) while each reviewing the 12 essays. In the verbal protocols, they will talk aloud about the features of the essay that place it in one of the three groups. In addition, the raters will draw on their knowledge as teachers and consider the advice they would give the test taker on how to improve. A standard qualitative theme analysis will be used to analyze the verbal report data (Patton, 2002). The reporting of results will focus on the key criterion that characterizes the 50 band score as basis for the diagnostic construct. Specific attention will be given to the information obtained from the borderline essays in order to form an in-depth understanding of the distinguishing features of the 50 band score from the 40 and 60 band scores, of which is also a theme proposed by the LTRC 2010 planning committee. To conclude, an initial framework of diagnostic feedback will be created from the raters’ comments. Presentation. In the presentation of the work in progress, I will discuss the current literature on diagnostic assessment, the methods used, and the preliminary findings. Since the data collection will conclude two months prior to LTRC 2010, I will welcome comments and suggestions from the audience regarding the study and initial findings. Furthermore, I look forward to engaging with others in a discussion about the test construct for a diagnostic assessment of writing.

Online training of raters of university-level writing scripts
Janet von Randow, *University of Auckland*
Janna Fox, *Carleton University, Ottawa*

The Diagnostic English Language Needs Assessment (DELNA) programme for first-year students at the University of Auckland includes a writing task that identifies students who need support in the area of academic literacy. Scripts are double-rated on an analytic scale and diagnostic feedback is available for the advising of students on appropriate forms of support. Although it is not a high-stakes assessment situation, the quality of the feedback depends on good rater training. All the raters attend face-to-face training at the beginning of each semester, but there is a need for periodic refresher training as batches of scripts require rating on an irregular basis through the semester. To meet this need, an online rater training system was introduced in 2005 as a convenient means for raters to re-train in situations where it was impractical to arrange a face-to-face session on campus. Two studies (Elder, Barkhuizen, Knoch & von Randow, 2005, 2007) have investigated how raters have responded to the online programme and both pointed to the need for further research and streamlining of the system. The present study, which is a collaborative project between the University of Auckland and Carleton University in Canada, is designed to explore the extent to which the online system can work effectively with novice raters. The first stage of the project will involve beginning raters (unfamiliar with DELNA) at each location, along with experienced DELNA raters in Auckland. After completing the online training, all the participants will provide think-aloud protocols as they rate six scripts, followed by a short interview. This will provide a basis for comparing what the novice and experienced raters attend to in each script and how they arrive at their scores. It is expected that the analysis will show that some revision and elaboration of the online training programme is required to meet the needs of novices using the system for initial training. Then, in the second stage of the project, we will explore from various perspectives whether in fact the remodelled online system can be used exclusively to train novice raters, without being complemented by conventional face-to-face sessions. In this session we will discuss the rationale for the project and present a progress report on the first stage.

Linking words used on various levels of proficiency in argumentative writing
Henna Tossavainen, *University of Jyväskylä*

Linking words used on various levels of proficiency in argumentative writing This work in progress aims to describe the use of linking words in argumentative writing on various levels of proficiency in the Finnish National Certificates (NC) English language test. The work is part of my study that analyzes the use of metadiscourse markers when expressing opinion on different levels of proficiency.
These six levels are linked to the Common European Framework of Reference (A1-C2) and are covered through three different tests. I will focus specifically on the use of linking words in relation to language proficiency development. The aim is to develop assessment criteria for writing on the various levels of proficiency. The data are collected from the NC Corpus. There are 80 scripts from the advanced, 81 scripts from the intermediate and 59 from the basic level, each from a different candidate. The corresponding test tasks are also used in the analysis to discover such differences and similarities in task content that may have effect on test performance. The rationale for the study is that argumentative writing has generally not been assumed to be much in evidence at basic level language proficiency and research on argumentative writing has focused on academic and business writing. The language development perspective has not commonly been used in studying argumentative writing either. The metadiscourse markers used in the study were collected from various studies on argumentative writing and focus on such markers that bear relevance to writing in a foreign language.

Lecture Room 4
Chaired by: Elvis Wagner

An in-process meta-analytic study of language criteria across frameworks – opportunities and challenges
Eric Surface, SWA Consulting
Kathryn Nelson, SWA Consulting

With a global economy, organizations need to identify an increasing number of individuals with foreign language capability (Jackson & Malone, 2009). These organizations rely on language assessments to select, place, classify, certify and evaluate individuals for various purposes. Language capability measures vary in terms of quality, construct definition and empirical support. They can include proficiency tests, specific performance tests, and achievement tests among others. Choosing the right evaluative criteria can be difficult, and often depends on organizational goals and practical constraints (Ross, 1998). However, using only one criterion can lead to errors as it may not adequately measure all areas of proficiency. Therefore, high-quality practice suggests using multiple measures of language capability. Although some studies have looked at the relationship between one or two types of criteria, the relationships among many language capability measures have yet to be determined systematically. The knowledge of relationships between language criteria can help to inform organizational practices and to direct future primary research. It can even be used to compare criteria across different language testing domains and frameworks, providing macro-level insights. One way to examine the relationship between language criteria variables is through the use meta-analysis to integrate previous empirical research related to language criteria. This Work in Progress outlines the methodology used and the issues found while conducting a comprehensive meta-analysis of foreign language criteria as well as initial findings and suggestions for future research. Meta-analysis is a statistical summary of existing research studies; however, unlike primary research studies, it is not subject to the biases and limitations (e.g., sample size, population characteristics, etc.). Instead, meta-analyses combine the results of primary studies to examine and statistically determine the average relationships across all included studies, thus providing an unbiased understanding of the strength between variables (Hunter & Schmidt, 2004; Lyons, 1997). A meta-analysis also can correct the measures for unreliability, providing a better estimate of the true relationship between variables. Although this technique provides the potential for more robust answers, the quality of the information presented in the primary studies limits the quality of the meta-analysis. The accuracy of meta-analytic and narrative reviews depends on the complete reporting of information in primary studies. When primary studies fail to include enough information meta-analytic investigation, the results of the meta-analysis may be biased and incomplete, as they will not be based on the entire population of studies examining the relationships in question (Hunter & Schmidt, 2004). In addition to discussing our methodology and any preliminary findings, this Work in Progress examines the amount of missing information (e.g., correlations, sample sizes, reliabilities, etc.) in articles from language journals (e.g., Modern Language Journal, Foreign Language Annals, etc…) and other sources, and demonstrates that even after contacting authors and organizations to obtain the information, much important information was not presented and is unrecoverable. At this point in the study, following
standard meta-analysis procedure, 6377 abstracts have been reviewed. Of these, 1077 documents have been selected as meeting our initial screening criteria—abstract discusses language criteria—and reviewed to determine whether or not they fit our specific inclusion criteria—empirical studies with multiple language criterion measures with sufficient statistics. Of the 1077 reviewed, approximately 195 documents met our minimal data standards (correlations between criteria), but only 117 of the 195 published and unpublished articles, technical reports, and dissertations included reliability estimates for the measured variables. An additional 60 studies could not be included because they did not present either correlations between measured variables or the sample size (N) used in the study. Meta-analysis could be an important tool for improving our understanding of the relationships between language criteria, the effectiveness of language measurement, and the interrelationship of language testing domains and frameworks. However, our initial work suggests that many studies are not reporting sufficient statistics to be included in meta-analytic evaluations which negatively impacts the field. Suggestions for future statistical reporting standards in the foreign language literature as well as other options to address this issue will be discussed.

‘Can Do’ Project
Sarah Brewer, University of Reading
John Slaght, University of Reading

The Can Do Project has been designed to build up a database of evidence to support a re-evaluation of existing assessment criteria and descriptors used at entry points to Higher Education, particularly at Master’s level. It is headed by a working party of EAP lecturers / researchers who have established the research background and methodology and have recently been joined by a group of volunteer researchers / contributors from other universities or HE institutions. The ultimate aim of the research is to review and develop existing assessment measures and to create new measures of assessment to match current HE developments and needs. At present international students are assessed by meeting criteria which have, in many cases, been constructed intuitively and developed organically over a period of years. The authoritativeness of such criteria is now increasingly called into question as universities expand their international intake and introduce a much greater range of Master’s programmes. This has led to uncertainty as to how to decide when a student has reached the appropriate level to begin study and equally, uncertainty as to how far current assessment criteria / descriptors match the requirements of changing academic domains. The proliferation of pre-sessional programmes designed to prepare international students for university and their employment of a range of exit assessment measures may have added to the complexity of this situation since there is a lack of a common framework or standardisation. The Can Do project is based on a Literature Review providing a context and background to the research and is a qualitative study involving semi-structured interviews in a wide range of academic departments. The interviews are designed to elicit information on what Non Native Speaker students, who have succeeded in reaching the required language entry level, ‘can’ or ‘cannot’ successfully accomplish on their courses. The results of the interviews are summarised and formatted as a matrix. Currently more interviews are being carried out by volunteers in other universities allowing for the data base to be expanded and for more detailed analysis to be undertaken. Initially this analysis will focus on where the data matches current assessment criteria / descriptors and where there are gaps; but with further dissemination of information about the project it is hoped that the data could become available for other analytical and assessment purposes.

DIF investigation on gender, academic background, and university type differences in a large-scale high-stakes test
Xiaomei Song, Queen’s University, Ontario
Liying Cheng, Queen’s University, Ontario
Lianzhen He, Zhejiang University, P.R.C.

Concerns about test fairness among all test stakeholders are paramount in the milieu of high-stakes decision-making involving university admission, achievement, aptitude, certification, and licensure tests. These tests serve as gatekeepers to include or exclude individuals into academic communities and professional associations (Shohamy, 2001). One of the key processes in obtaining evidence of test fairness is to detect potentially biased test items that are in favour of or against test takers from
certain social groups (i.e., ethnicity, gender, linguistic status, or socio-economic status) in testing scores (Cole & Zieky, 2001). Over the past decades, predominant research of this kind has been conducted with tests developed and validated in North America (i.e., Angoff, 1989; Mahoney, 2008; Snetslier & Qualits, 2008). Studies of this nature are very limited within the Chinese high-stakes testing context which involves a huge number of test takers and other test stakeholders. This study examined statistically whether any systematic group differences in terms of gender, major of study (Arts/Sciences), and university type (key/non-key universities) existed in the 2009 Chinese Graduate School Entrance English Examination (GSEEE). The purposes of the GSEEE are to measure English proficiency of test takers and to provide information for educational institutions in selecting candidates into their Master’s programs. Only those students whose overall scores on the GSEEE are above the admission cut-score set by the country’s Ministry of Education may have a chance to attend university Master’s programs (He, 2009). The 2009 GSEEE consisted of three sections: Use of English (20 multiple-choice items), Reading Comprehension (20 multiple-choice items, 5 fill-in-the-blank items, and 5 translation items), and Writing (a practical writing and an essay writing task). The three grouping variables (gender, major of study, and university type), which represented the GSEEE’s major test taker group differences, have been documented in previous research having potential interaction with language proficiency test performance (Brantmeier, 2003; Hill & Liu, 2009; Pae, 2004). The 2009 GSEEE test scores from a random stratified sample of 13,745 test-takers from one major university in Southern China were collected and analyzed using differential item functioning (DIF) and differential bundle functioning (DBF). Preliminary results indicated that three grouping variables had a mixed effect on the GSEEE test score, with some items favoring the focal group and others favoring the reference group. Using psychometric approaches to investigate test fairness of the GSEEE, the study is engaged in an empirically driven discourse and has important implications in the area.

Comparison of the students’ performance in dynamic vs. static listening comprehension tests among EFL learners
Sahbi Hidri

Based on the traditional and psychometric static assessment (SA) and Vygotsky’s notion of dynamic assessment (DA) in the Sociocultural Theory of Mind, this study seeks to compare the students’ performance in dynamic vs. static LC tests in an EFL context. The study triangulated different data: a LC test with two parts, static and dynamic, and a retrospective interview (RI) which was administered to 60 test takers and six test raters and interviewers. The qualitative and quantitative results of the study indicated that learners opted for the dynamic nature of learning and assessment. The teachers somehow had the same opinion about DA but they remained quite doubtful about the scoring of the joint performance. In the dynamic LC test (DLCT), the teachers’ input impacted the students’ output. This impact was shaped by the mediation strategies (Lidz, 2002) the teachers were engaged in to interact with the learners. The static LC test (SLCT) also proved to be an appropriate measurement procedure to assess the individual performance in LC but it did not offer ample opportunities for learning. Recommendations were made to improve the status of teaching and testing LC dynamically by encouraging the teachers to consider the teaching practices from a dynamic perspective.

Investigating the yes/no vocabulary test: Input modality, context and response time
John Read, University of Auckland
Toshihiko Shiotsu, Kurume University

Lexical measures have been shown to function well not only to produce estimates of vocabulary size but also as general indicators of proficiency level, for purposes such as placement of learners in a language teaching program and for vocabulary acquisition research. The simplest kind of vocabulary measure is the Yes/No format, in which test-takers indicate whether they know each of a sample of target words. A proportion of the items are not real words in the target language, so that the scores of those who tend to overestimate their vocabulary knowledge can be adjusted accordingly. Several recent research studies have investigated the appropriate scoring procedures for the format (Beeckmans et al., 2001; Huibregtse et al., 2002; Mochida & Harrington, 2006). The project to be reported here extends the Yes/No research in three ways. First, it presents the target words in spoken as well as written form. Secondly, it explores how the addition of two types of sentence-based context influences performance on the Yes/No task. The third innovation is to investigate whether reaction
time adds a significant dimension to the measurement of vocabulary knowledge with this test format. The sample of target words has been drawn from the word frequency counts of the British National Corpus. Several computer-based forms of the test have been developed to address the various research questions (e.g. whether Yes/No format provides a reliable measure of listening vocabulary, how the provision of a co-text influences performance on both the oral and written versions of the Yes/No test, and whether reaction times account for test performance in a useful way, beyond what is measured by the accuracy scores), and the tests have so far been administered to 132 university students in Japan. In this session, the presenters will give a rationale for this approach to testing, discuss the design and development of the various versions of the test, and present a range of findings based on the data available to date. Participants will be encouraged to comment on the results and suggest further analyses.

Establishing standards for EFL tests in China: test construction, evaluation and use
Jinsong Fan

Lecture Room 5
Chaired by: Lorena Llosa

Understanding teachers’ classroom assessment practices in relation to students’ linguistic and affective changes: A case study in an EFL speaking course
Xiaoying Wang

Classroom assessment is a very important aspect of classroom teaching and learning. It is not only about grading but also about assisting the process of learning, and effective classroom assessment practices may enhance student learning and achievement (e.g., Black et al., 2004; Black & Wiliam, 1998, 2003; Brookhart & Durkin, 2003; Brookhart et al., 2006; Gipps, 1999; Rea-Dickins, 2006; Stiggins & Conklin, 1992). While a lot has been achieved concerning understanding of the influences of classroom assessment on students, three gaps have emerged through a review of the literature. First, previous studies have mainly focused on certain specific aspects of classroom assessment and examined their impact on students, while few studies have taken all the components of classroom assessment as a whole system into consideration to examine their effect on students. Secondly, previous studies have mainly taken students’ performances on some assessments as indicators of their achievement, while students’ voices about their experiences and perceptions of engaging in classroom assessment have not been sufficiently presented. Thirdly, classroom assessment is a highly localized practice, and the previous studies were conducted more in secondary education than in tertiary education, more in general education than in L2 education, and even within L2 education, more in ESL than in EFL contexts. Therefore, this proposed study will make an effort to examine the influences of classroom assessment as a system on students from a “whole person” perspective in an EFL context. More specifically, an in-depth multiple-case study will be carried out in order to gain a deep and comprehensive understanding of current assessment practices in a tertiary-level EFL speaking course in China as well as how students are engaged in such assessment practices and their opinions and perceptions. While this study will be guided by models and theories on classroom assessment (e.g., Brookhart, 1997; Cowie & Bell, 1999; Harlen, 2007), L2 motivation theory (Dornyei, 2005, 2009), and a social-cultural situated approach to learning (e.g., Wertsch, 1991), an emic approach will be pursued, which will enable the researcher to remain open and sensitive to the characteristics specific to the local context. This study will address the following research questions: What are the profiles of classroom assessment used in this course? Which one(s) is/are most frequently used? How does the daily formative assessment contribute to the final summative reporting? What do teachers perceive as classroom assessment? How do teachers look at the relationships between assessment, instruction, and learning? Do students’ self-confidence, goal orientations, self-identities, learning strategies, and perceived oral English proficiency change as a result of engaging in the classroom assessment activities in this course? If yes, how? Four teachers teaching oral English to first-year university students from four different universities in Beijing and their students will be recruited to participate in this study. Mixed methods will be used including classroom observation with recording, teacher interviews, student journals, student interviews,
Formative assessment practice in a Bachelor of Education/Teaching English as a Foreign Language Program (BEd/TEFL) in Mainland China: A case study
Yanming Wu

With the burgeoning movement towards Assessment for Learning and the increasing acknowledgement on the roles of self assessment and peer assessment in helping to create formative classroom assessment (Black & Wiliam, 1998; Assessment Reform Group, 1999) comes an urgent need to enhance preservice teachers' knowledge and training with self assessment and peer assessment as formative assessment. A few studies in the field of general education have begun to document that involving student teachers in self assessment and peer assessment in their preservice program can not only improve their own teaching, but also help student teachers transfer the skills of self assessment and peer assessment as formative assessment into their assessment practice with their students (Shin, Wilkins, & Ainsworth, 2006; Sluijsmans & Prins, 2006). However, there is still little published research on peer assessment and self assessment in the field of language teacher education (Bailey, 2006). Moreover, the implementation of self assessment and peer assessment as formative classroom assessment may face particular challenges in certain educational contexts such as China, as summative and teacher-dominant assessment have been the prevalent form of assessment for a long time (Cheng, 2008; Shing & King Fai, 2007; Zhao, 2008). To shed light on the aforementioned knowledge gaps, the current qualitative case study as a Master's thesis investigates a formative assessment practice implemented in a Bachelor of Education/Teaching English as a Foreign Language (BEd/TEFL) program in a university in mainland China. The assessment practice under study integrates instructor assessment/comments, peer assessment/comments, and student teachers' self assessment/reflection. The following research questions will be explored from the student teachers' perspectives: (1) Are there any differences or similarities among instructor assessment/comments, peer assessment/comments, and student teachers' self assessment/reflection? (2) How do student teachers perceive these three different forms of assessments, i.e., instructor assessment/comments, peer assessment/comments, and self assessment/reflection? (3) What impact does this formative assessment experience have on the students' learning to teach English as a Foreign Language in mainland China? Multiple research methods will be employed including content analysis of instructor comments, peer comments, and student teachers' self-reflection reports, student teachers' background surveys, and semi-structured interviews with individual student teachers. It is expected that this study will provide pedagogical implications for the use of formative peer assessment and self assessment in classroom setting in mainland China or other similar educational contexts, especially in preservice teacher education program specializing in Teaching English as a Second/Foreign Language. This session will present the preliminary results of the study and invite discussions on the interpretations of the data and future research directions.

Developing the assessment literacy of IELTS test users in higher education
Kieran O'Loughlin, The University of Melbourne

The rapid global growth of the IELTS test has resulted in a growing number of academic, administrative, marketing and other higher education staff providing information about the test, setting standards, interpreting scores and advising test takers both before and after they have sat for the test. Following recent theorising in the language testing field (e.g. Taylor, 2009), this research project aims to examine what are the assessment literacy needs of IELTS users in higher education, how well they are currently being met and how such efforts might be improved. The planned study builds on previous research into the use of the IELTS test in higher education conducted by the presenter. That earlier study concluded that there was a clear need for IELTS test users to be better educated. However, it also noted that exploring such issues as what type of and how much knowledge may be sufficient is pivotal to understanding the needs of different test users in higher education. The key
research questions to be investigated in this research project are: 1) what are the assessment literacy needs of different IELTS test users in higher education? and 2) what kinds of training program or other approach might best meet these needs? The study will take the form of a “proactive evaluation” (Owen, 2006) which will include a) an analysis of the assessment literacy needs of the various IELTS test users at the two Australian universities and how well these are matched by the currently available resources, b) a synthesis of research-based findings on the effectiveness of different kinds educational programs for staff in higher education and c) a review of international best practice in this area. These approaches will be used to arrive at a set of recommendations about developing the assessment literacy of test users in higher education. The findings will the contribute to the language testing field’s understanding of the evolving concept of assessment literacy (particularly in relation to standardised proficiency tests) and the ways in which it might best be developed amongst test users.


Investigating the use of English language proficiency tests for decision-making purposes in English-medium universities in Egypt
Elizabeth Arrigoni, American University in Cairo

English is increasingly becoming a language of instruction in higher education in much of the Middle East and North Africa (MENA) region. For example, of the 25 or so private universities established in the past two decades in Egypt, approximately 75% use English as the medium of instruction. Much of the expansion of English-medium instruction in Egypt results from the widespread belief that English language skills will improve access to job opportunities, both within Egypt and abroad, and improve the competitiveness of the Egyptian workforce. This relatively sudden increase in the number of universities offering instruction in English necessarily raises the following questions: What level of language proficiency is determined to be sufficient for study at an English-medium university in Egypt? On what basis do English-medium universities in Egypt make this determination? What instruments do Egyptian English-medium universities use to measure the language proficiency of its prospective and/or current students? What are some of the possible consequences and implications of the various uses of different English proficiency tests by English-medium universities in Egypt? Preliminary research into the types of English proficiency assessments used by English-medium universities in Egypt reveals that there is great variation in test use. Some universities require scores on specific tests, such as IELTS or TOEFL iBT, for admissions and placement. On the other hand, certain universities use English language proficiency tests only for placement in remedial or preparatory English courses, perhaps in recognition of the fact that the current system of secondary education in Egypt has not kept pace with the high demand for English language skills. The aim of this exploratory study is not only to determine whether, how, and why English-medium universities in Egypt use various English proficiency tests for informing decisions such as admissions and placement, but to identify some of the possible consequences, both positive and negative, of the various uses of such tests for decision-making purposes in this context. As McNamara and Roever (2006) have asserted, even tests that have been supported with validity evidence can have negative consequences; these may be compounded when such tests are used in different ways or in different contexts than those for which the test was intended. Qualitative data are being collected primarily through interviews with representatives of Egypt’s English-medium universities (N=31) in order to address the research questions. It is hoped that this session will elicit suggestions on the interpretation of the interview data, as well as identify ways that the current study can be expanded to examine English language proficiency test use in other contexts, both in Egypt and in other countries in the MENA region, based on findings from the current study.

Describing proficiency levels for young students of Italian as L2
Monica Barni, Università per Stranieri di Siena
Sabrina Machetti, Università per Stranieri di Siena
Carla Bagna, Università per Stranieri di Siena

This research investigates the possibilities and modalities of defining a level of L2 linguistic competence adequate for following school curricula. The public in question is that made up of foreign
students in compulsory education (6–16). This study has been produced within a context of worrying new legislation. The recent ‘Cota Bill’, which our research group has already discussed, at the 2009 EALTA Conference, proposes ‘bridge classes’ for foreign students who fail a language test: in other words, separate classes that, on paper, should lead to students’ eventually joining those for Italian-speaking students. This proposal is the clearest evidence of a centralistic language policy focused around Italian language acquisition, with the sole aim of creating forms of exclusion. Our research has two phases. It began in 2008 and lasts two years. The first phase (November 2008 – January 2009) took as its starting point a critical analysis of the Cota Bill, with its objective being to highlight the implications inherent in the tests proposed by the bill, which lacks any macroscopic description of them, any indication or definition of the competence level, any indication of who should design them, build them, experiment them or administer them. This lack reflects the ongoing situation in Italy whereby an assessment culture has yet to be established. Furthermore, politicians show an approach that is largely oblivious to scientific debate on the matter and gives a distorted reading of data on foreign students’ scholastic achievements. For this first phase, we also reviewed existing legislation, and discovered that with the Cota Bill Italy was for the first time at the centre of a highly discriminatory policy. The second phase (February 2009 –November 2010) has, and will continue to have, two main objectives: – the definition of a level of Italian L2 competence adequate for following classes at school. This definition seems to us to require reference both to the characteristics and motivations of the public in question and to the plurilingual context in which learning takes place; – a discussion of the possibility of assessing this level of competence with linguistic measuring and evaluation tools used not for questions of access but solely for descriptive and diagnostic purposes.

Language aptitude and the study of less commonly taught languages (LCTLs) by “average” learners in academic settings
Daniel Reed, Michigan State University
Seongmee Ahn, Michigan State University

This study investigates the relationship between language aptitude and the learning of less-commonly taught languages in academic contexts (in contrast to intensive classes in government language settings, which have been more thoroughly studied). There are two main research goals: 1) Documenting differences between student progress in government and academic settings for LCTL learning, with a focus on the progress of average and lower-than-average learners in LCTL programs; 2) Determining how learner motivation and language aptitude interact and relate to achievement in the learning of languages that are linguistically (and culturally) “distant from English” (the so-called “Category 4” languages). This study will target learners of Korean and Arabic in university language programs. The relationship between language aptitude and the attainment of various levels of language proficiency over a period of time has been studied extensively in the context of intensive instruction in government language programs, but less so in the typical academic programs in colleges and universities. There are interesting differences. In government programs it is common to select learners with high aptitude scores for the learning of more difficult languages (assuming a learner whose first language is English). In most colleges and universities, the students of these languages have other majors and interests and are studying the language to serve those interests, or merely as a secondary interest. These courses are also less intensive than government language courses. Overall, these programs have been successful (many students “pass” the courses), but the details of how successful the learning has been have yet to be well documented. Several related questions arise, including: Can individuals with lower-than-average language aptitude and perhaps average motivation learn these languages to useful levels in non-intensive, academic programs while simultaneously and seriously pursuing the study of other academic subjects? The design and analysis will assume Carroll’s classical framework of language aptitude and Carroll’s Model of School Learning. The results of the study will enable us to confirm whether learners with relatively high levels of language aptitude learn LCTLs that are dissimilar in structure to English more quickly and easily than their peers. The primary language aptitude measure used will be the Modern Language Aptitude Test (MLAT), and the proficiency instruments will include the Korean speaking proficiency test developed at a National Language Resource center.
EIKEN
Test in Practical English Proficiency

You spoke.
Educators and test users outside Japan increasingly request English-language information about EIKEN.

We responded.
STEP is pleased to announce the launch of our new English website, featuring:
- Overviews of the seven EIKEN grades
- Free downloads of test booklets and audio files
- Summaries of recent research
- Interactive registration system for schools recognizing EIKEN

stepeiken.org
Stop by. Discover EIKEN.
## Presenter Index

<p>| Ackermann, Kirsten | ........................................................................................................ | 74, 90 |
| Ahn, Seongmee | ........................................................................................................ | 109 |
| Alderson, Charles | ........................................................................................................ | 60, 71 |
| Alsborg, Jane | ........................................................................................................ | 94, 95 |
| Amaty, Pradyumna | ........................................................................................................ | 84 |
| Ameriks, Yoko | ........................................................................................................ | 55 |
| Arrigoni, Elizabeth | ........................................................................................................ | 108 |
| Bagni, Carla | ........................................................................................................ | 108 |
| Barker, Fiona | ........................................................................................................ | 77 |
| Barni, Monica | ........................................................................................................ | 108 |
| Baten, Lut | ........................................................................................................ | 93 |
| Beresford-Knox, Nick | ........................................................................................................ | 97 |
| Bernstein, Jared | ........................................................................................................ | 54 |
| Bhat, Suma | ........................................................................................................ | 101 |
| Black-Allen, Jesse | ........................................................................................................ | 85 |
| Boals, Tim | ........................................................................................................ | 51 |
| Borges-Almeida, Vanessa | ........................................................................................................ | 98 |
| Brewer, Sarah | ........................................................................................................ | 104 |
| Brown, Amanda | ........................................................................................................ | 69 |
| Brown, Annie | ........................................................................................................ | 58 |
| Brunfaut, Tineke | ........................................................................................................ | 88 |
| Butler, Yuko | ........................................................................................................ | 68 |
| Chen, Yuanshan | ........................................................................................................ | 81 |
| Cheng, Liying | ........................................................................................................ | 104 |
| Cheng, Peter | ........................................................................................................ | 96 |
| Chi, Youngshin | ........................................................................................................ | 51 |
| Cho, Bokyung | ........................................................................................................ | 82 |
| Clark, Martyn | ........................................................................................................ | 91 |
| Cline, Fred | ........................................................................................................ | 76 |
| Conso, Douglas | ........................................................................................................ | 98 |
| Cummins, Jim | ........................................................................................................ | 86 |
| Dakin, Je Wha | ........................................................................................................ | 55 |
| Davidson, Fred | ........................................................................................................ | 51, 54, 63, 101 |
| Dewey, Dan | ........................................................................................................ | 93 |
| Dewey, Kayleen | ........................................................................................................ | 93 |
| de Jong, John | ........................................................................................................ | 71, 74 |
| De Jong, Nivja | ........................................................................................................ | 64 |
| Doe, Christine | ........................................................................................................ | 101 |
| Douglas, Dan | ........................................................................................................ | 60 |
| Dusar, Lutgart | ........................................................................................................ | 93 |
| Davíd, Gergely A | ........................................................................................................ | 75 |
| Elder, Catherine | ........................................................................................................ | 84 |
| Emery, Henry | ........................................................................................................ | 90 |
| Fan, Jinsong | ........................................................................................................ | 106 |
| Fang, Felicia | ........................................................................................................ | 66 |
| Farris, Candace | ........................................................................................................ | 60 |
| ffrinch, Angela | ........................................................................................................ | 69 |
| Figuera, Neus | ........................................................................................................ | 51 |
| Fox, Janna | ........................................................................................................ | 102 |
| Friginal, Eric | ........................................................................................................ | 70 |
| Fujita, Tomoko | ........................................................................................................ | 96 |
| Fulcher, Glenn | ........................................................................................................ | 54, 63 |
| Galaczi, Evalina | ........................................................................................................ | 58 |
| Gao, Manman | ........................................................................................................ | 83 |
| Gottlieb, Margo | ........................................................................................................ | 51 |
| Grabowski, Kirby | ........................................................................................................ | 55 |
| Green, Anthony | ........................................................................................................ | 78 |
| Gu, Xiangdong | ........................................................................................................ | 87 |
| Halleck, Gene B | ........................................................................................................ | 66 |
| Hamp-Lyons, Liz | ........................................................................................................ | 52, 54 |
| Hasegawa-Johnson, Mark | ........................................................................................................ | 101 |
| He, Lianzhen | ........................................................................................................ | 104 |
| Hidri, Sahbi | ........................................................................................................ | 105 |
| Hildén, Raii | ........................................................................................................ | 100 |
| Hill, Kathryhn | ........................................................................................................ | 82 |
| Huensch, Amanda | ........................................................................................................ | 94, 95 |
| Huhta, Ari | ........................................................................................................ | 60, 68 |
| Hulstijn, Jan | ........................................................................................................ | 61 |
| Isaacs, Talia | ........................................................................................................ | 62 |
| Jamieson, Joan | ........................................................................................................ | 92 |
| Jang, Eunice | ........................................................................................................ | 86 |
| Jang, So-young | ........................................................................................................ | 94, 95 |
| Jaquith, Paul | ........................................................................................................ | 86 |
| Jones, Neill | ........................................................................................................ | 51, 96 |
| Jung, Wojin | ........................................................................................................ | 82 |
| Kemp, Jenny | ........................................................................................................ | 63 |
| Kenyon, Dorry | ........................................................................................................ | 51 |
| Khabbazbash, Nahal | ........................................................................................................ | 100 |
| Khalifa, Hanan | ........................................................................................................ | 69 |
| Kim, Ah-Young | ........................................................................................................ | 94 |
| Kim, Kyeongsuk | ........................................................................................................ | 94, 95 |
| Kim, Kyungyun | ........................................................................................................ | 82 |
| Kirsch, Irwin | ........................................................................................................ | 71 |
| Knoch, Ute | ........................................................................................................ | 60 |
| Lafontaine, Dominique | ........................................................................................................ | 71 |
| Laurier, Michael | ........................................................................................................ | 62 |
| Lepage, Sylvie | ........................................................................................................ | 51 |
| Liao, Yen-Fen | ........................................................................................................ | 55 |
| Lin, Chih-Kai (Cary) | ........................................................................................................ | 51, 96 |
| Liu, Min | ........................................................................................................ | 98 |
| Liu, Weiwei | ........................................................................................................ | 80 |
| Lockwood, Jane | ........................................................................................................ | 95, 97 |
| Lumley, Tom | ........................................................................................................ | 71 |
| Machetti, Sabrina | ........................................................................................................ | 108 |
| Malone, Margaret E | ........................................................................................................ | 78, 91 |
| Martinez, Ron | ........................................................................................................ | 72 |
| Matsushita, Hikotaku | ........................................................................................................ | 93 |
| May, Lyn | ........................................................................................................ | 58 |
| McCarthy, Michael | ........................................................................................................ | 65 |
| Mendelovits, Juliette | ........................................................................................................ | 71 |
| Meng, Fanna | ........................................................................................................ | 87 |
| Moder, Carol Lynn | ........................................................................................................ | 66 |
| Montee, Megan | ........................................................................................................ | 78 |
| Nakatsuhara, Fumiy | ........................................................................................................ | 58 |
| Nelson, Kathryn | ........................................................................................................ | 103 |
| Neumann, Heike | ........................................................................................................ | 80 |</p>
<table>
<thead>
<tr>
<th>Presenter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nier, Victoria</td>
<td>91</td>
</tr>
<tr>
<td>North, Brian</td>
<td>51</td>
</tr>
<tr>
<td>Nuttal, Martin</td>
<td>97</td>
</tr>
<tr>
<td>O’Loughlin, Kieran</td>
<td>107</td>
</tr>
<tr>
<td>Papageorgiou, Spiros</td>
<td>77</td>
</tr>
<tr>
<td>Papp, Szilvia</td>
<td>51, 69</td>
</tr>
<tr>
<td>Poonpon, Kornwipa</td>
<td>92</td>
</tr>
<tr>
<td>Present-Thomas, Rebecca</td>
<td>70</td>
</tr>
<tr>
<td>Purpura, James</td>
<td>55</td>
</tr>
<tr>
<td>Qian, David D</td>
<td>98</td>
</tr>
<tr>
<td>Raquel, Michelle</td>
<td>95, 97</td>
</tr>
<tr>
<td>Read, John</td>
<td>60, 105</td>
</tr>
<tr>
<td>Reed, Daniel</td>
<td>109</td>
</tr>
<tr>
<td>Revesz, Andrea</td>
<td>88</td>
</tr>
<tr>
<td>Saida, Chisato</td>
<td>99</td>
</tr>
<tr>
<td>Salamoura, Angeliki</td>
<td>51</td>
</tr>
<tr>
<td>Sanford, Ellie</td>
<td>76</td>
</tr>
<tr>
<td>Saville, Nick</td>
<td>51</td>
</tr>
<tr>
<td>Searle, Dara</td>
<td>71</td>
</tr>
<tr>
<td>Shiotsu, Toshihiko</td>
<td>105</td>
</tr>
<tr>
<td>Sim, Hyeonae</td>
<td>82</td>
</tr>
<tr>
<td>Slaght, John</td>
<td>104</td>
</tr>
<tr>
<td>So, Youngsoon</td>
<td>72</td>
</tr>
<tr>
<td>Song, Xiaomei</td>
<td>104</td>
</tr>
<tr>
<td>Spolsky, Bernard</td>
<td>54</td>
</tr>
<tr>
<td>Sproat, Richard</td>
<td>101</td>
</tr>
<tr>
<td>Stille, Saskia</td>
<td>86</td>
</tr>
<tr>
<td>Surface, Eric</td>
<td>103</td>
</tr>
<tr>
<td>Tamananen, Mirja</td>
<td>68</td>
</tr>
<tr>
<td>Taylor, Lynda</td>
<td>52</td>
</tr>
<tr>
<td>Tossavainen, Henna</td>
<td>102</td>
</tr>
<tr>
<td>Trim, John</td>
<td>51</td>
</tr>
<tr>
<td>Tsang, Carrie</td>
<td>66</td>
</tr>
<tr>
<td>Turner, Carolyn</td>
<td>62</td>
</tr>
<tr>
<td>Urston, Alan</td>
<td>66</td>
</tr>
<tr>
<td>Van Maele, Jan</td>
<td>93</td>
</tr>
<tr>
<td>van Moere, Alistair</td>
<td>58, 70</td>
</tr>
<tr>
<td>Verhelst, Norman</td>
<td>51</td>
</tr>
<tr>
<td>Vidakovic, Ivana</td>
<td>77</td>
</tr>
<tr>
<td>Vongpumivitch, Viphavee</td>
<td>75</td>
</tr>
<tr>
<td>von Randow, Janet</td>
<td>102</td>
</tr>
<tr>
<td>Wagner, Elvis</td>
<td>88</td>
</tr>
<tr>
<td>Wagner, Maryam</td>
<td>86</td>
</tr>
<tr>
<td>Wang, Xiaoying</td>
<td>106</td>
</tr>
<tr>
<td>Weigle, Sara</td>
<td>70</td>
</tr>
<tr>
<td>Wendler, Cathy</td>
<td>76</td>
</tr>
<tr>
<td>Wigglesworth, Gillian</td>
<td>58</td>
</tr>
<tr>
<td>Wilmes, Carsten</td>
<td>51</td>
</tr>
<tr>
<td>Wu, Yanming</td>
<td>107</td>
</tr>
<tr>
<td>Xi, Xiaoming</td>
<td>64</td>
</tr>
<tr>
<td>Xu, Jing</td>
<td>64</td>
</tr>
<tr>
<td>Yin, Chongbin</td>
<td>78, 91</td>
</tr>
<tr>
<td>Zeng, Wei</td>
<td>68</td>
</tr>
<tr>
<td>Zheng, Ying</td>
<td>90</td>
</tr>
<tr>
<td>Zulkifli, Putri AfzanxxMaria</td>
<td>96</td>
</tr>
</tbody>
</table>
Setting the New Standard!

Pearson Tests of English (PTE) is a suite of English language tests designed to assess the English language usage of students who are speakers of other languages. These include PTE Academic, a new computer-based academic English language test endorsed by GMAC®, PTE General and PTE Young Learners.

Pearson Test of English Academic - PTE Academic:
• assesses performance in all four language skills in one single three-hour session
• provides additional feedback on grammar, oral fluency, pronunciation, spelling, vocabulary and written discourse skills
• a true test of academic English using real-life examples and authentic academic content
• robustly aligned with the Common European Framework (CEF)

Pearson Test of English General - PTE General:
• a scenario-based English language test with emphasis on authentic, relevant situations
• offered at six levels ranging from A1 to 5 linked to the CEF
• has one written paper testing Listening, Reading and Writing and a separate Oral examination at each level
• awarded by Edexcel, one of the largest examination bodies in the UK

Presentations at LTRC | 10.30 am Thursday 15 April
Using the Common European Framework to rate test taker responses
Kirsten Ackermann

Getting the levels right: deriving item writer guidelines from CEFR descriptors
John de Jong, Glyn Jones

Poster presentation
Knowing the test takers: investigating test taker background, test taker characteristics and their test performance
John de Jong, Ying Zheng, Kirsten Ackermann

Visit our stand outside the conference room or visit www.pearsonpte.com for more information

PEARSON LANGUAGE TESTS
www.pearsonpte.com
Additional Information for Delegates

Badges
Please make sure you wear your badge at all times throughout the conference and social events. This makes for easier social networking and it also assists with security.

Attendance at the Gala Banquet or any Pre-Conference Workshops will be marked on the badge.

If you have any questions or require assistance at any time during the conference, please ask a member of the conference administration team who can be identified by a staff badge including the conference logo and white badge ribbon.

Exhibiting staff can be identified by a blue badge ribbon.

Parallel sessions – Thursday and Friday
On Thursday and Friday, the conference will begin with parallel sessions. Please go directly to the room of your chosen parallel session for 8.30. Please refer to the programme for details.

Lunch
Please note there are no restaurants or coffee shops located on or near West Road. The nearest places for food and drink are based in the city centre which is approximately a 10–15 minute walk from West Road.

Pre-Conference Workshops (Monday and Tuesday)
Mid-morning/mid-afternoon tea, coffee, etc will be provided free of charge. A sandwich lunch is included in the Pre-Conference Workshop fee on Monday and Tuesday served in the Faculty of Music. A vegetarian option is available.

Main Conference (Wednesday, Thursday and Friday)
Mid-morning/mid-afternoon tea, coffee, etc will be provided free of charge. A subsidised, snack lunch can be purchased on Wednesday, Thursday and Friday at the Faculty of Music. Lunch includes a bap, foccacia bread or baguette with a choice of 2 fillings (meat or vegetarian option), a piece of fruit, a piece of cake and a soft drink.

Lunch costs £5 (Sterling). Please note that payment can only be made in cash and no other currency can be accepted.

It is not necessary to order your lunch in advance. An appropriate number of lunches will be ordered in advance for delegates to purchase during the conference on a first-come, first-served basis.

Exhibition
The Exhibition is open every day of the Main Conference. Refreshments and lunch are also served in this room on Wednesday, Thursday and Friday.

Wednesday 14 April  10.00 – 5.00
Thursday 15 April  8.15 – 4.30
Friday 16 April  8.15 – 4.00

Exhibiting staff can be identified by a blue badge ribbon.
Computer Room with internet access

A computer room with free WIFI internet access is available at the Faculty of Music:

Wednesday 14 April  8.15 – 5.00
Thursday 15 April  8.15 – 4.30
Friday 16 April  8.15 – 4.00

Contact the Registration Desk in the Faculty of Music Foyer with queries or technical issues.

Conference Photo

A group photo will be taken of all delegates who attend the conference. This will be taken during the morning refreshment break on Wednesday. The group photo will be available on the conference website (www.cambridgeesol.org/LTRC2010) for free download after the conference.

Cloakroom and Luggage Storage

Coat racks and hangers are available at the Faculty of Music. Please note these facilities are not secure and the cloakroom is not staffed. We strongly recommend that you do not leave valuables in this area but keep them with you at all times. The Conference Organisers cannot be held responsible for any loss of personal belongings.

An area for storing luggage in the Faculty of Music is only available on the final day of the conference (Friday 16 April). This area will not be locked.

Noticeboard

A noticeboard is provided in the Foyer of the Faculty of Music for delegates to display personal or business messages.

Access to the venue through King's College

Conference delegates can walk through the grounds of King's College between the hours of 7.00 – 6.00 from Monday to Friday. Delegates should make sure that their conference badge is visible at all times when walking through the grounds.

Cambridge University Press Bookshop Discount

A 15% discount voucher for the bookshop (valid for the duration of the conference) is included in all delegate bags.

Cambridge University Press Bookshop
1 Trinity Street
Cambridge
CB2 1SZ

University of Cambridge Merchandise

A 10% discount has been arranged at one of the best known University Outfitting companies in Cambridge. Ryder & Amies are the original University outfitter and the largest stockist of Cambridge T-Shirts anywhere; all shirts are top quality, with all University designs approved and licensed by the University of Cambridge.

A 10% discount voucher will be available to all LTRC delegates valid for the duration of the conference.

Ryder & Amies
22 King's Parade
Cambridge
CB2 1SP
Information about Cambridge

The Cambridge Tourist Information Centre
Peas Hill
Cambridge
CB2 3AD

Opening hours:
Monday – Saturday  10.00 – 5.00
Tel 0871 226 8006

Smoke Free Legislation

Smoke free legislation became effective from 1 July 2007 stating that virtually all enclosed public places and workplaces in England will be smoke free. This includes cafes, restaurants, bars, hotels and conference venues.

Smoking is only permitted on public footpaths at the conference venues.

Smoking is only allowed at King’s College in the Chetwynd Courtyard.

Smoking is only permitted on public footpaths outside of the Museum of Zoology grounds.
FREE LECTURER COPIES AVAILABLE!

Understanding Language Testing
DAN DOUGLAS, Professor of TESL and Applied Linguistics, Iowa State University, USA

An ideal title for student teachers taking modules in applied linguistics, TESOL and students of linguistics taking modules in SLA and language teaching:

■ Unique – the first book of its kind to include a chapter on technology and language testing

■ Approachable – Theories are presented in accessible language ensuring no prior knowledge of the subject matter is required

■ Comprehensive – Exercises consolidate learning helping students put what they have learnt into practice

978 0 340 98343 0 • January 2010 • 176pp • PB • £19.99

Practical Language Testing
GLENN FULCHER, Senior Lecturer in Education (TESOL), University of Leicester, UK

‘Dr Fulcher’s book is a good, practical synthesis of “how-to” skills, theoretical knowledge about testing, and principles of test development and use, and makes all these relevant to classroom-based ESL assessment.’

F. Scott Walters, Queens College, City University of New York, USA

■ Covers all stages of test construction from theory to designing tests, to prototyping and piloting, analysis and dealing with external standards

■ Provides examples from a variety of real life settings relevant to all students and practitioners with an interest in language testing

■ Offers individual and group activities to consolidate learning

978 0 340 98448 2 • March 2010 • 304pp • PB • £19.99

To request a free lecturer copy, e-mail academic.marketing@hodder.co.uk with your name and institution details, complete with course name and student numbers.
In Memoriam – Caroline Clapham

Dr Caroline Margaret Clapham (1942–2009)

It was with great sadness that many of us learned of the death of our former colleague Dr Caroline Clapham who passed away at her home in Mepal, near Cambridge, on 14th December 2009.

Caroline worked in Cambridge ESOL from September 2002 until the end of March 2004 as a member of the team working on IELTS within the Research and Validation Group. During her time with us she made many friends and is fondly remembered for her modesty and friendly nature, as well as for her wide knowledge of language testing. Unfortunately, not long after starting to work for Cambridge ESOL, Caroline was diagnosed with the illness which led to her retirement on grounds of ill-health, cutting short a distinguished career in such an untimely way.

Before joining Cambridge ESOL, Caroline had already established herself as a leading expert in language testing and as a much-loved member of the language testing community at large. She was an experienced language tester and in her early career in the 1970s constructed the first versions of the “PLAB test” for the Professional and Linguistics Assessment Board of the General Medical Council in the UK for testing the listening comprehension of overseas doctors.

Caroline was an established researcher and in coming to Cambridge after many years working for the University of Lancaster she was well-placed to join the IELTS research team, having been the research coordinator for the IELTS development project in the 1980s. Her doctoral dissertation focussed on the development of IELTS investigating the effect of background knowledge on reading comprehension. This work contributed significantly to the revision of IELTS in 1995 and was published in 1996 as Volume 4 in the Studies in Language Testing series: *The Development of IELTS: A study of the effect of background knowledge on reading comprehension*. The quality of this work was recognised in the same year when she was presented with the Jacqueline A. Ross Dissertation Award.

Caroline was always generous with her time and supported people in many different ways. She regularly reviewed for academic journals, contributed to editorial advisory boards and served the International Language Testing Association (ILTA) as both Secretary/Treasurer and President. Together with her colleague Dianne Wall, she edited *Language Testing Update* for many years and in recognition of her outstanding service, she was awarded an honorary membership of ILTA (see the original citation below). Caroline will remain an inspiration to the many who knew her, and especially to those who worked with her or who were taught by her. For over 20 years she advised aspiring language testers from around the world and was widely respected by her students for her supportive supervision and fair judgements as a doctoral examiner.

After her retirement, colleagues from Cambridge ESOL continued to visit her at home and she was always eager to catch up with news from the world of language testing. Those who visited her commented on her enduring good humour and dignity in the face of deteriorating health. We are now considering ways in which we can honour and mark Caroline’s contribution to the field of language testing. All of us who knew and respected Caroline extend our deepest sympathy to her husband Christopher, and to her children Phoebe and Tom.

(The above text is slightly adapted from the tribute posted to the ESOL website a few days after Caroline’s death in December 2009: www.cambridgeesol.org/what-we-do/newsroom/2009/caroline-clapham.html)
In 2005 the ILTA Executive Board awarded Caroline Clapham honorary membership of ILTA with the following citation:

In recognition of her outstanding service to the International Language Testing Association (ILTA), and her distinguished contribution to the field of language testing in general, the Executive Board has voted unanimously to recommend that Dr. Caroline Clapham be awarded an honorary membership of ILTA (in accordance with the ILTA By-Laws, Article 1, Section 3).

Caroline Clapham has made a long and dedicated commitment to ILTA and to language testing in general. She was an Officer of ILTA – Secretary/Treasurer and President – for longer than any other person to date. She performed exemplary service in all those roles and made tremendous contributions to our professional organization while maintaining the most cheerful and gentle demeanour.

Caroline’s contributions to the profession started in the 1970s with her work on the UK General Medical Council’s test for doctors, the Professional and Linguistic Assessment Board (PLAB). She was also involved in the development of the original ELTS test, now the IELTS test and with the many revisions this has since undergone. She wrote an excellent Ph.D, that raised and elaborated significant questions about the testing of English for specific purposes and was joint-winner of the TOEFL award for best dissertation in 1996; it is now a much referred to volume in the SILT series published by UCLES/CUP. The Encyclopaedia volume that she edited with David Corson (Kluwer) is also much cited. She was Editor of Language Testing Update with Dianne Wall for many years and was also a long serving member of the Editorial Advisory Board of Language Testing, offering reliable and constructive reviews for the journal.

For three decades Caroline has given her support and careful advice unstintingly to a new generation of language testers across the world and is widely revered by her former students as an amazing doctoral supervisor. Her last appointment was as IELTS validation officer for Cambridge ESOL until ill health sadly forced her to resign.

As part of our memorial to Caroline, there will be a small table set aside at LTRC 2010 with a photograph and a selection of her publications.
ASSESSMENT FROM OXFORD UNIVERSITY PRESS

Find out what Oxford has to offer in assessment theory and practice

The Oxford Online Placement Test available at www.oxfordenglishtesting.com

A test designed with busy teachers and administrators in mind:

- Save time and resources with automatic marking and instant results.
- Test students where it’s most convenient: assign by email to be taken remotely, or print ready-made login slips for students in class.
- Use a variety of scores to make your placement decision: CEFR level, time taken, score between 0-120, ‘Use of English’ section as well as ‘Listening’, plus customizable columns for additional information.
- Be confident you’ll make the right decision: The test has been pretested and validated by more than 10,000 students in more than 40 countries.

Hear more about the test during the talk *How do we define grammatical knowledge in terms of form and meaning dimensions at six different CEFR Proficiency Levels?* presented by Professor James Purpura and Jee Wha Dakin.

Meet the authors – morning coffee break on 15th April

Meet Adrian Palmer and Lyle Bachman at the OUP stand in the Recital Room to discuss *Language Assessment in Practice*, the fully revised second edition of the bestselling *Language Testing in Practice*:

- Takes into account the most recent developments in research in applied linguistics, language teaching, and language assessment.
- Draws on the authors’ extensive experience in language teaching and assessment.
- Illustrative test development projects guide readers through the process.
- Additional website material to accompany the book.

Receive a 20% discount on your book order at the OUP stand

32nd Language Testing Research Colloquium, Cambridge – 12–16 April 2010
More than four decades of research validate the TOEFL® test

Read the latest studies, monographs and reports on the TOEFL® test.


www.ets.org/toeflresearch
Maps and plans

West Road site overview

Faculty of Music

Ground floor