Exploring Diverse Methodologies and Conceptualizations in Language Testing Research

University of Barcelona, Barcelona, Spain

UNIVERSIDAD DE BARCELONA
C/ Gran Via de les Corts Catalanes, 585
08007 Barcelona
Catalonia, Spain

Sincere thanks to Cambridge ESOL Exams for the design and production of the LTRC 2007 Program Book.
Table of Contents

LTRC 2007 Conference Organizers and Committee Members 4
LTRC Special Events 7
LTRC 2007 Conference Sponsors 8
Message from the ILTA President 9
About ILTA 11
  • ILTA Goals 11
  • ILTA Executive Board 2007 11
  • ILTA Committees 12
  • ILTA Institutional Affiliates 13
  • A Little LTRC History 13
  • ILTA Presidents 15
LTRC 2007 Program Overview 16
LTRC 2007 Full Program 20
  • Thursday, June 7 20
  • Friday, June 8 20
  • Saturday, June 9 22
  • Sunday, June 10 24
  • Monday, June 11 28
UCLES /ILTA Lifetime Achievement Award 2007 31
2007 Samuel J. Messick Memorial Lecture 33
Abstracts: 34
  • Workshop 1 34
  • Workshop 2 34
  • Workshop 3 35
  • Workshop 4 36
  • Symposium 1 37
  • Symposium 2 38
  • Symposium 3 40
  • Papers 42
  • Works-in-Progress 60
  • Posters 65
News from the ILTA Affiliates 74
LTRC 2007 Conference Organizers and Committee Members

LTRC 2007 Organizing Committee

Conference Co-Chairs
James E. Purpura, Teachers College, Columbia University
Carolyn E. Turner, McGill University

Local Co-Chairs
Teresa Navés i Nogués, University of Barcelona
Laura Sanchez Perez, University of Barcelona
Carmen Velasco, Spanish Embassy, Washington D.C.

Members of the Organizing Committee
Lorena Llosa, New York University
Nick Saville, University of Cambridge ESOL Examinations
Elvis Wagner, Teachers College, Columbia University
Sarah Creider, LTRC 2007 Webdiva, Teachers College, Columbia University

Lifetime Achievement Award Committee
Chair: John L.D. Clark, Defense Language Institute
Glenn Fulcher, University of Leicester
James E. Purpura, Teachers College, Columbia University
Nick Saville, University of Cambridge ESOL Examinations

ILTA Best Paper Award Committee
Chair: Annie Brown, Ministry of Higher Education, UAE
Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Miyoko Kobashi, University of Warwick
Dorry Kenyon, Center for Applied Linguistics
Lado Award Committee for Outstanding Student Paper at LTRC

Chair: Liz Hamp-Lyons, University of Hong Kong/University of Nottingham
Rob Schoonen, University of Amsterdam
Xiaoming Xi, Educational Testing Service

ILTA Student Travel Award Committee

Chair: Glenn Fulcher, University of Leicester
Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Lynda Taylor, University of Cambridge ESOL Examinations

Conference Proceedings

James E. Purpura, Teachers College, Columbia University
Carolyn E. Turner, McGill University

ILTA Business Office
Prime-Management, Inc.
3416 Primm Lane
Birmingham, AL 35216 USA
(1) (205) 823-6106
ILTA@primemanagement.net (Robert Ranieri)

LTRC 2007 Volunteers

It takes many people to organize a conference like LTRC. We would like to thank all the student volunteers—those who volunteered early enough to make it into the program book and those who stepped up to the plate later on.

Yorena Alcalà, University of Barcelona
Marcela Chapeton, University of Barcelona
Jee Wha Dakin, Teachers College, Chapetón University
Mar Galindo Merino, University of Alicante
Kristen di Gennaro, Teachers College, Columbia University
Kirby Grabowski, Teachers College, Columbia University
Cecilia Guanfang Zhao, New York University
Ah-Young (Alicia) Kim, Teachers College, Columbia University
Hyunjoo Kim, Teachers College, Columbia University
Hyun Jung Kim, Teachers College, Columbia University
Yen-Fen Liao, Teachers College, Columbia University
Danica Salizar, University of Barcelona
Taejoon Park, Teachers College, Columbia University

Proposal Reviewers

Lyle Bachman, University of California, Los Angeles
Vivien Berry, The University of Hong Kong
Geoff Brindley, Macquarie University
Annie Brown, Ministry of Higher Education, UAE
Gary Buck, University of Michigan
Frances Butler, Language Testing Consultant
Nathan Carr, California State University, Fullerton
Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Carol Chapelle, Iowa State University
Christine Coombe, Dubai Men’s College, UAE
Alister Cumming, OISE/University of Toronto
Fred Davidson, University of Illinois at Urbana-Champaign
Alan Davies, University of Edinburgh
John H.A.L. De Jong, Pearson Language Assessments
Craig Deville, Measurement, Inc.
Barbara Dobson, University of Michigan
Dan Douglas, Iowa State University
Usman Erdösy, Carleton University
Hossein Farhady, Iran University of Science & Technology
Janna Fox, Carleton University
Evelina Galaczi, University of Cambridge ESOL Examinations
Liz Hamp-Lyons, University of Hong Kong / University of Nottingham
Ari Huhta, University of Jyväskylä
Dorry Kenyon, Center for Applied Linguistics
Antony Kunnan, California State University, Los Angeles
Constant Leung, King’s College London
Jo Lewkowicz, The University of Warsaw
Sari Luoma, Ballard & Tighe Publishers
Meg Malone, Center for Applied Linguistics
Tim McNamara, University of Melbourne
Barry O’Sullivan, Roehampton University
Adrian Palmer, University of Utah
India Plough, University of Michigan
David Qian, Hong Kong Polytechnic University
John Read, University of Auckland
Carsten Roever, University of Melbourne
Steven Ross, Kwansei Gakuin University
Yasuyo Sawaki, Educational Testing Service
Charles Stansfield, Second Language Testing, Inc.
Lynda Taylor, University of Cambridge ESOL Examinations
Randy Thrasher, Okinawa Christian University
Yoshinori Watanabe, Sophia University
Sara Weigle, Georgia State University
Gillian Wigglesworth, University of Melbourne
Xiaoming Xi, Educational Testing Service
LTRC 2007 Special Events

Welcome Reception
Friday, June 8, 7:15-9:00 p.m.
Sponsored by University of Cambridge ESOL Examinations

Wine-tasting
Saturday, June 9, 7:45 p.m.
With support from Educational Testing Service

Student-led Session
Sunday, June 10, 7:45-9:30 p.m.

LTRC Banquet & Awards Presentations
Monday, June 11, 9:00 p.m.

Can Cortada: Av. de l’Estatut de Catalunya, s/n Barcelona (Horta)

- The 2007 ILTA Travel Award for Student Presenters
  Winners: Khaled Barkaoui, Ontario Institute for Studies in Education, University of Toronto and Taejoon Park, Teachers College, Columbia University
- The 2006 Best Paper in Language Testing Award
- ILTA’s 2007 Lado Award for Outstanding Student Paper at LTRC
- The IELTS Masters Award (sponsored by the IELTS Partners)
  Winner: Youn-Hee Kim, McGill University
The LTRC 2007 Organizing Committee would like to express our most sincere gratitude to our sponsors for their generous financial support of LTRC 2007.

Cambridge University ESOL Examinations
Cambridge University Press
Educational Testing Service
International Language Testing Association
McGill University
NetworkOmni Multilingual Communications
Pearson Language Assessments
Sage Publications
Second Language Testing, Inc.
Taylor & Francis
Teachers College, Columbia University
Universitat de Barcelona
Message from the ILTA President

Dear Friends and Colleagues,

On behalf of the Executive Board and all members of the International Language Testing Association, I would like to welcome you to the 2007 Language Testing Research Colloquium in Barcelona. We have wanted to take LTRC to Barcelona for a long time and the day has come.

This year’s conference theme is: “Exploring diverse methodologies and conceptualizations in language testing research.” Some 211 proposals were submitted for review. Of these, 3 symposia, 29 papers, 16 posters and 9 works-in-progress were selected. This year’s LTRC offers 4 pre-conference workshops: one on basic and one on advanced statistical analyses for language testing, one on qualitative methods in test development and validation, and one that examines the social context as a means of understanding assessment as a social practice. This is the first year that two workshops extend over a two-day period. Finally, this year the Samuel J. Messick Memorial Lecture, sponsored by Educational Testing Service, will be delivered by Bob Mislevy. Bob will discuss the promises that an emerging synthesis of cognitive and sociocultural insights might hold for the design and analysis of language assessments.

As we have done in years past, ILTA, together with the Cambridge University ESOL Exams, is proud to acknowledge long-established members who have distinguished service and scholarship records in the field of language testing. The UCLES/ILTA Lifetime Achievement Award will be given to Charlie Stansfield at LTRC 2007. Charlie’s lecture will discuss where the field of language testing has been and where we need to go. Also, the best paper in language testing for 2006 will be announced at the LTRC banquet.

ILTA has always encouraged students to attend LTRC, present their research, join ILTA and get involved in the organization. Many of us began as student LTRC volunteers. In this effort, ILTA is proud to provide support for two students to travel to LTRC. The ILTA 2007 travel award is given to those students with the most promising abstracts.
The 2007 awardees are Khaled Barkaoui and Taejoon Park. ILTA also gives the Lado Award for the outstanding student paper at LTRC.

One new initiative this year has included a formal agreement between AAAL and ILTA to hold an annual joint symposium, where researchers from language testing and other fields of applied linguistics can “explore areas of common interest, both substantive and methodological, from a variety of perspectives” (Bachman & Cohen, 1998, p. xi). This interest in mutual dialogue led to the organization of a symposium at AAAL 1992 in which SLA and language testing researchers discussed areas of interface. The first AAAL/ILTA joint symposium was held at LTRC 2000 in Vancouver. Since then, other successful joint symposia have been organized; however, no formal arrangement between the organizations has been made, until now. This year, Lynda Taylor and Gillian Wigglesworth have organized the symposium at AAAL 07 in Costa Mesa, California. The title was: “Are two heads better than one? Pair work in L2 learning and assessment.”

This is an exciting time for ILTA. This is the first time that we have contracted the services of a business office. Beginning in October 2006, Prime Management’s first charge was to help us update our membership database and encourage lapsed members to renew. In this effort, PM programmed the ILTA website to allow members to renew on-line. Members are now reminded by e-mail when their membership is about to lapse. New members can also join on-line. This service has already begun to pay off as membership has increased to 190 with members representing 26 countries. If you are not an ILTA member, I encourage you to join. If you are a member, please encourage your students to join. Without a strong membership, we would be financially pressed to support our awards, the new website, the newsletter, on-line voting and conference registration, professional development activities, and of course, future LTRCs.

On that note, I hope you have a thoroughly enjoyable and enriching conference in Barcelona.

With best wishes,

Jim Purpura

President of ILTA 2007
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 Executive Board 2007

President: James E. Purpura, Teachers College, Columbia University
Vice-President: Carolyn E. Turner, McGill University
Secretary: Randy Thrasher, Okinawa Christian University
Treasurer: Craig Deville, Measurement Inc.
Immediate Past President: Glenn Fulcher, University of Leicester

Members-at-Large 2007
Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Gillian Wigglesworth, University of Melbourne
Annie Brown, Ministry of Higher Education, UAE
Lynda Taylor, University of Cambridge ESOL Examinations
ILTA Committees 2007

ILTA Nominating Committee 2007
Elvis Wagner, Chair, Teachers College, Columbia University
Yong-Won Lee, Seoul National University
Lorena Llosa, New York University
John Norris, University of Hawai‘i

ILTA Committee on Affiliation 2007
Dan Douglas, Chair, Iowa State University
Fred Davidson, University of Illinois, Champaign-Urbana
Francis Butler, Language Testing Consultant
Antony Kunnan, California State University, Los Angeles
Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Liz Hamp-Lyons, University of Hong Kong/University of Nottingham

ILTA Newsletter 2007
Vivien Berry, Editor-in-Chief, The University of Hong Kong
Michael Chau, Webmaster, Hong Kong Polytechnic University
Dan Douglas, Iowa State University
Yang Lu, Reading University
Elvis Wagner, Teachers College, Columbia University

A special thanks to Liz Hamp-Lyons for her leadership and hard work on the ILTA Newsletter since its inception

ILTA Archivists 2007
Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Craig Deville, Measurement Inc.

ILTA Task Force on Testing Standards Update 2007
Samira Elatia, Chair, University of Alberta
Nathan Carr, California State University, Fullerton
Fred Davidson, University of Illinois, Champaign-Urbana
Alexis A. Lopez, Universidad de los Andes
Paul Jaquith, Ministry of Education, United Arab Emirates
Ana Oscoz, University of Maryland, Baltimore County
Sang-kuen Shin, Ewha Women’s University

ILTA Task Force on the Code of Practice
Alan Davies, Chair, University of Edinburgh
Charles Alderson, Lancaster University
Fred Davidson, University of Illinois at Urbana-Champaign
Antony Kunnan, California State University, Los Angeles
Liz Hamp-Lyons, Univ. of Hong Kong/University of Nottingham
Charles Stansfield, Second Language Testing, Inc.
Randy Thrasher, Okinawa Christian University
ILTA Institutional Affiliates

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

A Little LTRC History

This is the 29th annual LTRC. Every now and again, it helps to look back to see how far we've come. The following piece, taken from the 10th LTRC Program Book (1988), was written by Lyle Bachman and Buzz Palmer.

Nearly ten years ago a relatively small group of applied linguists from several countries, with varied backgrounds and perspectives, braved the rigors of the New England winter and the vagaries of convention planning, which forced them to squeeze into a single hotel room for two days, in order to participate in a discussion of the issues and problems of language testing. For many who were at that first colloquium, it was both surprising and gratifying to find that there were others who shared their interest in this abstruse and relatively unpopular enterprise. Equally amazing was the discovery that not only did we share common professional interests, but we were in many other ways “kindred spirits”. As Palmer, Groot and Trosper (1981) observed in their forward to the volume of papers that came out of that first colloquium, “the colloquium has enabled people with a common narrowly-defined interest to get to know each other and to develop the closeness and the lines of communication that allow each to profit more fully from the work of others” (p. vii). That observation is as true today as it was then. Indeed, it is this spirit of camaraderie, of personal closeness, based on mutual respect, that has nourished the spirited annual debates of professional issues and year-around communication among colleagues who are also friends that have become the hallmarks of the Language Testing Research Colloquium.

The past ten years have seen many developments in language testing, and the Colloquium, with its focus on research, has provided the sounding board for many of these. The use of confirmatory factor analysis as an approach to the construct validation of language tests was forged by the debate and collective effort among the participants at the first Colloquium. Research in oral testing has been a frequent topic at the Colloquium, and it is safe to say that many refinements in the measurement of oral ability have been influenced by discussions at colloquia. More recently, the applications of item-response theory and multidimensional scaling to language testing research have provided fodder for the Colloquium’s cannon, and have emerged all the stronger and more promising in the process. Not all of these developments have been chronicled, but the four published volumes that have come out of the Colloquium (Palmer, Groot, Troesper 1981, Jones, DesBrissay & Paribakhp 1985, Stansfield 1986, and Bailey, Dale & Clifford 1987) provide an overview of the issues that have captivated the field over the past ten years.
The first Colloquium came at a time when John Oller's research into the nature of language ability was redefining our view of language testing and at the same time raising questions that would lead to the emergence of language testing as a subfield of applied linguistics in its own right, with its own research questions, and with a research methodology that would contribute to other areas of applied linguistics. It also came at a time when Mike Canale and Merrill Swain were formulating the ideas that would emerge in their seminal paper on teaching and testing communicative competence. As a result of these cross-currents, it is not surprising that the focal points that emerged from the first Colloquium were an interest in a broader view of language proficiency as communicative competence, and a determination to embark on a program of empirical research into the then relatively unknown realm of construct validation.

The Colloquium has had a variety of themes over the years, and in some years has had no particular theme, other than a focus on research. We felt it was timely, in this tenth year of our persistence, to return to the theme of the first colloquium: the validation of tests of communicative language ability. Timely, because although “communicative” as a buzz-word has lost a certain amount of cachet in language teaching, it appears that language testers are beginning to come to grips with what the characteristics of “communicative” tests are. Thus, language testing may offer one avenue for investigating the nature of both communicative language use and the very abilities that make such use possible. One question we may ask ourselves, then, is, “How far have we really come in the past ten years toward understanding the language abilities that we profess to measure?”

We might also look at technological developments as an indication of our emergence as a field. In 1979, technical sophistication focused primarily on research design and the analysis of results-- Clifford’s examination of multitrait-multimethod correlations and Engelskirchen, Cottrell and Oller’s principal components analysis were “state-of-the-art”. In the past four or five years, we have seen the increasing application of technology to test design and administration, and we now have at our disposal not only a wider range of analytic tools, but more powerful ones as well. Appropriate questions to ponder in this regard are, “How much have these technical advances contributed to our understanding of the fundamental issues of language testing?” and “Are we simply probing the same questions in greater detail, or are we asking new questions?”

While the tenth annual Colloquium is perhaps no more special than was the ninth or than will be the eleventh, we feel this is an occasion to celebrate the remarkable “staying power” of the Colloquium. As Stansfield (1986) noted in his introduction to the papers from the seventh Colloquium, we have no charter, no officers, no dues. Every year there are new faces and new perspectives which add to the fabric of our collective identity. The Colloquium endures in spite of our resistance to becoming formally organized. It thrives because of our common interest in and commitment to the field of language testing, and because we truly enjoy our work, especially bashing heads once a year. And we have fun together. After all, some of our best friends are language testers.
Chronology of Language Testing Research Colloquium and Collections of Colloquium Papers

First  LTRC: Boston, TESOL, 1979

Second  LTRC: San Francisco, TESOL, 1980

Third  LTRC: Honolulu, TESOL, 1981

Fourth  LTRC: Ann Arbor, University of Michigan, 1982

Fifth  LTRC: Ottawa, University of Ottawa, 1983

Sixth  LTRC: Houston, TESOL, 1984


Eighth  LTRC: Monterey, Defense Language Institute, 1986


Tenth  LTRC: Urbana, University of Illinois 1988

ILTA Presidents

<table>
<thead>
<tr>
<th>Year</th>
<th>President</th>
<th>Year</th>
<th>President</th>
<th>Year</th>
<th>President</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007-8</td>
<td>James Purpura</td>
<td>2002</td>
<td>Fred Davidson</td>
<td>1997</td>
<td>John Clark</td>
</tr>
<tr>
<td>2006</td>
<td>Glenn Fulcher</td>
<td>2001</td>
<td>Caroline Clapham</td>
<td>1996</td>
<td>Lyle Bachman</td>
</tr>
<tr>
<td>2005</td>
<td>Dan Douglas</td>
<td>2000</td>
<td>Alan Davies</td>
<td>1995</td>
<td>Bernard Spolsky</td>
</tr>
<tr>
<td>2004</td>
<td>Antony Kunnan</td>
<td>1999</td>
<td>Elana Shohamy</td>
<td>1993-4</td>
<td>Charles Alderson</td>
</tr>
</tbody>
</table>
# LTRC 2007 Program Overview

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Thursday, June 7:</strong></td>
<td>Pre-Conference Workshop</td>
<td></td>
</tr>
<tr>
<td>9:00 - 9:30</td>
<td>Registration</td>
<td>Outside Paranimf Hall</td>
</tr>
<tr>
<td>9:30 - 5:00</td>
<td><strong>Workshop 1:</strong> Language Testing: The Social Dimension</td>
<td>Room 1.1</td>
</tr>
<tr>
<td>1:00 - 2:30</td>
<td>(Lunch)</td>
<td></td>
</tr>
<tr>
<td>9:30 - 5:00</td>
<td><strong>Workshop 2:</strong> A Short Course in Structural Equation Modeling (Day 1)</td>
<td>Room 1.2</td>
</tr>
<tr>
<td>1:00 - 2:30</td>
<td>(Lunch)</td>
<td></td>
</tr>
<tr>
<td>9:30 - 5:00</td>
<td><strong>Workshop 3:</strong> Statistical Analyses for Language Assessment (Day 1)</td>
<td>Room 1.4</td>
</tr>
<tr>
<td>1:00 - 2:30</td>
<td>(Lunch)</td>
<td></td>
</tr>
<tr>
<td><strong>Friday, June 8:</strong></td>
<td>Pre-Conference Workshops</td>
<td></td>
</tr>
<tr>
<td>9:00 - 9:30</td>
<td>Registration</td>
<td>Outside Paranimf Hall</td>
</tr>
<tr>
<td>9:30 - 4:00</td>
<td><strong>Workshop 2:</strong> A Short Course in Structural Equation Modeling (Day 2)</td>
<td>Room 1.2</td>
</tr>
<tr>
<td>1:00 - 2:30</td>
<td>(Lunch)</td>
<td></td>
</tr>
<tr>
<td>9:30 - 4:00</td>
<td><strong>Workshop 3:</strong> Statistical Analyses for Language Assessment (Day 2)</td>
<td>Room 1.4</td>
</tr>
<tr>
<td>1:00 - 2:30</td>
<td>(Lunch)</td>
<td></td>
</tr>
<tr>
<td>9:30 - 4:00</td>
<td><strong>Workshop 4:</strong> Using Qualitative Research Methods in Language Test</td>
<td>Room 1.1</td>
</tr>
<tr>
<td>1:00 - 2:30</td>
<td>Development and Validation</td>
<td></td>
</tr>
<tr>
<td><strong>Friday, June 8:</strong></td>
<td>The 29th Annual LTRC – Opening Session</td>
<td></td>
</tr>
<tr>
<td>4:00 - 7:00</td>
<td>Registration</td>
<td>Outside Paranimf Hall</td>
</tr>
<tr>
<td>5:00 - 6:05</td>
<td><strong>Welcome</strong></td>
<td>Room: Aula Magna</td>
</tr>
<tr>
<td>5:05 - 7:05</td>
<td><strong>Symposium 1:</strong> The Measurement of Language Aptitude and High Level</td>
<td>Room: Aula Magna</td>
</tr>
<tr>
<td></td>
<td>L2 Abilities</td>
<td></td>
</tr>
<tr>
<td>7:15 - 9:00</td>
<td><strong>Welcome Reception</strong> (Sponsored by University of Cambridge, ESOL)</td>
<td>Garden</td>
</tr>
<tr>
<td></td>
<td>Examinations</td>
<td></td>
</tr>
<tr>
<td><strong>Saturday, June 9:</strong></td>
<td>Book Exhibit</td>
<td></td>
</tr>
<tr>
<td>10:00 - 6:00</td>
<td><strong>BOOK EXHIBIT</strong></td>
<td>Josep Carner Bldg.</td>
</tr>
<tr>
<td>8:00 - 9:00</td>
<td>Registration</td>
<td>Outside Paranimf Hall</td>
</tr>
<tr>
<td>9:00 - 9:15</td>
<td><strong>Welcome to LTRC 2007</strong></td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>Exploring Diverse Methodologies and Conceptualizations in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language Testing Research</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Sponsored by Educational Testing Service)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Toward a Test Theory for the Interactionalist Era, Robert J Mislevy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presentation of the 2007 Messick Award</td>
<td></td>
</tr>
<tr>
<td>10:25 - 10:45</td>
<td><strong>Break</strong></td>
<td>Galeria Paranimf</td>
</tr>
<tr>
<td>10:45 - 12:45</td>
<td><strong>Paper Session 1</strong> (4 papers)</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• Mixed-Methods Designs in Language Testing Research:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Review and Suggestions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Multidimensionality-Based DIF Approaches to L2 Vocabulary:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mixed Methods</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Investigating Perceptions of Language Ability in the CEFR Scales</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Towards a Reproducible and Evaluator-Independent Assessment of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Translation Ability</td>
<td></td>
</tr>
<tr>
<td>12:45 - 12:50</td>
<td><strong>Launch of “Language Testing Reconsidered” (LTRC 2005, Ottawa)</strong></td>
<td></td>
</tr>
<tr>
<td>12:45 - 2:30</td>
<td><strong>Lunch</strong></td>
<td>Room: 1.6</td>
</tr>
</tbody>
</table>

**Note:** Times are in 24-hour format. All locations are within the Paranimf Hall unless otherwise specified.
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:30 - 4:30</td>
<td><strong>Symposium 2</strong>: Current Perspectives on Language Assessment for Migration and Citizenship</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td>4:30 - 4:50</td>
<td>Break</td>
<td>Galeria Paranimf</td>
</tr>
<tr>
<td>4:50 - 6:20</td>
<td><strong>Paper Session 2</strong> (3 papers)</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• Reconsidering the Measurement of Pragmatic Knowledge Using a Reciprocal Task Format</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Validating TOEFL iBT Speaking and Setting Score Requirements for ITA Screening</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• The Effect of Rater Background on the Evaluation of Speech Samples</td>
<td></td>
</tr>
<tr>
<td>6:30 - 7:30</td>
<td><strong>Works in Progress</strong> (9 WIPs)</td>
<td>Room: Vestibulo Principal</td>
</tr>
<tr>
<td></td>
<td>• Impact of Varieties of English on Rater Judgment in Oral Tests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• The Effect of Accent on an EAP Listening Test: A DIF Perspective</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Applying Survey Results to an Online Tutorial for Language Test Selection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Investigating Task Difficulty in ESL Reading Assessment: Combining Research Approaches</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• When There is a Will, There is a Way</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Developing a Diagnostic Assessment of Academic Writing: The Construct Definition Phase</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Bias and the Native Speaker</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Testing Receptive Vocabulary Size: X_Lex and Y_Lex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• New Methodologies and Concepts in Language Testing</td>
<td></td>
</tr>
<tr>
<td>7:45 -</td>
<td>Wine-tasting</td>
<td></td>
</tr>
<tr>
<td></td>
<td>With support from Educational Testing Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Sunday, June 10</strong></td>
<td></td>
</tr>
<tr>
<td>10:00 - 6:00</td>
<td>BOOK EXHIBIT</td>
<td>Josep Carner Bldg.</td>
</tr>
<tr>
<td>8:30 - 9:15</td>
<td>Registration</td>
<td>Outside Paranimf Hall</td>
</tr>
<tr>
<td>9:15 - 11:15</td>
<td><strong>Symposium 3</strong>: Cognitive Diagnosis Approaches in Language Assessment</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td>11:15 - 11:30</td>
<td>Group Photograph</td>
<td>Garden</td>
</tr>
<tr>
<td>11:30 - 11:45</td>
<td>Break</td>
<td>Galeria Paranimf</td>
</tr>
<tr>
<td>11:45 - 1:15</td>
<td><strong>Paper Session 3</strong> (3 papers)</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• (Where) Is the Construct in an Interpretive Argument?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• How the Use of the Visual Channel Affects L2 Listening Test-Taker Performance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Construct Validation of Bachman and Palmer’s (1996) Strategic Competence Model over Time in EFL Reading Tests</td>
<td></td>
</tr>
<tr>
<td>1:15 - 3:00</td>
<td>Lunch</td>
<td>Room 1.6</td>
</tr>
<tr>
<td></td>
<td>Language Assessment Quarterly EAB Meeting (for LAQ EAB Members)</td>
<td></td>
</tr>
<tr>
<td>3:00 - 4:30</td>
<td><strong>Paper Session 4</strong> (3 papers)</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• The Value of On-Line Training for Essay Marking</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Validation of Automated Essay Scoring for ESL Writers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Investigating the Construct Validity of the Community English Program (CEP) Writing Test</td>
<td></td>
</tr>
<tr>
<td>4:30 - 4:50</td>
<td>Break</td>
<td>Galeria Paranimf</td>
</tr>
<tr>
<td>4:50 - 6:20</td>
<td><strong>Paper Session 5</strong> (3 papers)</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• Patterns of Test-Taker Interaction in the First Certificate in English</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Facets of Speaking Proficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• The Use of Contextualization Cues in a Semi-direct Speaking Test</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td>Location</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>6:30 - 7:30</td>
<td><strong>Poster Displays</strong> (16 posters)</td>
<td>Galeria Paranimf</td>
</tr>
<tr>
<td></td>
<td>• Revising the Writing Scale of a Large-Scale EFL Examination</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• The LPATE: Revising a High Stakes Assessment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• A Speaking Scale Revision Project—Anticipated and Unanticipated Ramifications</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• The Differential Effects of Time on Accommodated vs. Standard Versions for ELLs: A Meta-Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Test Generation and Online Delivery System (TGOD)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• The Development and Maintenance of a Test of English for Language Teachers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Measuring Grammatical Knowledge with Limited-production Tasks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Validation of an Expanded Definition of the Grammar Construct</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Assessing Writing Holistically and Analytically to Investigate an Early Foreign Language Start</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Health Care Access for Linguistic Minorities: Identifying Nurse Speech Acts for Assessment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Validating the Catalan and Spanish Versions of the MLAT-E in Catalonia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Construct and Task Considerations in Assessing Young EFL Learners</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Validation of the English Diagnostic Grammar Test: Taking Quantitative and Qualitative Approaches</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Candidate Awareness of Performance Features of Spoken Interaction in Peer-peer Test Performance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Features of English L2 Children’s Oral Narrative Proficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• The Certificado Internacional del Español (CIE). A Hispanic Certification Project</td>
<td></td>
</tr>
<tr>
<td>7:45 - 9:30</td>
<td><strong>Student-led Session</strong></td>
<td>Room: 1.3</td>
</tr>
<tr>
<td>Monday, June 11</td>
<td><strong>BOOK EXHIBIT</strong></td>
<td>Josep Carner Bldg.</td>
</tr>
<tr>
<td>10:00 - 1:30</td>
<td><strong>Registration</strong></td>
<td>Outside Paranimf Hall</td>
</tr>
<tr>
<td>9:15 - 9:50</td>
<td><strong>UCLES/ILTA Lifetime Achievement Award Lecture</strong> (Sponsored by Cambridge University, ESOL Examinations and ILTA): Where We Have Been And Where We Should Go</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>Charles W. “Charlie” Stansfield</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presentation of UCLES/ILTA Lifetime Achievement Award</td>
<td></td>
</tr>
<tr>
<td>9:50 - 11:20</td>
<td><strong>Paper Session 6</strong> (3 papers)</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• Writer Reader Relationships in Undergraduate Textbooks and in Academic Reading Tests</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Effects of Thinking Aloud on ESL Essay Rater Performance: A FACETS Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Effects of Source Texts on Summarization Performance</td>
<td></td>
</tr>
<tr>
<td>11:20 - 11:40</td>
<td><strong>Break</strong></td>
<td>Galeria Paranimf</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td>Location</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>11:40 - 12:40</td>
<td><strong>Paper Session 7</strong> (3 papers)</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• Capturing Accuracy in Second Language Performance: Exploring Global and Local Measures</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Formulaic Sequences and Perceived Oral Proficiency: Putting a Lexical Approach to the Test</td>
<td></td>
</tr>
<tr>
<td>12:40 - 3:10</td>
<td><strong>Lunch</strong></td>
<td>Room: 1.6</td>
</tr>
<tr>
<td></td>
<td>ILTA Annual Business Meeting (for ILTA Members)</td>
<td></td>
</tr>
<tr>
<td>3:10 - 5:10</td>
<td><strong>Paper Session 8</strong> (4 papers)</td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• E-portfolio Assessment and Individual Item Selection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Looking at Assessment and Feedback Practices from Multiple Perspectives</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Using Activity Theory to Investigate Constructs of Competence in the MFL Classroom</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Assessment for Literacy Education in Secondary Schools: A Comparison of 12 Cities</td>
<td></td>
</tr>
<tr>
<td>5:10 - 5:30</td>
<td><strong>Break</strong></td>
<td>Galeria Paranimf</td>
</tr>
<tr>
<td>5:30 - 7:00</td>
<td><strong>Paper Session 9</strong></td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td></td>
<td>• Discourse Context, Complexity, and Task Variation in the ESP Assessment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• English for Occupational and Academic Purposes: Same or Different?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Testing Academic Language or Academic Content? An Exploration Using CFA</td>
<td></td>
</tr>
<tr>
<td>7:00 - 7:15</td>
<td><strong>Closing Remarks</strong></td>
<td>Room: Paranimf</td>
</tr>
<tr>
<td>9:00</td>
<td><strong>Banquet &amp; Awards Presentations</strong></td>
<td>Can Cortada Restaurant</td>
</tr>
</tbody>
</table>
## LTRC 2007 Full Program

### Thursday, June 7

#### Pre-Conference Workshops

<table>
<thead>
<tr>
<th>Registration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location: Outside Paranimf Hall</td>
</tr>
<tr>
<td>9:00 - 9:30</td>
</tr>
</tbody>
</table>

**Workshop 1: Language Testing: The Social Dimension**  
*Tim McNamara and Carsten Roever, The University of Melbourne*  

<table>
<thead>
<tr>
<th>Location: Room 1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30 - 5:00 (Lunch: 1:00 to 2:30)</td>
</tr>
</tbody>
</table>

**Workshop 2: A Short Course in Structural Equation Modeling** *(Day 1)*  
*Barbara M. Byrne, University of Ottawa and Ardeshir Geranpayeh, University of Cambridge ESOL Examinations*  

<table>
<thead>
<tr>
<th>Location: Room 1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30 - 5:00 (Lunch: 1:00 to 2:30)</td>
</tr>
</tbody>
</table>

**Workshop 3: Statistical Analyses for Language Assessment** *(Day 1)*  
*Lyle F. Bachman, University of California, Los Angeles and Antony John Kunnan, California State University, Los Angeles*  

<table>
<thead>
<tr>
<th>Location: Room 1.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30 - 5:00 (Lunch: 1:00 to 2:30)</td>
</tr>
</tbody>
</table>

### Friday, June 8:

#### Pre-Conference Workshops

<table>
<thead>
<tr>
<th>Workshop 2: A Short Course in Structural Equation Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Barbara M. Byrne, University of Ottawa and Ardeshir Geranpayeh, University of Cambridge ESOL Examinations</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location: Room 1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30 - 4:00 (Lunch: 1:00 to 2:30)</td>
</tr>
</tbody>
</table>
Workshop 3: Statistical Analyses for Language Assessment (Day 2)
Lyle F. Bachman, University of California, Los Angeles and Antony John Kunnan, California State University, Los Angeles

Location: Room 1.4
9:30 - 4:00 (Lunch: 1:00 to 2:30)

Workshop 4: Using Qualitative Research Methods in Language Test Development and Validation
Lynda Taylor, University of Cambridge ESOL Examinations

Location: Room 1.1
9:30 - 4:00 (Lunch: 1:00 to 2:30)

Friday, June 8:
Registration
Location: Outside Paranimf Hall
4:00 - 7:00

The 29th Annual LTRC—Opening Session
Location: Aula Magna
5:00 - 7:05

5:00-5:05 Welcome

Symposium 1: The Measurement of Language Aptitude and High-level L2 Abilities
Organizer: Mike Long  Chair: Lorena Llosa

Location: Aula Magna
5:05 - 7:05

Issues in Developing New Aptitude Measures
Catherine J. Doughty, University of Maryland, Center for Advanced Study of Language

Issues in the Diagnostic Assessment of Advanced Abilities in Less Commonly Taught Languages
Michael H. Long and Kira Gor, University of Maryland, School of Languages, Literatures, and Cultures

Welcome Reception (Sponsored by University of Cambridge, ESOL Examinations)
Location: Garden
7:15 - 9:00
Saturday, June 9

Book Exhibit

Location: Josep Carner Bldg.
10:00 - 6:00

Registration

Location: Outside Paranimf Hall
8:00 - 9:00

Welcome to LTRC 2007

Location: Paranimf
9:00 - 10:25

9:00 - 9:05 Welcome to LTRC 2007

9:05 - 9:15 Exploring Diverse Methodologies and Conceptualizations in Language Testing Research
James E. Purpura, Teachers College, Columbia University and Carolyn E. Turner, McGill University

Lyle F. Bachman, University of California, Los Angeles

Robert J. Mislevy, University of Maryland

Presentation of the 2007 Samuel J. Messick Memorial Lecture Award by ETS

10:25 - 10:45 Break — Galeria Paranimf

Paper Session 1:
Geoff Brindley (Chair)

Location: Paranimf
10:45 - 12:45

10:45 - 11:15 Mixed-Methods Designs in Language Testing Research: Review and Suggestions
Jiyoung Kim, University of Illinois at Urbana-Champaign

11:15 - 11:45 Multidimensionality-Based DIF Approaches to L2 Vocabulary: Mixed Methods
Eunice Eunhee Jang, Ontario Institute for Studies in Education of the University of Toronto
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:45 - 12:15</td>
<td>Investigating Perceptions of Language Ability in the CEFR Scales&lt;br&gt;&lt;i&gt;Spiros Papageorgiou, Lancaster University&lt;/i&gt;</td>
</tr>
<tr>
<td>12:15 - 12:45</td>
<td>Towards a Reproducible and Evaluator-Independent Assessment of Translation Ability&lt;br&gt;&lt;i&gt;Winibert Segers, Anvers Katholieke Universiteit Leuven, Philippe Anckaert, Haute École, Francisco Ferrer Bruxelles, and June Eyckmans, Erasmushogeschool Brussel &amp; Vrije, Universiteit Brussel&lt;/i&gt;</td>
</tr>
<tr>
<td>12:45 - 12:50</td>
<td>Launch of &quot;Language Testing Reconsidered&quot; (LTRC 2005, Ottawa)</td>
</tr>
<tr>
<td>12:50 - 2:30</td>
<td>Lunch&lt;br&gt;Language Testing Editorial Advisory Board Meeting&lt;br&gt;(for LT EAB Members)&lt;br&gt;(Room: 1.6)</td>
</tr>
</tbody>
</table>

**Symposium 2: Current Perspectives on Language Assessment for Migration and Citizenship**

Organizer: Nick Saville  
Chair: Jee Wha Dakin  
Location: Paranimf  
2:30 - 4:30

**Introduction: Overview and Key Issues**

<i>Nick Saville, University of Cambridge, ESOL Examinations</i>

**Ethicality and Language Testing For Citizenship**

<i>Elana Shohamy, Tel Aviv University</i>

**Which Level for Integration?**

<i>Piet Van Avermaet, Centre for Intercultural Education, University of Ghent</i>

**Testing for Access: Comparing the Cases of Belgium and the Netherlands**

<i>Sara Gysen, Catholic University of Leuven and Henk Kuijper, Cito</i>

**The U.S. Naturalization Test Redesign: Being Set up to Fail**

<i>Antony J. Kunnan, California State University, Los Angeles</i>

**Discussants:** Antony J. Kunnan & Nick Saville

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:30 - 4:50</td>
<td>Break—Galeria Paranimf</td>
</tr>
</tbody>
</table>

**Paper Session 2:**

<i>David Qian (Chair)</i>

Location: Paranimf  
4:50 - 6:20

**4:50 - 5:20**  
Reconsidering the Measurement of Pragmatic Knowledge Using a Reciprocal Task Format  
<i>Kirby Grabowski, Teachers College, Columbia University</i>
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:20 - 5:50</td>
<td>Validating TOEFL iBT Speaking and Setting Score Requirements for ITA Screening</td>
</tr>
<tr>
<td></td>
<td><strong>Xiaoming Xi, Educational Testing Service</strong></td>
</tr>
<tr>
<td>5:50 - 6:20</td>
<td>The Effect of Rater Background on the Evaluation of Speech Samples</td>
</tr>
<tr>
<td></td>
<td><strong>Paula Winke, Susan Gass, and Dan Reed, Michigan State University</strong></td>
</tr>
</tbody>
</table>

### Works in Progress

**Location:** Vestibulo Principal  
**Time:** 6:30 - 7:30

<table>
<thead>
<tr>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact of Varieties of English on Rater Judgment in Oral Tests</td>
</tr>
<tr>
<td><strong>Huei-Lien Hsu, Carsten Wilmès, Kadeessa Kadir, and So-young Jang, University of Illinois at Urbana-Champaign</strong></td>
</tr>
<tr>
<td>The Effect of Accent on an EAP Listening Test: A DIF Perspective</td>
</tr>
<tr>
<td><strong>Luke Harding, University of Melbourne</strong></td>
</tr>
<tr>
<td>Applying Survey Results to an Online Tutorial for Language Test Selection</td>
</tr>
<tr>
<td><strong>Margaret Malone, Megan Montee, and Colleen Gallagher, Center for Applied Linguistics</strong></td>
</tr>
<tr>
<td>Investigating Task Difficulty in ESL Reading Assessment: Combining Research Approaches</td>
</tr>
<tr>
<td><strong>Geoff Brindley, Stephen Moore and Helen Slatyer, Macquarie University</strong></td>
</tr>
<tr>
<td>When There is a Will, There is a Way</td>
</tr>
<tr>
<td><strong>Neus Figueras, Department of Education, Catalonia, Spain and Kaftandjieva Felianka, University of Sofia, Bulgaria</strong></td>
</tr>
<tr>
<td>Developing a Diagnostic Assessment of Academic Writing: The Construct Definition Phase</td>
</tr>
<tr>
<td><strong>Lorena Llosa and Sarah W. Beck, New York University</strong></td>
</tr>
<tr>
<td>Bias and the Native Speaker</td>
</tr>
<tr>
<td><strong>Alan Davies, University of Edinburgh</strong></td>
</tr>
<tr>
<td>Testing Receptive Vocabulary Size: X_Lex and Y_Lex</td>
</tr>
<tr>
<td><strong>Imma Miralpeix, Universitat de Barcelona</strong></td>
</tr>
<tr>
<td>New Methodologies and Concepts in Language Testing</td>
</tr>
<tr>
<td><strong>John H.A.L. de Jong, Pearson Language Assessments</strong></td>
</tr>
</tbody>
</table>

**7:45 -** Wine-tasting *(With support from ETS)*
## Sunday, June 10

### Book Exhibit

**Location:** Josep Carner Bldg.  
**Time:** 10:00 - 6:00

### Registration

**Location:** Outside Paranimf Hall  
**Time:** 8:30 - 9:15

### Symposium 3: Cognitive Diagnosis Approaches in Language Assessment

*Organizer: Yong-Won Lee  Chair: Kristin DeGennaro*

**Location:** Paranimf  
**Time:** 9:15 - 11:15

<table>
<thead>
<tr>
<th>Title</th>
<th>Presenter</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Diagnosis in Language Assessment: An Overview</td>
<td>Yong-Won Lee</td>
<td>Seoul National University</td>
</tr>
<tr>
<td>Q-matrix Construction: Defining the Link between Constructs and Test Items in Cognitive Diagnosis Approaches</td>
<td>Hae-Jin Kim</td>
<td>Educational Testing Service</td>
</tr>
<tr>
<td>Application of Three Cognitive Diagnosis Models in ESL Reading and Listening Tests</td>
<td>Yasuyo Sawaki</td>
<td>Educational Testing Service</td>
</tr>
<tr>
<td>Cognitive Diagnostic Approaches to L2 Reading: Dialectical Tracking between Psychometric- and Classroom-validity Evidence</td>
<td>Eunice Eunhee Jang</td>
<td>OISE, University of Toronto</td>
</tr>
</tbody>
</table>

**Discussant:** Antony J. Kunnan, California State University, Los Angeles

### Group Photograph

**Location:** Garden  
**Time:** 11:15 - 11:30

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:30 - 11:45</td>
<td>Break—Galeria Paranimf</td>
</tr>
</tbody>
</table>
Paper Session 3:
Robert Edwards (Chair)

Location: Paranimf
11:45 - 1:15

11:45 - 12:15 (Where) Is the Construct in an Interpretive Argument?
Carol Chapelle, Iowa State University, Mary K. Enright, Educational Testing Service, and Joan Jamieson, Northern Arizona University

12:15 - 12:45 How the Use of the Visual Channel Affects L2 Listening Test-Taker Performance
Elvis Wagner, Teachers College, Columbia University

12:45 - 1:15 Construct Validation of Bachman and Palmer’s (1996) Strategic Competence Model Over Time in EFL Reading Tests
Aek Phakiti, University of Sydney

1:15 - 3:00 Lunch
Language Assessment Quarterly Editorial Advisory Board Meeting (for LAQ EAB Members) (Room: 1.6)

Paper Session 4:
Elvira Swender (Chair)

Location: Paranimf
3:00 - 430

3:00 - 3:30 The Value of On-Line Training for Essay Marking
Paul Jaquith, Annie Brown, and Karen Caldwell, Ministry of Higher Education, UAE

3:30 - 4:00 Validation of Automated Essay Scoring for ESL Writers
Sara Cushing Weigle, Yanbin Lu, and Amanda Baker, Georgia State University

4:00 - 4:30 Investigating the Construct Validity of the Community English Program (CEP) Writing Test
Taejoon Park, Teachers College, Columbia University

4:30 - 4:50 Break — Galeria Paranimf

Paper Session 5:
Jayanti Banerjee (Chair)

Location: Paranimf
4:50 - 6:20

4:50 - 5:20 Patterns of Test-taker Interaction in the First Certificate in English
Evelina Galaczi, University of Cambridge ESOL Examinations
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 5:20 - 5:50| Facets of Speaking Proficiency  
*Rob Schoonen, Nivja de Jong, Margarita Steinel, Arjen Florijn, and Jan Hulstijn, University of Amsterdam* |
| 5:50 - 6:20| The Use of Contextualization Cues in a Semi-direct Speaking Test  
*Hyunjoo Kim, Teachers College, Columbia University* |

**Poster Displays**

**Location:** Galeria Paranimf  
**6:30 - 7:30**

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revising the Writing Scale of a Large-Scale EFL Examination</td>
<td><em>Fernando Fleurquin, Aaron Ohlrogge, and Barbara Dobson, University of Michigan ELI</em></td>
</tr>
<tr>
<td>The LPATE: Revising a High Stakes Assessment</td>
<td><em>Alan Urmston, Hong Kong Polytechnic University, Tom Lunley, Australian Council for Education Research, and Mark Knight, Hong Kong Polytechnic University</em></td>
</tr>
<tr>
<td>A Speaking Scale Revision Project—Anticipated and Unanticipated Ramifications</td>
<td><em>Sarah Briggs, Melinda Matice, and Fernando Fleurquin, University of Michigan</em></td>
</tr>
<tr>
<td>The Differential Effects of Time on Accommodated vs. Standard Versions for ELLs: A Meta-Analysis</td>
<td><em>Maria Pennock-Roman, MPR Psychometric Research &amp; Consulting and Charlene Rivera, The George Washington University Center for Equity &amp; Excellence in Education</em></td>
</tr>
<tr>
<td>Test Generation and Online Delivery System (TGOD)</td>
<td><em>Gary Buck, Pacia Platzek, and Forest Platzek, Lidget Green Inc.</em></td>
</tr>
<tr>
<td>The Development and Maintenance of a Test of English for Language Teachers</td>
<td><em>Robert Edwards, Université de Sherbrooke</em></td>
</tr>
<tr>
<td>Measuring Grammatical Knowledge with Limited-production Tasks</td>
<td><em>Jee Wha Dakin and Hyun Jung Kim, Teachers College, Columbia University</em></td>
</tr>
<tr>
<td>Validation of an Expanded Definition of the Grammar Construct</td>
<td><em>India C. Plough and Jeff S. Johnson, University of Michigan</em></td>
</tr>
<tr>
<td>Assessing Writing Holistically and Analytically to Investigate an Early Foreign Language Start</td>
<td><em>Teresa Naves, Universitat de Barcelona</em></td>
</tr>
</tbody>
</table>
Health Care Access for Linguistic Minorities: Identifying Nurse Speech Acts for Assessment
Carolyn E. Turner, McGill University, Michel D. Laurier, Université de Montréal, and Xuelian Wei, McGill University

Validating the Catalan and Spanish Versions of the MLAT-E in Catalonia
Maria del Mar Suarez Vilagran, Universitat de Barcelona and Charles Stansfield, Second Language Testing Foundation

Construct and Task Considerations in Assessing Young EFL Learners
Elana Shohamy, Tel Aviv University and Ofra Inbar, Beit Berl College, Tel Aviv University

Validation of the English Diagnostic Grammar Test: Taking Quantitative and Qualitative Approaches
Rie Koizumi, Tokiwa University, Hideki Sakai, Shinshu University, Takahiro Ido, Waseda Junior & Senior High School, Hiroshi Ota, Komazawa Women’s University, Megumi Kimura, Dokkyo University, Masatoshi Sato, Ichikawa High School, and Akiko Nemoto, Atago Junior High School

Candidate Awareness of Performance Features of Spoken Interaction in Peer-peer Test Performance
Anna Maria Ducasse, La Trobe University

Features of English L2 Children’s Oral Narrative Proficiency
Sari Luoma, Ballard & Tighe, Publishers

The Certificado Internacional del Español (CIE):
A Hispanic Certification Project
Martha Jurado, Universidad Nacional Autónoma de México and Ramón Parrondo, Instituto Cervantes

Student-led Session
Organized by Jee Wha Dakin and Kristen Di Gennaro, Teachers College, Columbia University

Location: Room 1.3
7:45 - 9:30

Monday, June 11

Book Exhibit
Location: Josep Carner Bldg.
10:00 - 1:30

Registration
Location: Outside Paranimf Hall
8:30 - 9:00
UCLES/ILTA Lifetime Achievement Award Lecture
Sponsored by Cambridge University, ESOL Examinations and ILTA:

Location: Paranimf
9:00 - 9:50

Introduction of the UCLES/ILTA Lifetime Achievement Award Lecturer
Alan Davies, University of Edinburgh

Where We Have Been and Where We Should Go

Presentation of UCLES/ILTA Lifetime Achievement Award

Paper Session 6:
Ulla Gudnason (Chair)

Location: Paranimf
9:50 - 11:20

9:50 - 10:20     Writer Reader Relationships in Undergraduate Textbooks and in Academic Reading Tests
Cyril Weir, University of Bedfordshire and Sara Gysen, Universiteit Leuven/Centrum voor Taal en Migratie

10:20 - 10:50     Effects of Thinking Aloud on ESL Essay Rater Performance: A FACETS Analysis
Khaled Barkaoui, Ontario Institute for Studies in Education of the University of Toronto

10:50 - 11:20     Effects of Source Texts on Summarization Performance
Guoxing Yu, University of Bristol

11:20 - 11:40     Break—Galeria Paranimf

Paper Session 7:
Rachel Brooks (Chair)

Location: Paranimf
11:40 - 1:10

11:40 - 12:10     Capturing Accuracy in Second Language Performance: Exploring Global and Local Measures
Gillian Wigglesworth, University of Melbourne and Pauline Foster, St. Mary’s College

12:10 - 12:40     Formulaic Sequences and Perceived Oral Proficiency: Putting a Lexical Approach to the Test
June Eyckmans, Erasmushogeschool Brussel & Vrije Universiteit Brussel, Hélène Stengers, Vrije Universiteit Brussel, and Frank Boers, Erasmushogeschool Brussel & Universiteit Atwerpen
### Paper Session 8:
*Youn-Hee Kim (Chair)*

**Location: Paranimf**

**3:10 - 5:10**

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Presenter(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:10 - 3:40</td>
<td>E-portfolio Assessment and Individual Item Selection</td>
<td>Vivien Berry, The University of Hong Kong</td>
</tr>
<tr>
<td>3:40 - 4:10</td>
<td>Looking at Assessment and Feedback Practices from Multiple Perspectives</td>
<td>Ari Huhta and Mirja Tarnanen, University of Jyväskylä</td>
</tr>
<tr>
<td>4:10 - 4:40</td>
<td>Using Activity Theory to Investigate Constructs of Competence in the MFL Classroom</td>
<td>Kathryn Hill, University of Melbourne</td>
</tr>
<tr>
<td>4:40 - 5:10</td>
<td>Assessment for Literacy Education in Secondary Schools: A Comparison of 12 Cities</td>
<td>Alister Cumming, Ontario Institute for Studies in Education of the University of Toronto</td>
</tr>
</tbody>
</table>

| 5:10 - 5:30 | Break — Galeria Paranimf |

### Paper Session 9:
*Carmen Velasco (Chair)*

**Location: Paranimf**

**5:30 - 7:00**

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Presenter(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:30 - 6:00</td>
<td>Discourse Context, Complexity, and Task Variation in the ESP Assessment</td>
<td>Carol Lynn Moder and Gene B. Halleck, Oklahoma State University</td>
</tr>
<tr>
<td>6:00 - 6:30</td>
<td>English for Occupational and Academic Purposes: Same or Different?</td>
<td>Catherine Elder, University of Melbourne</td>
</tr>
<tr>
<td>6:30 - 7:00</td>
<td>Testing Academic Language or Academic Content? An Exploration Using CFA</td>
<td>Dorry Kenyon and Bokyung Cho, Center for Applied Linguistics</td>
</tr>
</tbody>
</table>

| 7:00 - 7:15 | Closing Remarks |

### Banquet & Awards Presentations

**Location: Can Cortada Restaurant**

**9:00 -**
UCLES/ILTA Lifetime Achievement Award 2007

Awarded to
Dr. Charles W. (Charlie) Stansfield

On behalf of the review committee for the UCLES/ILTA Lifetime Achievement Award (comprised of Glenn Fulcher, Jim Purpura, Nick Saville, and myself—John Clark) it is a great pleasure to announce that after carefully considering the several and highly meritorious nominations received from the field, Dr. Charles W. (Charlie) Stansfield has been selected to receive the Lifetime Achievement Award to be presented in Barcelona, Spain, at the 2007 Language Testing Research Colloquium (LTRC), sponsored by the International Language Testing Association (ILTA).

Dr. Stansfield is currently president of Second Language Testing, Inc. (SLTI), a company which he founded in 1994 and through which he has continued to engage in numerous test development projects and related activities on behalf of the foreign/second language profession. Prior to assuming his SLTI responsibilities, Dr. Stansfield served from 1986-1994 as director of the Foreign Language Education and Testing Division at the Center for Applied Linguistics (CAL), where he played a major role in the development and validation of the Simulated Oral Proficiency Interview (SOPI), a semi-direct measure that has been well accepted and widely used as an alternative to face-to-face speaking proficiency testing.

Dr. Stansfield’s work at CAL was preceded by a five-year tenure as associate director of the Language Programs division at Educational Testing Service, during which time he was intensively involved in the development and introduction of the Test of Spoken English (TSE) and the Test of Written English (TWE) into the expanding TOEFL program. In this connection, Charlie conducted important research studies on test rater performance and on the development of scoring rubrics that could be uniformly applied across diverse speaking and writing topics.

Other positions held by Dr. Stansfield include an associate professorship in Spanish at the University of Colorado; directorship of the Peace Corps training center in Managua,
Nicaragua; and directorship of the University of Colorado's study abroad programs in Xalapa, Mexico and Valencia, Spain.

The totality of Dr. Stansfield's thirty-year career has involved extensive, hands-on experience in virtually every aspect of the foreign/second language testing endeavor. This has ranged from a thorough grounding in classroom testing techniques and associated issues, up to significant development and administrative roles in large-scale, high-stakes testing programs. His operational contributions to these numerous and varied testing activities have been supported and conceptually informed by similarly impressive professional and academic accomplishments, as attested to by some eighty monographs, book chapters, and articles on second/foreign language testing published in major journals in the field.

In addition to all of the preceding, Charlie has consistently and very willingly rendered diligent and collegial service to the advancement of the language testing profession as a whole. His contributions in this regard have included serving on the editorial boards of five major journals (TESOL Quarterly, Modern Language Journal, Journal of Second Language Writing, Language Testing, and Journal of Language for International Business) and on the advisory boards for the National Foreign Language Resource Centers at the University of Hawaii and the University of Minnesota, as well as serving as the director of a similar center at CAL. Most particularly, the birth and formative development of ILTA itself is correctly—and very appropriately—credited to Dr. Stansfield, who not only conceptually spearheaded but also willingly and successfully took on the considerable coordination and practical work needed to bring the organization into existence and to serve as its first president.

In sum, our profession owes a large debt of gratitude to Dr. Charles Stansfield, not only for his past contributions to the field but also for his present and continuing involvement in important language testing activities and associated measurement issues at both the national and international levels.

UCLES/ILTA Lifetime Achievement Award Lecture
Monday, June 11th, 9:15 – 9:50
Where We Have Been and Where We Should Go
Charles W. (Charlie) Stansfield
Second Language Testing, Inc.

Abstract
This presentation will review how the field of language testing has increased in stature and professionalism in the last 20 years. It will also provide some ideas for how we can increase our presence and importance in the future. We are well beyond the days when we were not even recognized as a field of applied linguistics. We are recognized now and gaining in stature. A number of factors are converging to make our perspectives and our work more important each year. But to reach our full potential, we must rise to the occasion. We must offer our services to the broader field of education, to government, and to private industry. Then we must deliver high quality work in a timely way. The opportunity to become involved in policy decisions is also there. But larger policy decisions involve more than a knowledge of language testing. One must demonstrate sensitivity to the political, economic, and contextual factors surrounding tests, test use, and test-related policies. And, one must communicate beyond the language testing community.
2007 Samuel J. Messick Memorial Lecture

Toward a Test Theory for the Interactionist Era
Robert J. Mislevy, University of Maryland

Abstract
An interactionalist perspective on language ability emphasizes the contextualized and social nature of communication. This view conflicts with the segregation between characteristics of persons and characteristics of situations that is seen (perhaps oversimplistically) to underlie language tests cast in trait and behaviorist perspectives. An emerging synthesis of cognitive and sociocultural insights holds promise for a testing approach that both acknowledges interactionalism and provides generalization for claims about persons—in ways, and to extents, that require a theory of context in communication to explicate. Resulting implications for the design and analysis of language tests are considered in this presentation. In particular, I explore the nature of assessment arguments and ways they can be embodied in the artifacts and processes of operational testing—some new, some extensions of current practice, and others familiar but appropriately reconceived. The starting point for these ruminations is Messick’s 1994 bellwether article, “The interplay of evidence and consequences in the validation of performance assessments.”

About Robert J. Mislevy
Robert Mislevy is professor of Measurement and Statistics at the University of Maryland, and was previously Distinguished Research Scientist at ETS. His research applies developments in statistics, technology, and psychology to practical problems in assessment. His work includes the multiple-imputation analysis in the National Assessment of Educational Progress and an evidence-centered assessment design framework. He has served as president of the Psychometric Society and received the NCME award for Career Contributions to Educational Measurement.

Abstracts

Pre-conference Workshops

Workshop 1: Language Testing: The Social Dimension
Tim McNamara and Carsten Roever, University of Melbourne

The workshop examines language testing as a social practice, a perspective which is at odds with the dominant cognitive, individualistic perspective of language testing and measurement more generally. The workshop is in two halves. In the morning, the social dimension of language assessment is considered in terms of current approaches to test validity in the work of Messick, Kane and Bachman, and we explore approaches to the social within the psychometric tradition, especially DIF and test fairness. In the afternoon, we consider tests at the societal level, as used in the management of education systems, and as instruments of social policy, particularly in contested intergroup settings. We discuss language tests and identity, and the role of tests in immigration and citizenship contexts, arguing that we need a richer model of the social context in order to fully understand language testing as a social practice.


About Tim McNamara
Tim McNamara is Professor of Applied Linguistics at the University of Melbourne. His interests are in the social meaning of language testing, Rasch measurement, performance assessment, validity theory and the testing of language for specific purposes. His book with Carsten Roever, Language Testing: The Social Dimension, was published by Blackwell in 2006.

About Carsten Roever
Carsten Roever is a Senior Lecturer in Applied Linguistics at the University of Melbourne. He holds a Ph.D. in Second Language Acquisition from the University of Hawai‘i at Manoa. His research interests are language testing, interlanguage pragmatics, and second language acquisition of less commonly taught languages.

Workshop 2: A Short Course in Structural Equation Modeling
Barbara M. Byrne, University of Ottawa and Ardeshir Geranpayeh, University of Cambridge, ESOL Examinations

SEM Workshop
This 2-day workshop provides an introduction to the basic concepts and applications of structural equation modeling (SEM) using the EQS statistical program.

Day 1 (June 7) focuses on critical aspects of model specification, model testing, and model misspecification related to single-group analyses. Following a review of EQS notation, participants will be “walked through” the input and output files related to both confirmatory factor analytic and path analytic models.

Day 2 (June 8) focuses on multiple-group analyses and examines applications designed to test for the equivalence of measurement and structural parameters across groups, as well as for differences in latent factor means.
Although this workshop is intended as an introduction to SEM, some knowledge of multiple regression and factor analysis is recommended.

About Barbara M. Byrne
Barbara M. Byrne is Professor Emeritus in the School of Psychology at the University of Ottawa, Canada. Substantively, Dr. Byrne's research focuses on construct validation issues related to the structure and measurement of self-concept, burnout, and depression. Methodologically, her research centers on the sound application of structural equation modeling (SEM) in the validation of measuring instruments and psychological constructs. She has conducted over 60 workshops related to the application of SEM and is the author of five popular introductory books on the topic, the most recent of these being *Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming* (Second Edition; in press). In addition to the publication of over 80 book chapters and scholarly journal articles, most of which have addressed SEM application issues, Dr Byrne is the author of an important reference book related to self-concept measurement, *Measuring Self-concept Across the Lifespan: Issues and Instrumentation* (1996). She is the recipient of three Distinguished Teaching Awards presented by the Canadian Psychological Association (1995), the American Psychological Association (APA; 2002), and the APA, Division 5 (Jacob Cohen Award; 2002).

About Ardeshir Geranpayeh
Ardeshir holds an MSc and PhD in Applied Linguistics from the University of Edinburgh (UK). His role as a senior coordinator involves standardising operational validation tasks, developing quality-related IT systems and supporting quantitative analysis of the examinations. Current interests include: IRT and item banking, differential item functioning, and construct investigation using structural equation modeling.

Workshop 3: Statistical Analyses for Language Assessment

*Lyle F. Bachman, University of California, Los Angeles and Antony John Kunnan, California State University, Los Angeles*

This workshop will introduce participants to basic and intermediate statistical analyses relevant for language assessment professionals. The workshop is aimed at the following audiences:
- language teachers from school, college and university levels
- government officers involved in the development of language tests
- students in linguistics, applied linguistics and allied fields

Participants should purchase the books listed below and preferably read the chapters and work through the exercises in the Workbook prior to attending the workshop. This will make it easier to understand the concepts and procedures that will be introduced in the workshop. In addition, participants should bring along a laptop computer with SPSS (version 11 or above) (also available for free two-week download at www.spss.com) and any language test data that they might want to analyze as part of the workshop. Some time will be set aside on both days for discussion and help with individual data sets. Participants may choose to attend one or both days.

**Workshop topics:**
**Day 1 (June 7):** Basic concepts and descriptive statistics (Bachman Chapters 1, 2, 3, 7, 8)
No prior knowledge of statistics required.
Day 2 (June 8): Test analysis (Bachman Chapters 4, 5, 6, 9, 10)
Prerequisite: Day 1 workshop or prior knowledge of descriptive statistics, correlation, hypothesis testing and tests for statistical significance (t, z, F)


Laptop:
PC or Mac computer with CD player and SPSS with version 11 or above.

About Lyle F. Bachman
Lyle Bachman has published numerous articles and books in language assessment and teaches courses and conducts practitioner training workshops in language assessment. His current research interests include validation theory, issues in assessing the academic achievement and academic English of ELLs in schools, and the interface between language testing research and second language acquisition research.

About Antony John Kunnan
Antony John Kunnan is the author of many journal articles, book chapters, edited and authored books. He has also conducted workshops on many topics in language testing in Argentina, Bulgaria, Egypt, India, Taiwan, UK, and USA and presented conference papers in an additional 20 countries. He is currently the editor of Language Assessment Quarterly.

Workshop 4: Using Qualitative Research Methods in Language Test Development and Validation
Lynda Taylor, University of Cambridge ESOL Examinations

This workshop will introduce the underlying premises of qualitative research, with a focus on qualitative research used in the context of language testing and assessment. We shall begin by overviewing briefly some of the noteworthy research in this field that has made use of qualitative research methods, and then go on to examine the application of qualitative methodologies in relation to the testing of speaking and writing, where such methods have proved particularly fruitful over the past 10-15 years.

Through discussion and worked practice activities, we shall explore the following topics:
• Using qualitative analysis to gain insights into Test-taker behaviour
  • Investigating the PROCESS of speaking and writing assessment
  • Investigating the PRODUCT of speaking and writing assessment
• Using qualitative analysis to gain insights into Interlocutor and Rater behaviour
  • Investigating Interlocutor and Rater strategies and behaviour in speaking and writing assessment (both PROCESS and PRODUCT)

The workshop sessions are designed to help participants develop their confidence and skills so that they feel better equipped to conduct their own qualitative research studies in the field of language testing and assessment.
About Lynda Taylor
Lynda Taylor holds a Master's and PhD in applied linguistics and language testing from the University of Cambridge (UK). She has extensive experience of the theoretical and practical issues involved in language assessment. As Assistant Director of Research and Validation with Cambridge ESOL she helps coordinate the overall research programme and manages the presentation/publication of research outcomes.

Symposia

Symposium 1: The Measurement of Language Aptitude and High-Level L2 Abilities

Organizer
Michael H. Long, University of Maryland

Presenters
Catherine Doughty, University of Maryland, Center for Advanced Study of Language, Michael H. Long and Kira Gor, University of Maryland, School of Languages, Literatures, and Cultures.

Powerful geopolitical forces have greatly increased the need for professionals with high-level L2 abilities, especially in hitherto less commonly taught languages (LCTLs). New language aptitude measures are required to identify individuals with potential to obtain such advanced abilities, and thereby make expensive training programs more cost-effective. New achievement measures are required to assess proficiency at advanced levels, often in languages for which psychometrically sound assessment tools are lacking, even at elementary levels. New diagnostic measures are required to provide information about the linguistic profiles of typical heritage and non-heritage learners at various proficiency levels (i.e., about what has been mastered or remains problematic at each level).

The symposium will feature presentations of research at the University of Maryland Center for Advanced Study of Language on a new computer-delivered aptitude measure, and of research at UMD's School of Languages, Literatures, and Cultures on the acquisition and testing of four LCTLs, illustrating from our work on Russian.

Issues in developing new aptitude measures
Catherine Doughty, University of Maryland, Center for Advanced Study of Language

The research and development of CASL's aptitude battery for predicting near-native attainment (Hi-LAB) has uncovered a number of important aptitude construct issues. Since Hi-LAB aims to make a prediction different from existing aptitude tests -- that is, who among adult starters can be expected to reach a very high level of ultimate attainment -- we have postulated a number of innovative aptitude constructs (e.g., memory, induction, perceptual acuity, processing speed, pragmatic sensitivity, and fluency). In adapting existing measures or programming new ones, we have encountered at least four issues that must be resolved through empirical research:

1. A new construct is really a system, the components of which need to be defined and measured separately. For example, working memory is a complex system that has at least two higher-order constructs (WM-control, WM-storage); and the higher-order con-
Structs sometimes have more than one primary construct associated with them (WM-control: task-switching and control-updating/inhibition; WM-storage: active and passive).

2. A primary construct may be a component of more than one measure. For example, perceptual acuity is one component of Hi-LAB’s processing speed measure (the time it takes to perceive a stimulus) and the only component of Hi-LAB’s phonemic discrimination measure.

3. As in other tests, constructs may overlap. Ideally, we want Hi-LAB’s constructs to be separate.

4. Some measures could be considered either proficiency measures (in the sense of processing) or aptitude measures (in the sense of an inherent talent). Furthermore, it has been proposed that aptitude is really a set of complexes (processing + talent).

These four issues will be exemplified through demonstration of the relevant Hi-LAB components, and implications for measurement will be discussed.

**Issues in the diagnostic assessment of advanced abilities in less commonly taught languages**
Michael H. Long and Kira Gor, University of Maryland, School of Languages, Literatures, and Cultures

The Linguistic Correlates of Proficiency (LCP) project at the University of Maryland seeks to provide descriptions of the linguistic abilities and difficulties at very advanced proficiency levels of English-speaking heritage and non-heritage learners of four LCTLs: Arabic, Korean, Persian, and Russian. Inventories of linguistic features and listening and speaking abilities of typical learners at levels 2-4 on the ILR scale will provide useful information for syllabus designers, materials writers, and classroom teachers. Diagnostic tests focusing on those features and abilities will be useful for testers and for identifying the linguistic abilities and remaining needs of individual learners.

The combination of LCTLs, advanced proficiency, and constraints on the form the eventual diagnostic measures may take has presented several problems for the project team. Issues addressed have included the paucity of work on the acquisition or teaching of most LCTLs at any proficiency level, the need for the listening and speaking tests to be computer-deliverable and usable with learners who may have no command of the writing systems of the languages concerned; relationships between control of discrete linguistic features and global speech processing; lexical frequency and structural complexity as variables in item writing; and the desirability of cross-linguistic comparability in test design. Since many tests measure reaction times in addition to accuracy, a calibration on both these behavioural measures for different ILR levels, and in comparison to native speakers’ performance, is one of the goals of the pilot stage of the project.

**Symposium 2: Current Perspectives on Language Assessment for Migration and Citizenship**

Organizer
Nick Saville, University of Cambridge, ESOL Examinations
Demonstrating language proficiency as part of the process of entering a country as a newcomer is not new, but in the last 20 years economic and socio-political developments such as globalisation and trans-nationalism have accelerated the mass movement of people and increased the use of language tests for entry, residency and citizenship purposes.

Important reasons for mobility now include asylum, family reunion, access to education, employment and other business opportunities and the resulting demographic changes have brought a variety of challenges for policy makers and practitioners in different parts of the world. A common response from many governments, however, has been to introduce new policies relating to language requirements in national languages and the use of tests. Test developers are thus faced with the task of producing appropriate tests or of challenging the policies.

**Introduction: Overview and key issues**

*Nick Saville, University of Cambridge, ESOL Examinations*

This brief introduction provides an overview and highlights a number of the major concerns which will be illustrated in the other papers.

**Ethicality and language testing for citizenship**

*Elana Shohamy, Tel Aviv University*

This paper critiques the *language testing for citizenship* policy in terms of its rationale, purposes, consequences and ethicality; it presents the arguments posed by policy makers and demonstrates the problems with each. It then proposes alternative approaches and urges language testers to engage in activism to resist misuses of tests and languages.

**Which level for integration?**

*Piet Van Avermaet, Centre for Intercultural Education, University of Ghent*

In this paper the *ethical aspects* of using language tests as a condition for *integration* are addressed: to what extent is knowing the language of a country a token of integration? how should language be measured and at which level? To determine the proficiency level most countries in Europe now refer to the CEFR and this in itself raises several questions. These questions are addressed through comparison of concrete examples of policies in several European countries.
Testing for access: Comparing the cases of Belgium and the Netherlands
Sara Gysen, Catholic University of Leuven and Henk Kuijper, Cito

The neighbouring countries Belgium and the Netherlands share Dutch as an official language, yet remarkable differences in assessment traditions and policy exist between them. These differences are most obvious with regard to language testing in the area of immigration and citizenship. This paper presents the current language testing policies and describes the differences from historical, political and societal perspectives. The role of tuition in integration and citizenship contexts, as well as the way professional language testers might deal with dilemmas of ethics and good practice is also considered.

The U.S. naturalization test redesign: Being set up to fail
Antony J. Kunnan, California State University, Los Angeles

The U.S. Citizenship History and Civics and English language test for Naturalization has been in use from 1950 in accordance with Section 312 of the Immigration and Naturalization Act. In the past decades, immigrant groups have complained about the fairness of this test and as a result the Naturalization Test Redesign Project was started in 2001 to be implemented in 2008. Will this project address the concerns or does the policy itself need to change?

Symposium 3: Cognitive Diagnosis Approaches in Language Assessment

Organizer:
Yong-Won Lee, Seoul National University

Presenters:
Yong-Won Lee, Seoul National University
Hae-Jin Kim, Educational Testing Service
Yasuyo Sawaki, Educational Testing Service
Eunice Eunhee Jang, OISE, University of Toronto

Discussant:
Antony Kunnan, California State University in Los Angeles

Score users of language assessments (including test-takers, teachers, and admission officers) often request more detailed information that goes beyond what a single section score or a total test score can provide about test-takers’ test performance. Such information can inform course instructors for appropriate intervention and test takers for self-learning. Recently, cognitive diagnosis approaches (CDA) have been used to extract more detailed, fine-grained information about test takers’ test performance including their mastery of various language subskills. Although these approaches have been developed, they are relatively new to language testers and have not been widely researched or discussed in the field of language testing. The primary goal of this symposium is to give information on the state of the art in CDA in language assessments from multiple perspectives: we will give a general overview of history and major components of CDA along with challenges facing CDA in language assessments, describe the skill definition and Q-matrix (a skill-by-item matrix) (Tatsuoka, 1983, 1990) construction processes, present results of comparing the results of CDA analyses based on three different psychometric models of CDA when they are applied to language test data, and finally discuss validity evidence for CDA results in language assessments.
Cognitive diagnosis in language assessment: An overview
Yong-Won Lee, Seoul National University

What is cognitive diagnosis? What steps are involved in the development and evaluation of CDA-based assessments? What psychometric models are available for CDA? How do we interpret the results? What are the implications for language learning and testing? The main goal of this talk is to give a general overview of cognitive diagnosis in language assessments. In this talk, we will first overview history, previous research, and the major components of CDA. Then, we will discuss the general process of Q-matrix construction, characteristics of psychometric models, and various issues of consideration and future challenges facing CDA in language assessments.

Q-matrix construction: Defining the link between constructs and test items in cognitive diagnosis approaches
Hae-Jin Kim, Educational Testing Service

A cognitive diagnosis analysis starts with identifying the target skills of interest and specifying the relationships between the skills and individual test items. Such relationships are summarized in a skill-by-item matrix often referred to as a ‘Q-matrix’. The soundness of this skill definition and item coding process largely determines the quality and interpretability of the CDA results. The main goals of this talk are to give an overview of the steps involved in the skill definition and Q-matrix construction process and discuss various theoretical and practical issues involved by using TOEFL iBT reading and listening sections as a test case.

Application of three cognitive diagnosis models in ESL reading and listening tests
Yasuyo Sawaki, Educational Testing Service

This paper discusses the findings of a research project which compared the functioning of three emerging cognitive diagnosis models (the general diagnostic, fusion, and latent class analysis models) in the context of ESL listening and reading assessments. Test taker item response data from multiple forms of TOEFL iBT were analyzed by using these models. Results of analyses revealed that the three models produced similar results in terms of examinee classification consistency across test forms, psychometric dimensionality, and skill mastery profiles. However, some notable differences were also identified among these models. Implications are discussed for cognitive diagnosis in language assessments.

Cognitive diagnostic approaches to L2 reading: Dialectical tracking between psychometric- and classroom-validity evidence
Eunice Eunhee Jang, OISE, University of Toronto

Despite increasing popularity of cognitive diagnosis approaches in educational assessment, little empirical evidence exists to show whether and how these approaches achieve their intended goals in real learning contexts. The purpose of this presentation is to present validity evidence for and against cognitive diagnosis based on results of the LanguEdgeTM data and classroom observations. Positive evidence included high correct examinee skill classification rates, discriminating power of items, positive correlations between estimated skill mastery and students’ self-assessment, and positive impact on students’ learning. Based on the negative evidence, there will be a discussion on the desirable characteristics of an ideal diagnostic assessment.
Mixed-methods designs in language testing research: Review and suggestions
Jiyoung Kim, University of Illinois at Urbana-Champaign

Influenced by the emergence of the interpretive research paradigm, the communicative approach to language use and ability, and Messick’s unitary view of test validity, qualitative approaches have recently received attention by language testers. However, little discussion has been made about how the newly introduced qualitative approaches can be effectively mixed with predominant quantitative approaches so as to increase our understanding of complex testing practices. Aiming to facilitate our conversation on mixed-methods designs, this paper reviewed previous mixed-methods studies published in the Language Testing journal for last 6 years analyzing three features: 1) types of study, 2) purpose of mixing, and 3) data analysis strategies. Four types of study were found to often use a mixed-method design: 1) washback, 2) validity, 3) classroom assessment, and 4) test-taking process. Even though they are not mutually exclusive, three major purposes of mixing were observed: 1) understand a complex phenomenon, 2) enhance a validity argument, and 3) reflect diverse views. Examples of studies pursuing different purposes were illustrated. For a review of data analysis strategies, Caracelli and Greene’s (1993) framework was used, and three major strategies often employed in the previous studies were 1) data transformation, 2) typology development, and 3) extreme case analysis. In spite of the increased use of mixed-methods designs, the review revealed that language testers often do not use the design in explicit and systematic ways, thus limiting the potential benefits of the design. Based on the review, the paper provides several suggestions for the future use of mixed-methods designs.

Multidimensionality-based DIF approaches to L2 vocabulary: Mixed methods
Eunice Eunhee Jang, Ontario Institute for Studies in Education of the University of Toronto

This paper presents results of a Differential Item Functioning (DIF) study using Shealy and Stout’s (1993) multidimensionality-based DIF analysis framework. In this framework, differences in score distributions across different groups of examinees are a result of multidimensionality in that the secondary dimension rather than the primary dimension differentially affects examinee performance. Thus, this framework requires both statistical and substantive judgments for identifying potential DIF items and substantiating the causes of DIF. It allows researchers to approach DIF as part of comprehensive construct validity argument through mixed-method analyses (Greene & Caracelli, 1997; Tashakkori & Teddlie, 2003).

This paper illustrates step-by-step procedures of multidimensionality-based DIF analyses of Language reading comprehension test data. Qualitative data from think-aloud verbal protocols were used to generate DIF hypotheses about differential functioning of vocabulary items between two groups of Indo-European and non Indo-European L2 learners. Statistical Simultaneous Item Bias Test (SIBTEST) and Crossing SIBTEST (Li & Stout,
1996) were used to test the DIF hypotheses. The DIF results supported the hypotheses by flagging four uniform DIF items and one crossing DIF item. Post-hoc analyses of the DIF-flagged items were performed by revisiting the qualitative verbal data and analyses of cognate types. The results showed that DIF items with large effect sizes were associated with the following: (a) translation-equivalent cognates; (b) word meaning determined independent of context, and (c) less frequent, so more difficult distracter words than requested ones. In light of empirical evidence, the paper discusses implications for test development and validation process.

Session 1, Paper 3: Sat., June 9th, 11:45 - 12:15

**Diverse methodologies in investigating perceptions of language ability in the CEFR scales**  
*Spiros Papageorgiou, Lancaster University*

The draft Manual of the Council of Europe (2003) provides guidance for relating an exam to the Common European Framework of Reference (CEFR). In this context, investigating the understanding of the CEFR scales is crucial, because results are based on judgements by participants who have been trained in using the CEFR. However, research has shown that agreement in expert judgements is hard to achieve (Alderson, 1993).

This paper examines how different levels of language ability in the CEFR descriptors are perceived by the CEFR users in the context of such a linkage process. Two studies were conducted in order to investigate the use of the CEFR scales. In the first study, a verbal protocol analysis (Ericsson & Simon, 1993; Green, 1998) revealed that there are specific ways in which users of the Framework identify the level of the scaled descriptors and the language ability they depict. The second study, part of a larger research project aiming at linking two exams to the CEFR, adopted a quantitative approach. Many–facet Rasch measurement (Linacre, 1989) was employed in order to investigate the scaling of the CEFR descriptors and the effect of training in using the CEFR.

The results of the studies are of particular importance for two main reasons: First, because the understanding of language ability in the scales affects decision-making when building a linkage claim to the CEFR. Second, because CEFR linking claims by exam providers are widespread in Europe and have important consequences on test users.

Session 1, Paper 4: Sat., June 9th, 12:15 - 12:45

**Towards a reproducible and evaluator-independent assessment of translation ability**  
*Winibert Segers, Anvers Katholieke Universiteit Leaven, Philippe Anckaert, Haute École Francisco Ferrer Bruxelles, and June Eyckmans, Erasmushogeschool Brussel & Vrije Universiteit Brussel*

Translation tests have always been a popular format for assessing language proficiency in schools and colleges throughout the world, although some language testers have raised serious objections to this practice (see Klein-Braley 1987 as opposed to Oller 1979). Ironically, translation tests that are so common in testing language proficiency skills are not available as tests of the ability to translate (Lado 1961). Up until recently, no method had been developed to relate performance indicators to the underlying translation competence in a psychometrically controlled way (Anckaert et al. 2006).
Today's practice of evaluating students' translations is still characterized by the use of analytical grids (Waddington 2001). The use of these analytical grids is motivated by the evaluator's wish to take the different dimensions that constitute translation competence into account. Although these instruments are supposed to reduce construct irrelevant variables, they fall short in adequately reducing the subjectivity of the evaluation since they are essentially criterion-related.

In our paper presentation we will put forward a norm-referenced method for evaluating translation ability that is based on the calibration of dichotomous items. The Calibration of Dichotomous Items-method (CDI) is a stable and evaluator-independent evaluation that bridges the gap between language testing theory and the specific epistemological characteristics of translation studies. The different steps that lead to the construction of a standardized test of translation will be illustrated by means of empirically assembled data. This will include pre-testing, establishing item difficulty indices and procedures for verifying the reliability of the translation test.

### Paper Session 2

**Session 2, Paper 1: Sat., June 9th, 4:50 - 5:20**

**Reconsidering the measurement of pragmatic knowledge using a reciprocal task format**

*Kirby Grabowski, Teachers College, Columbia University*

Current influential models of communicative language ability (CLA) (Bachman & Palmer, 1996; Purpura, 2004) represent language knowledge as a set of separate yet related components, including grammatical and pragmatic dimensions, which interact with background knowledge and metacognitive strategies in language use. Although some researchers have attempted to measure aspects of pragmatic knowledge (those following Hudson, Detmer, & Brown's 1995 framework; Roever, 2006), the vast majority have not incorporated a clearly articulated pragmatic component in the test construct. The purpose of this study is to address the shortcomings of prior pragmatics assessment research, namely, the issues of construct underrepresentation, the lack of highly contextually constrained tasks, and the use of less sophisticated statistical tools to support claims of validity.

This paper examines the construct validity of a test based on Purpura's (2004) theoretical model, which specifies the grammatical, semantic, and pragmatic (sociolinguistic, sociocultural, and psychological) components of language knowledge. This model accounts for the way in which grammatical resources are used to convey a range of pragmatic meanings in language use.

Six reciprocal tasks were used to elicit conversations from 100 participants. The linguistic behaviors elicited from the conversations were first analyzed using interactional sociolinguistic methods (Gumperz, 1982) in order to inform rubric construction. The conversations were then scored by two independent judges and analyzed using a variety of statistical methods. The analysis explored a broader range of features of pragmatic knowledge than had previously been investigated.
Session 2, Paper 2: Sat., June 9th, 5:20 - 5:50

Validating TOEFL iBT speaking and setting score requirements for ITA screening

Xiaoming Xi, Educational Testing Service

Although the primary use of TOEFL iBT Speaking scores is to inform admissions decisions at English medium universities, it may also be useful as a pre-arrival screening measure for international teaching assistants (ITAs). This study provides criterion-related validity evidence for TOEFL iBT Speaking scores for ITA screening and evaluates the effectiveness of using the scores for teaching assistantship (TA) assignment classification.

Local ITA screening tests or instructor recommendations were used as the criterion measures at four participating schools. The relationships between TOEFL iBT Speaking and the local ITA tests ranged from moderately strong to strong, supporting the use of the test for ITA screening. However, the strengths of the relationships were somewhat different across schools depending on the extent to which the local test engages and evaluates non-language abilities. Implications of this finding will be discussed.

Binary and ordinal logistic regression analyses were used to investigate how effectively TOEFL iBT Speaking scores separated students into distinct TA assignment categories. At all four schools, the TOEFL iBT Speaking scores were significantly related to students’ TA assignments and were fairly accurate in predicting TA assignments. ROC curves were used to determine the optimal TOEFL iBT Speaking cut scores for TA assignments at each school that would minimize classifying non-passes as passes.

This study provides a clear illustration of setting cut scores based on candidates’ performances on TOEFL iBT Speaking and the criterion measures. The results also have considerable practical value in guiding the use of the TOEFL iBT Speaking scores for ITA screening.

Session 2, Paper 3: Sat., June 9th, 5:50 - 6:20

The effect of rater background on the evaluation of speech samples

Paula Winke, Susan Gass, and Dan Reed, Michigan State University

Previous research has found rater characteristics, such as prior English teaching experience or familiarity with test takers’ ethnic accents, affect the scoring of performance-based tests (Barnwell, 1989; Chalhoub-Deville, 1995; Fayer & Krasinski, 1987; Hill, 1996; Major, Fitzmaurice, Bunta, & Balasubramanian, 2005; Ross, 1979; Santos, 1988). Individualized or group training has been shown to reduce rater bias (Weigle, 1994, 1998; Elder, Knoch, Barkhuizen, & von Randow, 2005); however, this needs more exploration. The present study, funded by a grant from ETS, investigates what raters, after training, attend to when evaluating TOEFL iBT speech samples.

After a 4-hour, online training session, 100 raters were randomly assigned to score one of six subsets of 432 speech samples from the TOEFL iBT. Using Multifaceted RASCH analysis, raters’ background characteristics (their language learning experiences and prior ESL/EFL teaching experiences) were compared to the scores they gave to the test takers’ speech samples, pre-categorized by L1 (Korean, Chinese, or Spanish). Stimulated recalls were performed with 25 raters; they were video taped while rating and subsequently interviewed,
with the videos taken during the ratings used as prompts to stimulate reports of thought processes during the rating. The interviews were tape recorded, transcribed, and coded. Results found an interesting and complex relationship between background characteristics and assigned ratings. Discussed are ways raters viewed their own personal biases in relation to the training they received, the scoring processes they had, and the rubrics they used. Discussed is how training could be amended to further limit subjectivity in scoring.

Paper Session 3

Session 3, Paper 1: Sun., June 10, 11:45 - 12:15

(Where) Is the construct in an interpretive argument?
Carol Chapelle, Iowa State University, Mary K. Enright, Educational Testing Service and Joan Jamieson, Northern Arizona University

Guidelines such as the Standards (APA/AERA/NCME, 1999) advise test designers and researchers to define the construct that the test is intended to measure and to use this construct as the basis for test design. Furthermore, test designers are required to collect various types of evidence to establish construct validity. However, the conception of construct as discussed in the Standards did not provide enough guidance for the design and validation of the new Internet-based Test of English as a Foreign Language (TOEFL® iBT). Attempting to follow the Standards, test designers spent years identifying relevant theoretical and empirical perspectives to define the construct of Academic English language proficiency. This approach succeeded in generating test tasks and research, but ultimately the validity of TOEFL iBT score interpretation and use was not justified on the basis of types of evidence pertaining to construct validity. Instead an interpretive argument was developed that specifies intended inferences (including a construct-related inference) that underlie score interpretation and use (Kane, 1992; 2001; 2006). Kane suggests that an interpretive argument will guide validation research and the construction of a validity argument, and it did so in the new TOEFL project. However, when an interpretive argument underlies the validity argument, what becomes of the construct and construct validity? This paper explains the role the construct played in the interpretive argument for the TOEFL. We explore the extent to which the construct’s place in the interpretive argument affects the language construct that underlies test interpretation and use.

Session 3, Paper 2: Sun., June 10th, 12:15-12:45

How the use of the visual channel affects L2 listening test-taker performance
Elvis Wagner, Teachers College, Columbia University

Traditionally, tests of second language (L2) listening ability have used only the aural channel to present spoken texts to the test-takers. While numerous researchers (e.g., Antes, 1996; Brown, 1995; Burgoon, 1994; Kellerman, 1992; Pennycook, 1985; Rost, 1990) have suggested that the non-verbal information that a speaker conveys using the visual channel might contribute to listeners’ comprehension and test-taker performance, there have been few studies (i.e., Baltova, 1994; Brett, 1997; Gruba, 1993; Parry & Meredith, 1984; Shin, 1998) that have investigated this empirically, and these studies have ambiguous results.

The current study used a quasi-experimental design to examine the effect of the utilization of the visual channel (through the use of video texts) in presenting spoken texts to L2
listening test-takers on test-taker performance. Using multivariate analysis of covariance (MANCOVA) to analyze the test results, it was found that the experimental (video) group scored higher on the post-test than the control (audio-only) group, and the difference in scores was statistically significant.

In addition, this study utilized verbal report methodology to investigate how test-takers reported attending to and utilizing the non-verbal information in the video texts. These verbal report data were then coded for the cognitive and metacognitive strategies the test-takers used while processing the text and answering the test items. These data suggested that test-takers vary in their ability to utilize the different components of the non-verbal information, and that this variance can be considered construct relevant when testing L2 listening ability in a communicative language ability framework.

Session 3, Paper 3: Sun., June 10th, 12:45 - 1:15

Construct validation of Bachman and Palmer’s (1996) strategic competence model over time in EFL reading test

Aek Phakiti, *University of Sydney*

This article reports on a large-scale study that aims to validate the theory of strategic competence proposed by Bachman and Palmer (1996) through the use of structural equation modeling (SEM). The present study examines the relationship of test-takers’ long-term strategic knowledge (i.e., trait strategies) and actual strategy use (i.e., state strategies) to second language (L2) reading test performance over time. The data were gathered on two occasions (during the midterm and final examination periods). 561 Thai university test-takers answered a trait strategy use questionnaire prior to the midterm and final reading achievement tests and, immediately after completing each test, they answered a state strategy use questionnaire. It was found that (1) trait metacognitive strategy use (MSU) directly and strongly affects trait cognitive strategy use (CSU); (2) trait CSU does not greatly affect state CSU; (3) trait MSU directly affects state MSU in a specific context, which in turn directly affects state CSU; and (4) state CSU directly affects a specific language test performance to a varying degree.

Paper Session 4

Session 4, Paper 1: Sun., June 10th, 3:00 - 3:30

The value of on-line training for essay marking

*Paul Jaquith, Annie Brown, and Karen Caldwell, Ministry of Higher Education, UAE*

In recent years it has become widely recognized that rubric-driven assessment is the most valid and reliable approach to assessing student essays. This approach requires that raters be carefully trained and calibrated, which can be difficult when raters and students are widely distributed geographically.

In recent years online rater training has been emerging as a viable approach to training geographically-distributed raters. There are several benefits. Raters access training materials via the internet and are able to practice and self calibrate by orienting themselves to the rating scales and marking sample scripts online. The online system provides immediate feedback on the rater’s performance.
However to date there has been little research into rater attitudes to this kind of training and its effectiveness in enhancing levels of inter- and intra-rater agreement, especially in contexts where raters are non-native speakers and unfamiliar with rubric-driven writing assessment.

This presentation will report on a country-wide project to train all Grade 12 English teachers, in a country where most are non-native speakers of English and few are familiar with rubric-driven assessment, to assess the writing component of a new English exam. Some trained independently, using the online program, and other trained in traditional assessment workshops. All completed a set of accreditation assessments. Following the training they completed a survey canvassing their attitudes to the training.

The paper will report on the findings of the study with regard to (a) the severity and stability of the ratings awarded by the two groups, and (b) attitudes to the two types of training.

Session 4, Paper 2: Sun., June 10th, 3:30 - 4:00

Validation of automated essay scoring for ESL writers
Sara Cushing Weigle, Yanbin Lu, and Amanda Baker, Georgia State University

Automated essay scoring has the potential to dramatically reduce the time and costs associated with writing assessment, but the use of automated scoring must be validated against a variety of criteria for it to be accepted by test users and stakeholders. Much of the research on automated scoring has compared automated scores on essays to the scores given by human raters. However, it is equally important to validate automated scores against other indicators of writing ability. The research presented in this paper investigates the validity of TOEFL iBT scores generated by ETS’s automated essay rating tool, e-rater, by examining relationships between e-rater scores on two different writing tasks and a set of independent indicators of writing ability.

Data for the study, collected from 400 students at various U.S. universities, include the following: e-rater and human ratings on two different TOEFL iBT tasks; student self-assessments of English language ability; instructor assessment of student language (and specifically writing) ability; and two writing samples from content courses. Relationships among variables were investigated through correlations between criterion variables (self-assessment, instructor assessment, and independent writing samples) and e-rater total and specific feature scores. For those e-rater variables having significant correlations with criterion variables, regression analysis was used to identify the variables that best predicted the criterion variables. Consistency across prompts (i.e., test-retest reliability) was investigated through multi-method correlational analysis, similar to that in Attali and Burstein (2004). Implications for improving automated scoring of ESL writing will be discussed.

Session 4, Paper 3: Sun., June 10th, 4:00 - 4:30

Investigating the construct validity of the Community English Program (CEP) writing test
Taejoon Park, Teachers College, Columbia University

Language testers have long recognized that a variety of factors affect test scores of performance-based writing tests. As Purves (1992) notes, these factors may interact with
each other in complex ways because “different tasks present different problems, which are treated differently by students and judged differently by raters” (p.112). Therefore, in order to better understand the complex nature of writing performance assessment, multiple sources of variance in test scores need to be investigated in a systematic analysis in which the effects of these different sources of variability can be disentangled.

In the present study, I constructed and piloted a writing test consisting of two writing tasks that were very different in the ways that they were contextualized. The primary purpose of this study was to investigate the extent to which the examinees' (N=278) performance on the writing test was affected by the aspects of writing ability being tapped into by the four analytic rating scales (i.e., task fulfillment, content control, organizational control, and language control) and the test methods (i.e., the tasks and raters) used to elicit and score test performance. I modeled the data for this multitrait-multimethod study by means of multivariate generalizability theory and structural equation modeling.

The results and implications of this study will be reported in detail with emphasis on:

1. the importance of task/context in writing performance assessment and
2. the role of test method facets from the interactionalist perspective of construct definition (e.g., Chapelle, 1998, 1999)

Paper Session 5

Session 5, Paper 1: Sun., June 10th, 4:50 - 5:20

**Patterns of test-taker interaction in the First Certificate in English**

*Evelina Galaczi, University of Cambridge ESOL Examinations*

Qualitative methodologies can offer the assessment community new and illuminating insights, as the growing field of discourse-based studies of oral language assessment attests. The present study situates itself within this interdisciplinary field and provides an illustration of the value of using a Conversation Analysis approach in the investigation of a paired speaking test (Cambridge ESOL’s First Certificate in English). Whilst the paired test format has been used by Cambridge ESOL for almost a decade, with numerous research endeavours investigating relevant issues, there is still a shortage of studies focusing on the nature of talk elicited by the paired task. The current study aims to fill that niche and offer a conversation analytic investigation of the talk which is co-constructed within dyads of FCE test-takers (N = 30 dyads). A further aim is to investigate a possible relationship between the CA findings and candidate scores on the test.

I will begin the presentation by providing an overview of the available research on paired tests and relevant issues at play. Next, I will present the FCE oral test and will move on to a discussion and illustration of the test-taker patterns of interaction which emerged from the analysis. I will focus on the dimensions of mutuality, equality and conversational dominance, and make specific reference to conversational devices which distinguished between each pattern, such as turn-taking strategies, follow-up questions, means of getting the floor, gaps and interruptions. Finally, I will discuss the study’s implications for assessment and its relevance to rating scale construction by focusing on the relationship between the CA findings and candidate scores.
Session 5, Paper 2: Sun., June 10th, 5:20 - 5:50

Facets of speaking proficiency
Rob Schoonen, Nivja de Jong, Margarita Steinel, Arjen Florijn, and Jan Hulstijn, University of Amsterdam

In this presentation, we address the following two questions: (1) What is the contribution of linguistic knowledge and psycholinguistic skills to (L2) speaking proficiency?, and (2) To what extent is the contribution of the above variables to speaking proficiency dependent on the level of speaking proficiency and the kinds of tasks the speaker has to perform?

Participants (200 L2 speakers of Dutch) performed eight speaking tasks varying in context (formal or informal), discourse mode (narrative or argumentative) and presumed content complexity (more or less complex). Furthermore, participants took two linguistic knowledge tests, one tapping their vocabulary knowledge and one tapping their morphological and syntactic knowledge. They also did two psycholinguistic tests, aimed at measuring the fluency or speed with which lexical and morphological and syntactic knowledge is accessible. In a correlation/regression analysis, we investigate the relationships between the (psycho)linguistic variables and performance on (different types of) speaking tasks. A group of 50 native speakers did the same (psycho)linguistic tests and (the same) speaking tasks, serving as a point of reference for the interpretation of the relationships found in L2 speakers.

Session 5, Paper 3: Sun., June 10th, 5:50 - 6:20

The use of contextualization cues in a semi-direct speaking test
Hyunjoo Kim, Teachers College, Columbia University

The earlier ‘static’ views on context seem to place a great deal of emphasis on listing a host of relevant features (e.g., settings, participants, etc.), and describing its unidirectional influence on language use. In contrast, relatively recent ‘dynamic’ notions of context focus on the interactional relationship between context and language use. In other words, within the dynamic view, it is no longer context alone that influences language use, but it is also the language that helps define what the context is.

Contextualization cues (Gumperz, 1981, 1992), such as the levels of prosody, paralinguistic signs, code choice, and choice of lexical forms or formulaic expressions, play a critical role in creating a context in a communicative situation. Much of the communication breakdown arises from differing uses and interpretations of contextualization cues. Thus, for constructing proper context and furthermore successful communicative acts, it is essential for speakers to be able to use and interpret the cues effectively.

In this study, using an interactional sociolinguistics approach, selected responses to a semi-direct speaking test of English were analyzed in order to address the following questions. First, what are the types of contextualization cues employed by test takers? Second, what communicative meanings and functions do the cues perform? Lastly, what are the differences in the use of contextualization cues observed between high and low proficiency students?

The theoretical and practical implications of the findings on the nature of second language speaking ability will also be discussed in the presentation.
Paper Session 6

Session 6, Paper 1: Mon., June 11th, 9:50 - 10:20

Writer reader relationships in undergraduate textbooks and in academic reading tests
Cyril Weir, University of Bedfordshire and Sara Gysen, Universiteit Leuven/Centrum voor Taal en Migratie

Text selection for reading tests is too often left to the connoisseurship of test developers. While we are readily able to monitor such subjectivity with computerised analysis of various linguistic features such as frequency of vocabulary, lexical density, complexity of structures, functional range, and through various readability indices, the critical dimension of writer-reader relationships has to date been left largely unexplored.

The dimensions of the writer reader relationship in both types of text, real life and test, are the focus of this study. This paper reports on initial research into a comparison of this parameter in texts used by first year undergraduates at British and Dutch universities with the corresponding English and Dutch texts used in reading tests to determine the suitability of candidates for tertiary medium study. Both qualitative and quantitative measures are employed to examine a number of salient aspects of the writer reader relationship including the interactive and interactional dimensions of this relationship (Hyland 2006), and the disposition of new and given information in text.

The “equivalence” of text features across test and real life in both Dutch and English will be addressed.

Session 6, Paper 2: Mon., June 11th, 10:20-10:50

Effects of thinking aloud on ESL essay rater performance: A FACETS analysis
Khaled Barkaoui, Ontario Institute for Studies in Education of the University of Toronto

Think-aloud protocols have been used in several studies to investigate the decision-making processes that raters employ and the essay aspects they attend to when marking compositions. However, to my knowledge, only one study (Lumley, 2005) has examined the effects of this data-collection technique on rater performance. This paper reports on a recent study to examine the effects of thinking aloud on the performance of raters of ESL essays. The study is part of a research project that used think-aloud protocols to examine the effects of scoring methods and rater experience on ESL essay rating processes and outcomes. The project included 33 novice and 31 experienced raters, who each marked a random sample of 24 essays (out of 200 essays) holistically and analytically. A subsample of 15 novice and 15 experienced raters marked 12 of the essays silently and 12 while thinking aloud, both holistically and analytically. All raters marked an anchor set of six essays to link data across raters and rating conditions (think-aloud and silent). Each rater was then interviewed about their perceptions of thinking aloud. A multifaceted Rasch analysis (FACETS) of essay scores was used to examine: (a) think-aloud effects on rater severity and self-consistency and essay-rater interactions, and (b) interaction effects of thinking aloud with rater experience and scoring method. The interviews were analyzed qualitatively to shed light on FACETS results. The results of these analyses will be discussed in the presentation.
Session 6, Paper 3: Mon. June 11th, 10:50 - 11:20

Effects of source texts on summarization performance

Guoxing Yu, University of Bristol

The cognitive demands of summarization tasks are dependent upon many factors (Hidi & Anderson 1986; Yu 2003, 2004, 2005, 2006, in press). This paper focuses on one of such factors – the qualities of source texts to be summarized. One hundred and fifty-seven Chinese undergraduates were asked to write summaries, in both English and Chinese, of one of the three extended English texts which were of similar length and readability, but differed in other discoursal features such as vocabulary density and macro-organization. The effects of the qualities of source texts were investigated through two parallel datasets: students' summarization task performance and their perception of such effects, as demonstrated in the post-summarization questionnaire and interviews. It was found that source text had significant and similar amount of effects as students' reading abilities on summarization performance, and that such effects were more pronounced on the Chinese than English summarization performance. These findings from the performance data were generally supported by the perception data which shed further light on the dynamics of the summarizability of a source text. The students held that the macro-organization, frequency of unfamiliar words, topic familiarity and interest, length of source texts were among the most influential to determine the summarizability of a source text. However, the effects of summarizability of a source text on students' summarization performance were idiosyncratic. Implications of these findings are discussed in relation to source text selection for designing summarization tasks as a measure of reading comprehension.

Paper Session 7

Session 7, Paper 1: Mon., June 11th, 11:40 - 12:10

Capturing accuracy in second language performance: Exploring global and local measures

Gillian Wigglesworth, University of Melbourne and Pauline Foster, St. Mary's College

An increasing number of research papers in language testing use discourse analytic techniques to assess candidate accuracy in their second language performances (McNamara Hill & May, 2002). In this light, it is timely to explore the two most commonly used measures: local measures which track particular grammatical features, and global measures which assess the overall accuracy of performance.

Both have drawbacks. With local measures there is the issue of which grammatical features will best give a proper reflection of underlying competence. How are they to be chosen? What confidence can we have that a chosen feature will occur often enough in a second language performance to make measurement worthwhile? For global measures to be calculated, language data needs to be segmented into units which are both reliable and valid. Longer units are likely to disadvantage learners who are attempting a more complex L2 performance because the likelihood of error increases with the length of the unit. Moreover, within a single unit, one minor error carries the same weight as repeated major errors.
While it is concluded that, on balance, global errors provide a more reliable evaluation of accuracy in L2 performance, this paper will discuss a more finely tuned global measure which classifies errors at different levels, i.e. those that seriously impede communication, those that impair communication to some degree, and those that do not impair communication at all. The paper discusses the problem of reliably identifying these levels and presents analysed samples from written and spoken L2 performance data.

Session 7, Paper 2: Mon., June 11th, 12:10 - 12:40

Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test

June Eyckmans, Erasmushogeschool Brussel & Vrije Universiteit Brussel, Hélène Stengers, Vrije Universiteit Brussel and Frank Boers, Erasmushogeschool Brussel & Universiteit Atwerpen

In recent years many applied and educational linguists have emphasized the importance of drawing second or foreign language learners' attention to standardized multiword expressions (such as collocations and idiomatic expressions), referred to in the literature as ‘lexical phrases’, ‘multiword units’, ‘formulas’, ‘prefabricated chunks’, ‘ready-made utterances’, etc. (e.g., Pawley and Syder 1983; Nattinger and DeCarrico 1992; Howarth 1998; Foster 2001; Wray 2002). This study reports a small-scale experiment that was set up to estimate (i) the extent to which the use of formulaic sequences (standardised phrases such as collocations and idiomatic expressions) can help learners come across as proficient L2 speakers and (ii) the extent to which an instructional method that emphasizes ‘noticing’ of L2 formulaic sequences can help language learners add such phrases to their linguistic repertoire.

Schmitt and his colleagues (Schmitt 2004) adopt ‘formulaic sequences’ as the overarching term for standardized phraseology. Formulaic sequences can be very diverse, in terms of lexical composition as well as function: they range from simple fillers (e.g., Sort of) and functions (e.g., Excuse me) over collocations (e.g., Tell a story) and idioms (e.g., Back to square one) to proverbs (e.g., Let’s make hay while the sun shines) and lengthy standardised phrases (e.g., There is a growing body of evidence that). Wray (2002:9) defines a formulaic sequence as “a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.” This definition clearly acknowledges the fuzzy nature of the category called ‘formulaic sequences’, since a stretch of words that is processed holistically by one individual need not be processed that way by another. Furthermore, ‘holistic processing’ itself may be a matter of degree. For example, the results of psycholinguistic experiments suggest that what appears to qualify as a formulaic sequence to a native speaker may be less likely to qualify as such for a non-native learner who still seems to process the given sequence in a word-for-word fashion (Underwood et al. 2004; Schmitt and Underwood 2004).

There are roughly three reasons why a command of formulaic sequences in L2 is believed to be beneficial to learners. Firstly, many standardised or institutionalised multiword expressions are predictable neither by ‘grammar rules’ nor by the properties of the individual words they are composed of. In other words, they reflect Sinclair’s (1991) ‘idiom principle’. Mastery of the ‘idiomatic’ dimension of natural language can help learners come across as ‘native-like’. Secondly, since formulaic sequences are believed to be retrieved from
memory holistically, i.e. as prefabricated, ready-made chunks, they are believed to facilitate fluent language production under real-time conditions (Skehan 1998). In fact, one of the signals to help recognise a formulaic sequence in a speaker's real-time discourse is the absence of hesitations within the sequence. In this view, hesitations should occur only in parts of discourse that connect the prefabricated chunks. This leads us to the third reason why mastery of formulaic sequences is believed to be beneficial to learners: formulaic sequences (at least those that are 'correctly' committed to memory) constitute ‘zones of safety’ and appropriate use of them may thus confine the risk of ‘erring’ to the spaces in between the formulaic sequences in one's discourse.

Given the above assumptions, the following hypotheses are put to the test in the reported study: (a) mastery of formulaic sequences can help learners come across as fluent L2 speakers (since stringing together pre-fabricated phrases is believed to be faster than generating sentences word by word and would involve fewer hesitations); (b) mastery of formulaic sequences can help learners come across as having acquired a good range of expression in their L2 (since formulaic sequences are believed to be part of native-like idiomatic language); and (c) mastery of formulaic sequences can help learners come across as accurate L2 speakers (since the prefabricated chunks are believed to be zones of safety).

The pedagogical message to draw learners' attention to formulaic sequences has been conveyed to the teaching community perhaps most successfully by Lewis (1993), for whom ‘chunk-noticing’ is at the heart of his Lexical Approach. In this approach, learners are systematically encouraged to notice recurring lexical chunks in the authentic L2 language they are exposed to. Lewis does not propose many mnemonic strategies to help learners commit those chunks to memory, but seems to rely mostly on the power of awareness-raising to trigger acquisition through imitation of sequences encountered either inside or outside the classroom. To our knowledge, no ‘hard’ empirical evidence of the effectiveness of ‘chunk-noticing’ has been published yet. It is true that Lewis' collaborators report action research (Lewis 1997; 2000) inspired by the Lexical Approach, in which students responded positively to the activities, but none of the described projects include any comparison with potential learning gains under control conditions, let alone any statistical analysis of such comparisons.

In our paper presentation we will report an experiment in which 32 college students majoring in English were exposed to the same authentic language input (audio, video and textual).

During exploration of those materials, the experimental students' attention (N 17) was directed to formulaic sequences, which means that the co-text of words was given special consideration. In other words, students were encouraged to appreciate the syntagmatic dimension of vocabulary. In the control group (N 15) the target language was analysed in a more traditional way: the distinction between grammar and vocabulary (which the learners were accustomed to) was upheld. The authentic language was used to exemplify grammar patterns (explained in one of the other English courses) and to draw students' attention to vocabulary at a paradigmatic level (i.e., by looking at the place of individual words in lexical networks alongside synonyms, antonyms, etc.). In other words, the central tenet of Lewis' Lexical Approach was applied in the experimental group by turning 'pedagogical chunking' into a frequent classroom activity, whereas this was not the case for the control group. The experimental and the control group were taught by the same teacher and received the same amount of class instruction (22 hours).
Afterwards, the participants’ oral proficiency in an interview was gauged by two blind judges. Both perceived the experimental group as more proficient than the control group. Two other blind judges counted the number of word combinations produced by the participants that they considered to be formulaic sequences. Their counts correlated well with the oral proficiency ratings, which suggests that helping learners build a repertoire of formulaic sequences can be a useful contribution to their oral proficiency. Further analysis also suggested that the experimental students’ benefited from their enhanced awareness of L2 formulaic sequences especially by turning it into a strategic advantage through noticing and ‘recycling’ word combinations they had been exposed to.

**Paper Session 8**

**Session 8, Paper 1: Mon., June 11th, 3:10 - 3:40**

**E-portfolio assessment and individual item selection**
*Vivien Berry, The University of Hong Kong*

Final year university students in Hong Kong are encouraged to take the IELTS test as an ‘exit-test’ in order to provide evidence of their language proficiency and thus assist potential employers in their selection processes. However, for a number of reasons, the strongest and weakest students decline to take IELTS and the majority of candidates therefore score Band 6 overall. In order to provide a means of discriminating more finely between students’ language proficiency, an inter-institutional project has recently been instigated to develop a sector-wide, flexible e-portfolio system, which will offer end-users a rich diachronic panorama of students’ progress and achievements in language learning and a record of their language proficiency levels on graduation. Each portfolio will have a uniform summary page.

This paper reports on a study which investigated how one of the key elements of portfolio assessment, item selection, can be accommodated within a standardized system. Sixty-nine students compiled e-portfolios over a period of twelve weeks. Although several items were compulsory, in order to allow for individual expression, a maximum of three additional items selected to provide evidence of ability could be included. Students also responded to pre and post-portfolio compilation questionnaires. Results showed surprisingly little relationship between test scores, attitudes towards language learning and number and range of individual items included. Analysis of their reflection statements showed that there was a divergence between teachers’ and students’ views of the language abilities demonstrated. Changing socio-cultural expectations also affected disparate item selection for demonstration of language proficiency.

**Session 8, Paper 2: Mon., June 11th, 3:40 - 4:10**

**Looking at assessment and feedback practices from multiple perspectives**
*Ari Huhta and Mirja Tarnanen, University of Jyväskylä*

The presentation reports on research into language assessment and feedback practices in the Finnish comprehensive schools. The study is part of a more extensive project that aims at exploring and interpreting students’ and teachers’ literacy practices in the first and second/foreign languages, both in and out-of-school contexts.
Questionnaire surveys were carried out in 2006 on (1) students at the end of compulsory education (15/16-year-olds), (2) their L1 teachers and (3) their foreign/second language teachers. All surveys were based on statistically representative samples; the results are thus generalisable to all students and language teachers at this stage of education in the country.

The study approaches the topic of interest from several complementary points of view and with complementary research methods. Firstly, practices in both the first and second/foreign languages are covered to see to what extent they match or differ. Secondly, both students and teachers are involved in the study, to compare their attitudes to assessment and to find out if they perceive the practices that take place at school in the same way. Thirdly, the survey will be complemented with other approaches, such as individual and focus-group interviews, and classroom observation, in the later phases of the study.

The presentation covers the main results of the surveys on assessment and feedback practices (e.g. division of responsibility, what is assessed, how feedback is given), and focuses on comparing the students’ and teachers’ views on the same questions, as well as on comparing the practices in the mother tongue and second/foreign languages.

Session 8, Paper 3: Mon., June 11th, 4:10 - 4:40

Using activity theory to investigate constructs of competence in the MFL classroom
Kathryn Hill, University of Melbourne

There is a growing body of research which suggests that the introduction of (foreign) languages (FL) programs in primary school does not necessarily result in higher proficiency outcomes in the longer term. ‘Discontinuity’ between primary and secondary school programs (e.g., in terms of content and approach) has emerged as a major contributor to this problem. The present study looks at the ‘continuity’ issue with respect to the way ‘competence’ (for assessment purposes) is constructed in primary and secondary school FL classrooms.

The notion of ‘competence’ is examined through the lens of classroom-based assessment. Paraphrasing Leung (2005), the first research question is

1. what do teachers do and look for, and what theory or standards do they use when assessing their students?

In line with a social constructivist view of classroom-based assessment, the second question is

2. do learners share the same understanding of how competence is constructed?

In order to investigate the issue of continuity, the third question asked

3. is competence constructed in the same way in late primary and early secondary languages classrooms?

The study was longitudinal, tracking students from their last year of primary school (Year 6) to their first year of high school (Year 7). Data include participant observation, classroom recordings, interviews with teachers and students and document analysis.

The findings have relevance for language policy as well as for classroom practice.
Session 8, Paper 4: Mon. June 11th, 4:40 - 5:10

Assessment for literacy education in secondary schools: A comparison of 12 cities
Alister Cumming, Ontario Institute for Studies in Education of the University of Toronto

What are the principal dimensions of variability in policies and practices for the assessment of literacy in secondary schools internationally? I approached this question through 54 face-to-face interviews with noted literacy researchers at universities as well as experienced teachers in secondary schools in two cities in each of six countries as well as through content analyses of local policy, curriculum, and assessment documents. The cities were selected to be of comparable size (one large and one smaller city in each country) and socio-economic circumstances as well as to have culturally diverse populations: Toronto and Vancouver (Canada), Amsterdam and Groningen (Netherlands), Madrid and Murcia (Spain), Athens and Rhodes (Greece), Nagoya and Kyoto (Japan), and Seattle and Washington, DC (United States). Themes identified across the interviews and document analyses indicate that formal and informal assessment policies and practices vary according to the status of languages in schools (i.e., majority, minority, and foreign languages), cultural factors in the local context (recent immigration and minority populations), the degree of national centralization or local autonomy for curriculum policies, the stakes and content of formal examinations, the extent of separate public and private education systems, the cultural history and characteristics of the dominant language of literacy, streaming into different school program types, specific curriculum frameworks, pedagogical innovations, social class distinctions, the importance attributed to foreign or international languages, and the perceived salience of gender differences.

Paper Session 9

Session 9, Paper 1: Mon., June 11th, 5:30 - 6:00

Discourse context, complexity, and task variation in the ESP assessment
Carol Lynn Moder and Gene B. Halleck, Oklahoma State University

Performance-based testing of English for Specific Purpose testing has focused on the identification of situation appropriate tasks that allow generalizations to real world discourse contexts (McNamara 1996, Douglas 2000) and on the variation in test-taker performance across task types (Elder et al. 2002, Fulcher & Reiter 2003, Skehan 1998, Bachman, 2002). Questions remain whether test-taker ability and task characteristics should be conceptualized independently. An additional area for concern is the whether task difficulty will be consistent across learners and contexts, or whether the discourse functions of the specific purpose contexts may be a determinant of task difficulty.

This study focuses on specialized radiotelephony and general ESP tasks used to assess the oral proficiency of 12 Central and South American Air Traffic Controllers. Test-taker performance was compared in terms of ratings on pronunciation, structure, vocabulary, and fluency, as well as through a discourse analysis of their test performance in comparison to their performance in authentic target language tasks.

Cognitive and code complexity did not adequately predict test-taker performance. Test-takers were often more fluent and effective on what would generally be conceived of as cognitively complex tasks and more disfluent on “simple” tasks. Furthermore, the
difficulty of the tasks was clearly related to the similarity of task items to frequently performed discourse functions. These results indicate that features of task complexity must be analyzed in relation to the normal target language context of use. We discuss the implications of these findings for models of ESP testing.

Session 9, Paper 2: Mon., June 11th, 6:00 - 6:30

**English for occupational and academic purposes: Same or different?**

*Catherine Elder, University of Melbourne*

Accrediting bodies differ in the English language requirements for overseas health professionals applying to practice in Australia with some mandating the International English Language Testing System (IELTS), others the health-specific Occupational English Test (OET) and others using the two tests interchangeably. The question of whether a candidate has the same likelihood of achieving the requisite “pass” threshold on either test clearly has important implications for policy-making and, ultimately, for community standards of health care.

The current study explores this issue by comparing the performance of a sample of health professionals with a range of specializations who sat an official version of both the IELTS and the OET within a 3-week time span and then filled out a questionnaire canvassing their views on each test's difficulty and fairness. Results revealed significant correlations between all skill components of the two tests except Writing. However, as expected, given the marked differences between each tests in both content and format, these relationships were not strong enough to support the notion that the two tests are measuring the same construct. While a slightly higher percentage of candidates reached the required threshold on the OET, there were some who “failed” the OET and “passed” the IELTS, raising doubts as to whether one test is indeed more difficult than the other. Test-taker feedback nevertheless showed a clear preference for the OET as a measure of professional language ability. Implications are drawn for communications with test users and for the theory and practice of LSP testing.

Session 9, Paper 3: Mon., June 11th, 6:30 - 7:00

**Testing academic language or academic content? An exploration using CFA**

*Dorry Kenyon and Bokyung Cho, Center for Applied Linguistics*

Current legislation mandates that all U.S. English language learners in kindergarten to grade 12 be tested annually to demonstrate progress in acquiring English. The WIDA Consortium’s ACCESS for ELLs test, an English language proficiency test used by 15 member states and now administered to over 400,000 students annually, is designed to allow students to demonstrate reaching five standards related to the acquisition of the English necessary “for academic success.”

This paper reports on a confirmatory factor analytic study investigating the degree to which ACCESS for ELLs listening and reading items appear to assess academic English language proficiency rather than academic content knowledge. Data comes from 18,000 students who took the first operational ACCESS for ELLs (Spring, 2005).
Based on the underlying theoretical design of the test, a structural model was developed that posits nine variables underlying test performance: listening and reading proficiency in English, proficiency in social and instructional English language, and one variable each for content area knowledge in math, science, social studies, and language arts. Using EQS, the model was fit to eight data sets: two levels of proficiency (low and high) in four grade level clusters (1-2, 3-5, 6-8, and 9-12).

There was very good model fit across all data sets. Overall, items loaded more strongly on the proficiency variables than on content ones. Some sets of items, however, had surprisingly high loadings on content-related variables. The CFA approach provided useful information to identify items, and consequently item specifications, needing revision to avoid testing content knowledge.
Impact of varieties of English on rater judgment in oral tests

Huei-Lien Hsu, Carsten Wilmes, Kadeessa Kadir, and So-young Jang, University of Illinois at Urbana-Champaign

Despite current discussion turning to the legitimacy of varieties of English in language testing, little empirical research attention has been focused on the impact of varieties of English on raters' judgment particularly in oral/written tests. As linguistic richness is brought by English varieties in oral/written tests, it leads to a complicated issue at the scoring stage, where raters may vary in their recognition and tolerance of varieties of English. Several studies have addressed issues on the status of English varieties as linguistic innovations or errors, and the challenge they face in codification. While most researchers concerned with this issue claim the recognition of varieties of English is still at the developing stage, the implication is clear: there is a potential difficulty raters face as to how they differentiate a test-taker who makes linguistic errors from a proficient learner who presents local linguistic variations. Raters in this study are currently rating an oral test, developed by University of Illinois at Urbana-Champaign and Berlitz International Inc., and eventually to be administered in over 60 countries. Insights gained from this study will also help determine what components may be needed to facilitate rater training. Study findings will provide empirical data to bridge the gap between sociolinguistics and language assessment to increase awareness among linguists, test developers, and educators of how rich linguistic variations impact language assessment.

The effect of accent on an EAP listening test: A DIF perspective

Luke Harding, University of Melbourne

Building on a theoretical rationale for the broader representation of diverse speaker accents on tests of academic listening (Harding & Wigglesworth, 2006), this paper reports on an empirical study of the effects of speaker accent on listening test scores. Test materials, drawn from the Diagnostic English Language Assessment (DELA) and recorded with Australian English, Japanese and Mandarin Chinese accented speakers, are trialed with candidates in Australia and New Zealand, many of whom have Japanese or Mandarin Chinese L1 backgrounds. Where previous studies have gauged the effect of accent on test scores through analysis of variance of total scores (e.g., Major et al., 2002; Tauroza and Luk, 1997), the current study draws on differential item functioning (DIF) methods to investigate whether a bias effect exists that advantages test-takers who share a speaker's accent. Items that show strong DIF are discussed with reference to the particular listening skills that they tap, and the question of whether DIF is evidence of bias is addressed.

Applying survey results to an online tutorial for language test selection

Margaret Malone, Megan Montee, and Colleen Gallagher, Center for Applied Linguistics

What do language instructors need to know about assessment to make informed choices for and about their students? While, to language assessment specialists, the list may be infinite, language instructors need helpful and accessible guidelines for language assessment and test selection as they make decisions about appropriate assessment of instruction.

Recently, attention has focused on the need to educate language instructors about assessment decisions (Thrasher, 2006; Stoynoff and Chapelle, 2004). This work-in-progress
session integrates the results of a survey of language assessment practices of U.S. language instructors (N= 1,613) and their applications to an on-line test selection instrument designed to improve assessment literacy.

The purpose of the survey was to determine the current assessment uses and needs of language teachers; it also provided insight into respondents' misconceptions about assessment and test uses. Common misconceptions included instructors' reported perceptions of student language ability (compared to reports of actual language ability), as well as reported uses of tests for purposes outside those stated by test developers.

A simultaneous project has focused on the development of a test selection tutorial for users of a language test database. The purpose of the database is to help language instructors define their assessment needs before selecting an assessment. The results of the survey, along with three focus groups, have informed the draft development of this tutorial. In this session, the presenters will outline the proposed test selection tutorial and solicit feedback from participants to improve the tutorial to confirm to ILTA Code of Practice standards.

**Investigating task difficulty in ESL reading assessment:**
**Combining research approaches**
*Geoff Brindley, Stephen Moore, and Helen Slatyer, Macquarie University*

In recent years, language learning programs worldwide have seen a rapid expansion in the use of outcomes-based curriculum frameworks which rely on teacher-developed tasks as a basis for assessing and reporting learner achievement. However, although such frameworks are consistent with current moves towards “curriculum-embedded” assessment, a number of problems have arisen in relation to the assessment tasks that accompany them. Not the least of these is the question of how to ensure that tasks assessing the same outcome are of comparable difficulty.

This Work in Progress session reports on the fourth of a series of interrelated studies that have explored factors affecting the difficulty of assessment tasks used to report learning outcomes in the Adult Migrant English Program, a large-scale government-funded adult ESL program in Australia. The study focuses on the tasks that are used to assess reading skills of beginning and lower intermediate learners and aims to investigate the question of task difficulty from multiple perspectives. To this end, psychometric analyses of scores on four reading assessment tasks from 169 learners have been conducted, focus group discussions have been held with 22 teachers, and think-aloud protocols have been collected from 20 learners as they completed the tasks. Information from each of these data sources is being analyzed in order to compare teacher and learner perceptions of task difficulty with actual test scores, and to isolate those factors that appear to be most consistently identified as key contributors to difficulty.

**When there is a will, there is a way**
*Neus Figueras, Department of Education. Catalonia, Spain and Kaftandjieva Felianka, University of Sofia, Bulgaria*

This paper will present the results of a study to link two test versions of German as a foreign language to the CEFR levels. The study is based on reading and listening test tasks with known psychometric values made available from different institutions. The estimated levels of the tasks ranged from levels A1 to C1 of the Common European Framework of
Reference for Languages (CEFR). Data were collected from 300 learners, who responded to the tasks, and from 13 experienced teachers (native and non native), who acted as judges in the standard setting procedures carried out. The report will include inter-judge consistency values, and consistency indices between judges’ assessment and item pre-estimation, between judges’ assessment and empirical results, and between pre-estimation values and observed values. Some of the procedures used in the study followed the ones in the Pilot Manual for relating examinations to the CEFR (Council of Europe), but some did not, and the same is the case with the statistical analyses carried out. Results show that careful planning of different methodologies and rigorous work can usefully complement the approach recommended in the Manual. Results also point at ways of complementing the Manual, which may be extremely useful in contexts with small samples and limited resources.

**Developing a diagnostic assessment of academic writing:**
**The construct definition phase**
*Lorena Llosa and Sarah W. Beck, New York University*

The ability to write effectively in a range of genres is crucial to students’ success not only in composition or English classes, but also in courses across the curriculum (Schleppegrell, 2004). Unfortunately, recent data indicate that approximately 1/3 of U.S. high-school students who are planning to continue on into post-secondary education do not meet standards of readiness for college-level writing. Among English language learners (ELL’s), the statistics are even more grim.

The purpose of our study is to develop a diagnostic assessment of high school ELL students’ difficulties with academic writing to help teachers target their instruction more effectively. During the session, we will discuss the design of the first phase of our study, the development of an evidentiary framework of students’ academic writing difficulties that will serve as the basis for the operational construct definition of our diagnostic assessment. Using the work of Bailey and Butler (2003) as a model, we will develop this framework using both a theoretical model of writing and empirical evidence, including (1) the types and characteristics of academic writing that are most relevant in high school contexts in the U.S., and 2) the types of difficulties students encounter in producing this type of writing. Specifically, we will investigate (a) the academic writing demands in local, national and state English Language Arts and ESL standards and standardized achievement tests; (b) teacher and student expectations for academic writing and perceptions of sources of difficulty; and (c) samples of student academic writing from classroom assignments and standardized tests.

**Bias and the native speaker**
*Alan Davies, University of Edinburgh*

Studies of global English refer to the *native speaker* as the model for all users. The question addressed in this research is how necessary the native speaker is. The theoretical insubstantiality of the World Englishes concept of *native user* who is not a *native speaker* is contradicted by practical experience on the post-colonial and English-as-a-lingua-franca ground. Evidence for this contradiction is said to be present in the bias against local varieties of English found in international tests of English Language proficiency (Lowenberg 1993, 2002). By comparing the judgements of Standard British English judges and local variety of English judges on tests of speaking and writing it is hoped to determine the presence (or not) of such bias. Tests of speaking and writing, based
on 2 speaking and 2 writing exemplar IELTS tasks will be constructed for 3 groups of educated non-native speakers of English (N = 3x50), in three locations, chosen to reflect native users (India), English as a lingua franca (Continental Europe) and a control (Australia). There will be 8 raters per group, 4 native speakers of British English and 4 educated speakers of the same variety of English as the test takers. Qualitative analyses (eg discourse organization) and quantitative analyses (eg FACETS) of the scripts as well as introspective interviews with the raters will be carried out and analysed for evidence of bias. Implications will be drawn as to which norms are criterial for new varieties of English.

**Testing receptive vocabulary size: X_lex and Y_lex**

*Imma Miralpeix, Universitat de Barcelona*

Vocabulary size has been shown to be a good indicator of overall language proficiency in a second language, quantifying the amount of words a learner knows is also crucial in the planning of successful learning programs. However, measuring receptive vocabulary size has proved to be extremely difficult. There are very few tests available and they present two basic problems. First, they are tedious to mark, as a test of vocabulary breadth should involve a large number of words. Secondly, most of these tests are not adequate for learners whose mother tongue is a Romance language, as the proportion of cognates overestimates the results (Meara, Lightbown & Halter 1994; Cobb 2000).

In order to overcome these problems, two testing tools have been developed to test vocabulary size using the Yes/No technique: X_Lex (for low-level students) and Y_Lex (for more proficient learners). These tools not only make an automatic estimation of the total receptive vocabulary size of the learner, but they also generate profiles of the test-taker’s vocabulary from 1,000 words to 10,000; that is, they give the percentage of words known at different difficulty levels.

The programs were piloted with two groups of Spanish-Catalan bilinguals studying at university: an intermediate group in Communication Studies (n=93) and an advanced group in Philology Studies (n=64). So far, preliminary results look promising: in an independent-samples t-test, there was a significant difference in favour of the advanced group, which indicates that the programs discriminate between these two proficiency levels. In addition, the profiles obtained were consistent with what was expected from each group.

**New methodologies and concepts in language testing**

*John H.A.L. de Jong, Pearson Language Assessments*

New electronic and internet technologies have had an enormous impact on many aspects of our daily life. For example, submitting transatlantic research proposals in time to meet deadlines is quite different nowadays from how it was 20 years ago. But also the way in which we gather information and purchase goods has undergone substantial change. In the light of these changes it is quite remarkable how language testing lags behind and testing programmes fail to make full use of what technology has to offer.

In this contribution I will offer information on an innovative approach to language assessment which intends to make full use of the latest technological developments. The methodology proposed will allow to look differently at some of the standard concepts in applied linguistics in general, and in language testing in particular.
Based on a wide range of different language tasks the approach aims to provide richer diagnostics and more coherent outcome reports than the standard four skills simplification. In addition, limited use of multiple choice along with a variety of open ended responses will allow to generate estimates of different components of measurement error thus providing means to neutralize their impact on outcome scores.

At the time of the conference apart from the test and item specifications, information about the sampling design for data collection will also be available.
Posters

Revising the writing scale of a large-scale EFL examination
Fernando Fleurquin, Aaron Ohlrogge, and Barbara Dobson, University of Michigan ELI

This poster reports the process and outcome of a project to revise the rubric used to rate the writing section of a large-scale EFL examination from a 4-point scale to a 5-point scale. The revision was undertaken as part of a larger project to improve the feedback provided to test takers about their performance and to relate the writing section of this examination to the Common European Framework of Reference (CEFR).

As the first step of the project, researchers worked to improve their understanding of the target level of the examination, CEFR level B2. They studied sections of the Common European Framework of Reference for languages: Learning, teaching, assessment relevant to written production, with special attention to the B2 level of illustrative scales, and examined tasks and graded writing papers of other examinations linked to the CEFR.

Next, they evaluated responses to the examination’s writing task from 70 test takers with a range of proficiency and language backgrounds. The researchers categorized each paper as “at or above B2 level” or “below B2 level.” After this initial classification, they subdivided these categories, resulting in five groups representing different performance levels, with level 3 representing a minimally-acceptable B2-level performance. Qualitative analyses of characteristics of these five groups of papers led to the development of the new scoring rubric.

This poster illustrates the old and new scales, outlines the development process, and provides information on results of implementing the new scale, including feedback from raters who have used both scales.

The LPATE: Revising a high stakes assessment
Alan Urmston, Hong Kong Polytechnic University, Tom Lumley, Australian Council, and Mark Knight, Hong Kong Polytechnic University

In Hong Kong, perceptions of declining English language standards are part of daily public discourse, and provoke a variety of responses, including major government policy initiatives. In 1995, the Hong Kong government decided that an important reason for this perceived decline lay with the teachers, because many of them lacked either adequate English proficiency themselves or the subject training needed to enable them to teach English effectively (or, perhaps, both). As a result, they set a deadline of September 2006 for all English teachers in state schools to have reached required ‘English language benchmarks’. The Language Proficiency Assessment for Teachers (English Language) (LPATE) was introduced in 2001 as one way in which the teachers could satisfy the government’s requirement.

In 2006, the government moved both to extend the life of the LPATE, with the idea of the Assessment now accepted by many stakeholders, and to commission a test revision project, in response to criticisms of the existing design. This poster presentation describes the revision project. It briefly examines the controversial nature of the implementation of the LPATE, before considering other issues raised in the process of revision, including...
weaknesses of the original test design, problems that arise when a language proficiency test attempts to assess pedagogical knowledge or ability, and the processes involved in trying to persuade stakeholders in Hong Kong of the importance of good language test design.

**A speaking scale revision project—Anticipated and unanticipated ramifications**  
*Sarah Briggs, Melinda Matice, and Fernando Fleurquin, University of Michigan*

Tests of language ability often involve the use of rating scales that, like other test features and components, need to be monitored and periodically revised. This poster documents the revision process of a rating scale used in a speaking test as part of an internationally administered English exam at the high-intermediate level. The original aim of the project was to revise an existing 4-level scale to a 5-level scale applied by interlocutor raters when evaluating the ability of test candidates to sustain a spoken interaction on a range of general, academic, vocational and leisure topics.

While the researchers anticipated the need to re-examine examiner interlocutor behavior and the need for rater training materials as part of the scale revision process, there were some unanticipated issues that needed to be addressed. These included continued qualitative analysis of candidate performance behavior as well as reconceptualizing the tasks in relation to the testing construct. Issues also emerged related to setting the standards and what aspects of candidate performance should be used to evaluate oral proficiency at this level.

The poster will also show a summary of feedback on application of the revised scale from the perspective of those training the oral examiners and the raters themselves. Additionally, it will show the distribution of scores from over 30,000 test candidates whose speaking test score was based on the revised 5-level scale as compared to a comparable group of candidates whose scores were based on the previous 4-level scale.

**The differential effects of time on accommodated vs. standard versions for ELLs: A meta-analysis**  
*Maria Pennock-Roman, MPR Psychometric Research & Consulting and Charlene Rivera, The George Washington University Center for Equity & Excellence in Education*

The objective was to analyze systematically the effects of test accommodations on ELLs’ achievement test scores, taking into account characteristics of the test administration conditions and sample characteristics. In the meta-analysis, only experimental studies in U.S. schools, grades K-12 were included. There were altogether 44 ELL subsamples receiving an accommodation that could be contrasted to ELLs receiving the standard version (control), and 30 non-ELL subsamples receiving an accommodation with a corresponding control. For each subsample, we calculated Glass’s d (effect size—the difference in means divided by the control group standard deviation) and its sampling variance. Then subsamples were classified into categories according to type of test accommodation, time allowed during test administration, and level of language proficiency of the subsample, where available. The weighted average of the effect sizes and its 95% confidence band was calculated for each category. There was evidence that providing generous time limits produced a differentially larger improvement for accommodation conditions for ELLs. For example, for the Dual Language Versions, the average effect size under standard time limits was 0.003 as compared with 0.299 when control and experimental groups were both allowed to have generous time limits. Average effect sizes for Spanish language versions.
were larger in a positive direction for subsamples having low English proficiency (EP) or content instruction in Spanish (1.37) as compared with subsamples having low proficiency in Spanish and/or content instruction in English (0.560).

**Test generation and online delivery system (TGOD)**

*Gary Buck, Pacia Platzek, and Forest Platzek, Lidget Green Inc.*

TGOD is an online test creation and delivery system. It has three main components: an online test creation component, an online test delivery component, and an administration component.

The system is written in JAVA and runs entirely on a powerful Sun Microsystems Sunfire X4100 server. The system automatically uploads a small program, called a thin client, on to the users machine. This functions somewhat similar to a browser, managing the interaction between the server and the client machine, although, unlike a browser, it only has the minimal functionally to run the test, and so it is very small, about 600kb.

The password protected administration module can be accessed from any computer connected to the internet, and allows the user to administer all the functions of the system; not only creating and administering tests, but also accessing scores, forwarding scores electronically, accessing data response files, and inputting descriptive data for test items and tests.

The test creation component has a number of blank templates, each for a different item type, from which new items can be created and stored in the database. All the items in the database are available for use. The test deliverer component offers the user a list of all available tests. A user ID and a password id automatically generated for each test user, the system schedules tests and the tests can be taken from any computer. After the test taker has finished the test, the system scores the test, and saves the results for the administrator.

**The development and maintenance of a test of English for language teachers**

*Robert Edwards, Université de Sherbrooke*

This poster session presents the development and on-going monitoring of a general proficiency test of English as a second language required by candidates in a second language teaching degree program at a Canadian university. The candidates within the program are majority non-native speakers. The test is criterion-referenced and performance-based and includes tasks drawn from an inventory of tasks routinely executed by teachers. The development of the test was collaborative and included actors from different sectors within the university (pedagogy, applied linguistics, and English) as well as actors from within the teaching community (teachers, program administrators, practicum supervisors). The procedures used for the validation of test content (analyses of task complexity and determination of task appropriacy by expert judgment) are described, as are the results of a differential groups study (N=84) that looked at writing ability and grammatical knowledge as part of the larger validity argument, particularly in regard to the determination of cut scores. Procedures for the monitoring of the dependability of the test and how these are reported to test stakeholders are also described.
Measuring grammatical knowledge with limited-production tasks
Jee Wha Dakin and Hyun Jung Kim, Teachers College, Columbia University

In the context of federally-funded, community-based organizations providing language instruction to adult ESL learners, the current study examined the construct validity of a grammar test designed to measure the grammatical knowledge of 100 low-proficiency learners. The test comprised two limited-production tasks, each eliciting six illustration-based responses scored by three trained raters. For the purposes of assessment, the theoretical construct of grammatical knowledge was defined in terms of two separate but related components of form and meaning (Purpura, 2004). Therefore, the responses were scored for grammatical effectiveness and semantic meaningfulness.

First, quantitative analyses were conducted to determine: (1) whether there was evidence of construct validity of the grammar test; and (2) whether the tasks adequately measured the grammatical knowledge of this particular population. Using the FACETS program, the patterns of examinees, raters, items, and components of form and meaning were analyzed. The quantitative results provided evidence that limited-production tasks adequately measured the grammatical knowledge of very low proficiency learners.

Based on the results of the quantitative analyses, it was found that the population was split into six distinct proficiency levels. In order to see the strengths and weaknesses of each proficiency level, in terms of form and meaning, a content analysis of the responses was conducted to gain a deeper understanding of the linguistic features produced in the responses. Findings from the content analysis might have implications related to: (1) pedagogy – principally for diagnostic purposes; and (2) test construction – particularly how grammar tests are developed for low-level learners.

Validation of an expanded definition of the grammar construct
India C. Plough and Jeff S. Johnson, University of Michigan

Recent conceptualizations of grammar have extended the construct definition beyond the traditional form-based model to include both form and meaning, and, in some instances, pragmatics (Larsen-Freeman, 2003; Purpura, 2004). Although some research has focused on the distinction between form and meaning (Saito, 2003), little empirical research has been done to shed light on the view that pragmatics is an essential, ‘testable,’ element of grammar.

The goal of the current research is to validate a definition of the grammar construct, proposed for the University of Michigan’s certificate exams, which incorporates pragmatic knowledge. In very general terms, the study asks: can pragmatic components of grammar be assessed through objective multiple choice items?

This poster session will report on two studies conducted to address the following questions: 1) What are the responses of learners to the new items? 2) How similar will the new grammar items discriminate between proficiency levels when compared to traditional grammar items? and 3) For practical test administration purposes, is there a difference in the amount of time needed to complete a traditional multiple choice grammar item and the amount of time needed to complete a new item?

Preliminary results indicate that learners view the new items favourably and that the amount of time required to complete the new items does not pose administrative concerns.
Creating items that discriminate well, however, has proved to be a challenge. This poster session will describe this challenge and ways to meet it.

**Assessing writing holistically and analytically to investigate an early foreign language start**  
*Teresa Naves, Universitat de Barcelona*

The "consensus view" on the effects of age on L2 learning (Long 1990) argues that in the long run, when time and exposure are kept constant, learners who receive natural exposure to an L2 during childhood attain higher levels of L2 proficiency than those beginning as adults—something which has not been found in foreign language (FL) contexts (Muñoz, 2006). Measures of L2 proficiency in the SLA literature, however, have only recently examined proficiency in the context of writing (Hamp-Lyons, 2002). Also, little research has investigated interlanguage writing ability in terms of accuracy, fluency, lexical and syntactic by means of both holistic and analytic measures proposed by Wolfe-Quintero et al. (1998).

The current poster will report the results of a study designed (1) to use both holistic and analytic measures of FL writing ability to investigate the long-term effects of an early start in a FL context in L2 writing, and (2), to examine how the analytic measures cluster by means of a exploratory factor analysis. The analyses showed that late starters significantly outperformed their younger peers although the effect size was small (p<.001) (See Norris and Ortega, 2006 in this respect). The EFLs revealed that the analytical measures cluster into four rather than five components. The fifth factor was composed of mean clause-, sentence- and t-unit- length, which overlapped with more traditional measures of syntactic complexity. This poster discusses the implications for SLA of assigning analytical metrics to one construct or another (See also Ortega, 2003).

**Health care access for linguistic minorities: Identifying nurse speech acts for assessment**  
*Carolyn E. Turner, McGill University, Michel D. Laurier, Université de Montréal, and Xuelian Wei, McGill University*

One of the most demanding situations in terms of communication for members of linguistic minorities is a conversation between a health professional and a patient (Bender et al., 2004). In Canada, this is the situation for many linguistic minority groups, especially French-speakers outside Quebec and English-speakers in some parts of Quebec. This also occurs with First Nation communities and immigrant groups.

This poster will present the first part of a community-based effort toward developing a suitable assessment instrument to help inform nurses of their language competence in oral interactions with patients. It will present the steps taken to identify speech acts used/needed in such contexts. The final instrument will be based on a real-life approach (i.e., test situations that replicate actual professional situations). The target group is in-service French speaking nurses serving areas in Quebec which include English speaking minority communities.

A mixed methods approach was used to identify relevant speech acts. It included: an extensive review of the literature (e.g., Bosher & Smalkoski, 2002; Cegala, 1997; Laurier et al., 2000); observation of “mock consultations” at a university’s medical simulation centre;
development of an initial list of speech acts; focus groups with in-service nurses to help
determine relevance, difficulty level, and generate examples. The final list was developed
into a questionnaire and administered to the target population of nurses. Results were
calibrated using Rasch analysis and used to derive a band scale corresponding to the
Canadian Language Benchmarks. The resulting analysis will inform the development of the
assessment instrument.

Validating the Catalan and Spanish versions of the MLAT-E in Catalonia
Maria del Mar Suarez Vilagran, Universitat de Barcelona and Charles Stansfield, Second Language
Testing Foundation

The desire to develop a test of language aptitude for Catalan-speaking children led to the
adaptation to Catalan of the Modern Language Aptitude Test-Elementary (MLAT-E) for
Spanish speakers (MLAT-ES) (Stansfield and Reed 2005). This project extends to Catalan
the framework that Stansfield and Reed (2003) spelled out for creating adapted versions
of Carroll and Sapon’s MLAT-E for children whose first language is not English. Although
Catalan and Castillian are closely related Romance languages, the adaptation of this test
posed problems due to linguistic cultural challenges. This poster will describe the
development of the Catalan version of the MLAT-ES and a validity study that involved
administering it and the MLAT-ES to over 500 Catalan and Castillian speaking children in
Cataluna. The results permit us to determine the reliability of the Catalan version, its
correlation with the MLAT-ES, to judge the validity of the Catalan version, and to compare
the validity of both versions for this bilingual group. The poster also portrays some of the
language and cultural problems that were encountered and how they were overcome.

The presenters are interested in making contact with others who would like to use
this instrument.

Construct and task considerations in assessing young EFL learners
Elana Shohamy, Tel Aviv University and Ofra Inbar, Tel Aviv University, Beit Berl College

With the expansion of programs for teaching EFL for young learners world wide, questions
arise as to the suitable assessment formats for this population. In order for the assessment
to be synchronized with the learning of EFL in early age, it needs to be based on a
thorough understanding of what the construct of language teaching in that age consists of.

This poster will present a proposal for a new direction of assessing young learners based on
a research study (Shohamy, Inbar, 2006) which compared home room vs. professional EFL
teachers, showing differences in the teaching approaches used by the two teacher groups.
While the English teachers’ teaching was somewhat removed from the contents of the
general curriculum, the home room teachers approached English teaching as integrated
with the content, conceptual knowledge and L1 literacy development. It was clear later that
the assessment instruments used and based on a view of ‘English as a subject’, (listening
comprehension, N=249) oral interaction, N=52), did not capture the integrative teaching
approach that emerged.

Analysis of the research findings brought forth the need to employ different assessment
measures which combine language with conceptual subject area knowledge in both L-1 and
L-2. This construct is based on a more appropriate theoretical model, reflects classroom
practices and is more in line with how language is used nowadays in societies with a strong
English presence. The poster will discuss the above and provide numerous examples of such assessment tasks.

**Validation of the English diagnostic grammar test:**
**Taking quantitative and qualitative approaches**

*Rie Koizumi, Tokiwa University, Hideki Sakai, Shinshu University, Takahiro Ido, Waseda Junior & Senior High School, Hiroshi Ota, Komazawa Women's University, Megumi Kimura, Dokkyo University, Masatoshi Sato, Ichikawa High School, and Akiko Nemoto, Atago Junior High School*

Despite the importance of using a diagnostic test to improve the quality of learning, teaching, and assessment, there are a limited number of tests specifically for a diagnostic purpose (Alderson, 2005). This poster illustrates the development and validation of the English Diagnostic Grammar Test (Edit Grammar Test), targeting Japanese beginning learners of English. It has been constructed in cooperation with researchers on SLA and language testing and secondary school teachers, with the aid of a non-profit organization, English Language Proficiency Assessment (ELPA).

The Edit Grammar Test has 72 multiple-choice items and is intended to assess understanding of the boundaries and internal structures of basic English noun phrases (NPs). NPs were selected among many grammatical items because teachers in Japan do not systematically teach them, and students’ insufficient knowledge is often unnoticed. The validity of inferences and uses based on the test scores is examined, using both quantitative and qualitative approaches, with further test revision in mind.

The overall results have provided much positive evidence for validity. From the quantitative perspective, it is shown that reliability is fairly high for a low-stakes test and that a high percentage of items and test-takers fit the Rasch model. From the qualitative perspective, a verbal protocol analysis demonstrates that most items elicit the test-taking processes intended by the test constructors. We also report other results including test score differences among students at different ability levels, students’ reactions to the test, and the quality and usefulness of the diagnostic information from the test.

**Candidate awareness of performance features of spoken interaction in peer-peer test performance**

*Anna Maria Ducasse, La Trobe University*

The validation of tests of speaking involving paired candidate formats is increasingly focusing on ways in which interaction between candidates is sustained, as ‘interaction’ emerges as an important criterion for assessing candidate performance. This paper, part of a larger project, focuses on the previously unexplored area of candidates’ awareness of features of interaction in such tests. It uses a Retrospective Stimulated Verbal Recall methodology with video-taped test performances from 25 participants in a beginners’ level test of Spanish.

The analysis addresses the following features of interaction: interactive listening, conversation management and nonverbal communication. These had been identified in a previous Verbal Protocol study in which Spanish teacher/raters demonstrated their orientation to features of interaction in evaluating the candidates’ performance. The aim of the current paper is to explore the correspondence between the features oriented to by raters and the features of interaction that candidates are aware of while performing on the test.
The participant reports are analysed in terms of student orientation to each of the above three features of interaction, and examples from the data are presented and discussed. The findings confirm that not only raters but candidates are significantly oriented to these specific features of interaction. The analysis additionally provides insight into candidates’ test preparation and their experience of interaction with their partners during the test. The paper has important implications for the validation of the construct of interaction in paired oral proficiency tests.

**Features of English L2 children’s oral narrative proficiency**
*Sari Luoma, Ballard & Tighe, Publishers*

The first extended oral presentation structure that children develop is the narrative. Research suggests that native English-speaking children differ in their ability to create narratives, with competent children generating longer, more complete stories that demonstrate knowledge of organization, use of elaborations, and syntactically more accurate sentences.

The present study investigates oral narratives told by non-native English speaking (NNES) children of preschool and kindergarten age. The aim is to examine what dimensions can be discovered in NNES children’s narratives, and whether these dimensions include length and complexity in the same manner that native English speaking students’ stories do. This is important because preschool and kindergarten age children’s proficiency in English is being measured in the U.S. using tasks such as story re-telling. To ensure valid measurement, rating must be focused on appropriate features of the performances.

In the study, the participants re-told a story after first hearing it told to them and answering comprehension questions about it. The children were given a picture sequence to help them remember the story. The performances were evaluated holistically on a 4 point scale.

The study concentrates on the analysis of low and high scorers. Transcripts of the stories are analyzed with respect to knowledge of narrative structure, amount of information, complexity of language, accuracy, and features indicating disfluency such as repetitions and mid-phrase pauses. The results inform a discussion of young learners’ developing L2 English narrative proficiency. Implications are drawn to the content of rating scale descriptors for young English language learners.

**The Certificado Internacional del Español (CIE). A Hispanic certification project**
*Martha Jurado, Universidad Nacional Autónoma de México and Ramón Parrondo, Instituto Cervantes*

Currently, there is no international Spanish assessment system exclusively for academic purposes, or one that includes language varieties equitably. A joint initiative by the Instituto Cervantes (IC) and the Universidad Nacional Autónoma de México (UNAM) has resulted in the development of the Certificado Internacional de Español (CIE). The CIE has been designed in accordance with international quality standards and contains Mexican and Peninsular varieties of Spanish; it seeks to diagnose the degree of suitability of students aged 17 to 30 – in terms of their language competence – for taking up academic courses in Spanish. The CIE will measure candidates’ language competence at a given point in time, irrespective of how they acquired it. The scores obtained could be interpreted according to CEFR and ACTFL scales.
The test proposes a multidimensional model that includes balanced assessment of language competence in standard Spanish and evaluation of the four skills using either the Mexican or the Peninsular variety; thus, candidates will be tested in the variety of their choice through practical tasks that involve listening, reading, writing or speaking. Content validity is ensured through specifications tables developed at the UNAM and analysed by external assessors; extensive pretesting, provides proof of consistency and quantitative validation; the range of indices spans CEFR levels B1, B2 and C. An item bank has also been designed at the UNAM, whose functionalities include input filters, item pretest and follow-up procedures, preliminary application of classic test theory and merging of items for test construction. The IC is developing the on-line test administration system.
News from the ILTA Affiliates

The Academic Committee for Research on Language Testing (ACROLT)
ACROLT will hold its annual conference in Israel in May 2008. The conference will focus on research and policy in language testing and assessment. Further information and the Call for Papers will be available on the ACROLT website www.smkb.ac.il/acrolt. To receive ACROLT updates please join our mailing list by contacting Isabel Berman: berman@idc.ac.il

The Association of Language Teachers in Europe (ALTE)
The ALTE 3rd International Conference will be held from April 10 to 12, 2008 at the University of Cambridge and hosted by Cambridge ESOL. The central conference theme is: The Social and Educational Impact of Language Assessment with sub themes on Assessment for Teaching and Learning, Intercultural Dialogue and Impact and Stakeholders. ALTE Cambridge 2008 is a unique event of unparalleled scope, depth and reach for all professionals with an interest in language assessment and associated issues. For further details, visit: www.alte.org/2008 or e-mail the ALTE Secretariat: alte2008@ucles.org.uk.

The East Coast Organization of Language Testers (ECOLT 2007)
With the support of the Center for Applied Linguistics (CAL), Second Language Testing, Inc. (SLTI), Georgetown University and The National Capital Language Resource Center (NCLRC), the East Coast Organization of Language Testers (ECOLT) will hold its fifth annual conference on November 2 and 3, 2007 in Washington, D.C. Dorry Kenyon from the Center for Applied Linguistics will be the plenary speaker.

Japan Language Testing Association (JALT)
JALT will hold its 11th Annual Research Meeting on October 28th at Aichi Gakuin University in Nagoya. The theme of the conference is: ‘The role of language testing in curriculum reform’. Please check our homepage < www.avis.ne.jp/~youichi/JLTA.html> for information on deadlines for submission of abstracts and other details. JALT is delighted to have had a very successful 10th Anniversary Conference in Kyoto last year with invited speakers from Israel, Korea, and China.

TESOL Arabia Testing, Assessment and Evaluation Special Interest Group (TESOL Arabia TAE SIG)
With the generous support of TESOL Arabia, the Higher Colleges of Technology and the U.S. Department of State, the TESOL Arabia TAE SIG will hold its 11th Annual Current Trends in English Language Testing (CTELT) Conference in November 2007 at Dubai Men’s College in Dubai, United Arab Emirates. The theme of this event is “Writing Assessment.” Date/time: TBD. For more information or to register, go to http://tesolarabia.org or contact Christine Coombe christine.coombe@hct.ac.ae. Hope to see you there!
Pearson Language Assessments

Assessing & recognising language skills
www.pearsonlanguageassessments.com

Raise Expectations!

Pearson Language Assessments provides examinations to assess and accredit the English language usage of students who are speakers of other languages.


Please come and visit us at our table to find out more about these exams.

For further information:
Pearson Language Assessments
One90 High Holborn
London WC1V 7BH
Tel: +44 (0) 207 190 5188
Email: plaenquiries@pearson.com
Language Testing Research Colloquium Attendees
Sign up at the LEA Booth Today and Receive a 20% Discount!

Language Assessment Quarterly

Editor
Antony John Kunnan
Associate Editors
Constant Leung
James Purpura
Nick Saville
Carolyn Turner

Language Assessment Quarterly: An International Journal is dedicated to the advancement of theory, research, and practice in first, second, and foreign language assessment for school, college, and university students; for employment; and for immigration and citizenship. The journal publishes original articles addressing theoretical issues, empirical research, and professional standards and ethics related to language assessment, as well as interdisciplinary articles on related topics, and reports of language test development and testing practice.

Forthcoming Contents

Teacher Assessment as Policy Instrument: Contradictions and Capacities, Leung and Rea-Dickins
Views from the Chalkface: English Language School-based Assessment in Hong Kong, Davison
Educational Reform and ESL Assessment in Australia: New Roles and New Tensions, McKay and Brindley
Grading, Feedback and Reporting in ESL/EFL Classrooms, Cheng and Wang
Investigating the Validity of Math Word Problems for English Language Learners Using DIF, Ockey

Publication Details
Volume 4, 2007, 4 issues per year
Print ISSN 1543-4303 Online ISSN 1543-4311

2007 Subscription Rates
Individual print and online discount rate: US$40/US&CAN/$80(ROW)
Institutional online-only rate: US$250/US&CAN/$525(ROW)
Institutional print and online rate: US$300/US&CAN/$630(ROW)
Visit: www.erlbaum.com

Lawrence Erlbaum Associates, Inc., Publishers (LEA) which publishes more than 100 journal titles and 200 new books a year under the imprints of Lawrence Erlbaum Associates and The Analytic Press has joined Informa’s Taylor & Francis Business.
The staff and consultants of Second Language Testing, Inc. congratulate ILTA on its 15th anniversary. We are pleased to support ILTA and the LTRC.

Be sure to visit our table in the exhibits area to learn more about the Modern Language Aptitude Test–Elementary in Spanish.
The staff and consultants of Second Language Testing, Inc. congratulate ILTA on its 15th anniversary. We are pleased to support ILTA and the LTRC. Be sure to visit our table in the exhibits area to learn more about the Modern Language Aptitude Test–Elementary in Spanish.

Charles W. Stansfield, President
Second Language Testing, Inc.
6135 Executive Boulevard
Rockville, MD 20852-3901
United States of America
301-231-6046, fax 301-231-9536
www.2lti.com
cstansfield@2lti.com
“Nothing today has more significant implications for ELT than the increasing recognition that English is a deterritorialised lingua franca. Andy Kirkpatrick’s arguments will appear striking and original to many readers, although in time we will wonder how we ever thought anything different.”

Peter Grundy, University of Leicester

- discusses the implications of World Englishes on English language learning and teaching
- explores the roles and contributions of multilingual and multicultural English language teachers
- includes an audio CD of authentic examples of World Englishes

www.cambridge.org/elt/worldenglishes
Setting the Standard for Measuring Communicative English

High-Quality Testing and Objective Scoring
Set the New TOEFL® iBT Apart from the Rest

The new Internet-based TOEFL® test (Test of English as a Foreign Language™) measures how well students have mastered the integrated language skills that communicative English requires.

TOEFL test takers demonstrate their mastery of the English-language skills they’ll use every day — in the classroom, at the registrar’s office and in the bookstore. Because it simulates real university life, the TOEFL test helps admissions officers determine if students have the skills they need to succeed.

The TOEFL iBT also sets the standard by providing fair, unbiased and objective scoring that has made ETS the world leader in educational measurement:

- Multiple raters anonymously score responses.
- Raters score responses from more than one country.
- Constant monitoring by scoring leaders ensures quality control and reliability.

Learn more about the TOEFL® test.

Web: www.ets.org/toefl
Phone: +1-609-683-2008
E-mail: TOEFLnews@ets.org

Copyright © 2007 by Educational Testing Service. All rights reserved. ETS, the ETS logo and TOEFL are registered trademarks of Educational Testing Service (ETS) in the United States of America and other countries throughout the world. TEST OF ENGLISH AS A FOREIGN LANGUAGE is a trademark of ETS. 50965
An exciting new collection, Language Testing Reconsidered provides a comprehensive account of important issues in language testing from the perspectives of researchers who have significantly influenced the direction of the field.

Each of the contributors to this collection has re-evaluated language testing by looking back at its history, reflecting on personal experiences, taking stock of current practices and looking ahead to future possibilities.

As a whole, Language Testing Reconsidered captures the history of language testing, highlights useful methodological approaches in test development and validation, and elaborates a research agenda for the future.

Available Summer 2007

To order, call 1-800-565-9523. Visit our online catalogue at www.uopress.uottawa.ca
Research Notes
Quarterly publication reporting on Cambridge ESOL’s research, test development and validation

An invaluable resource with up-to-date information to support your teaching, teacher training or research activity.

Research Notes 27
Focus on testing English for the workplace
- Cambridge ESOL and tests of English for Specific Purposes
- Publishing vocabulary lists for BEC Preliminary, PET and KET examinations
- Using simulation to inform item bank construction for the BULATS computer adaptive test
- The comparability of computer-based and paper-based tests: goals, approaches and a review of research
- Modelling facets of the assessment of Writing within an ESM environment
- Broadening the cultural context of examination materials

Research Notes 28
Focus on testing young learner English
- Reviewing the Cambridge Young Learner English (YLE) Tests
- Research into the marking of spelling for the revised YLE tests
- Cambridge ESOL YLE tests and children’s first steps in reading and writing in English
- Linking language assessments for younger learners across proficiency levels (Phase1)
- Developing Can Do statements for young learners

Research Notes Online
Articles matching your particular area of interest, searchable by:
- exam (e.g. YLE, IELTS) • skill (e.g. listening, speaking) • area (e.g. candidates, research methods)
Visit the website to find articles, subscribe or download back issues of Research Notes.

www.CambridgeESOL.org/researchnotes
Language Testing

Editors:
Catherine Elder The University of Melbourne, Australia
Glenn Fulcher University of Leicester, UK

“Language Testing... is a central issue in second and foreign-language teaching. A substantial literature has developed on the topic... it makes sense to concentrate more of this work into a single journal... Language Testing does this admirably.”

Times Higher Educational Supplement

Language Testing provides a forum for the exchange of ideas and information between people working in the fields of first and second language testing and assessment. This includes researchers and practitioners in EFL and ESL testing, and assessment in child language acquisition and language pathology.

In addition, special attention is focused on issues of testing theory, experimental investigations, and the following up of practical implications.

Recent key articles

• Validity evidence in a university group oral test
  Alistair Van Moere

• How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective
  André A. Rupp, Tracy Ferne, and Hyeran Choi

• Examining the relationship between differential item functioning and differential test functioning
  Tae-Il Pae and Gi-Pyo Park

• Measuring grammatical complexity: the Gordian knot
  Wayne Rimmer

• On the use of the immediate recall task as a measure of second language reading comprehension
  Yuh-Fang Chang

Published Quarterly: January, April, July, October
ISSN: 0265-5322
SIGN UP FOR SAGE CONTENTS ALERTS TODAY http://online.sagepub.com/cgi/alerts

http://ltj.sagepub.com

Subscription Hotline +44 (0)20 7324 8701  Email subscription@sagepub.co.uk

SAGE Publications
www.sagepublications.com
Being one of the ancient capitals of China with a history of 2200 years, Hangzhou has long been a famous tourist attraction. The widely expressed Chinese proverb—“in heaven there is paradise and on earth there are Hangzhou and Suzhou,” combined with the words of Marco Polo, who described this place as “the most beautiful and magnificent city in the world,” explains why Hangzhou attracts millions of tourists from home and abroad each year.

Carefully chosen for you is the conference site, Wang’s Villa, also known as Zhejiang Xizi Hotel, a hotel located at the foot of the hill famous for the former Leifeng Pagoda offering one the ten best views of the well-known West Lake. The hotel borders the lake on three sides and Jinci Temple to the south. With its richly endowed environment, Xizi Hotel has become the ideal Shangri-La.

Welcome to LTRC 2008!
ALTE 3rd INTERNATIONAL CONFERENCE
Cambridge 2008

The ALTE 3rd International Conference will be held from 10 - 12 April 2008 at the University of Cambridge and hosted by Cambridge ESOL.

The central conference theme is The Social and Educational Impact of Language Assessment with subthemes on Assessment for Teaching and Learning, Intercultural Dialogue, and Impact and Stakeholders. ALTE Cambridge 2008 is a unique event of unparalleled scope, range, depth and reach for all professionals with an interest in language assessment and associated issues.

For further details visit www.alte.org/2008 or e-mail the ALTE Secretariat: alte2008@ucles.org.uk

The Association of Language Teachers in Europe (ALTE)

ALTE will hold its thirty-fourth meeting at the University of Vilnius in Lithuania from 7-9th November, 2007 and its 3rd International Conference, The Social and Educational Impact of Language Assessment, from 10-12 April 2008 at the University of Cambridge and hosted by University of Cambridge ESOL Examinations. The call for papers for this international conference is now open at www.alte.org/2008/index.php
www.alte.org/2008/index.php

For additional details on this event visit www.alte.org/2008 or e-mail the ALTE Secretariat: alte2008@ucles.org.uk

If you would like to receive further information about other ALTE events and courses, please write to the ALTE Secretariat at alte@ucles.org.uk
Studies in Language Testing series

published jointly by: University of Cambridge ESOL Examinations/Cambridge University Press
Series editors: Michael Milanovic and Cyril Weir

Studies in Language Testing is a series of published volumes focusing on important developments in language testing and is produced by Cambridge ESOL in conjunction with Cambridge University Press. The series profiles work of direct relevance to Cambridge exams (e.g. by describing test revision projects or validation studies); it also contributes to and supports the wider field (e.g. by publishing quality research papers and doctoral dissertations on language testing/assessment).

Titles in the series will be of considerable interest to test users, language test developers and researchers. More information on all currently available volumes can be found at: www.cambridge.org/elt/catalogue/linguistics/silt

1 An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL comparability study
   Lyle F Bachman, Fred Davidson, Katherine Ryan, and Inn-Chull Choi (1995)

2 Test Taker Characteristics and Test Performance: A structural modeling approach
   Antony John Kunnan (1995)

3 Performance Testing, Cognition and Assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem
   Michael Milanovic and Nick Saville (Eds) (1996)

4 The Development of IELTS: A study of the effect of background knowledge on reading comprehension
   Caroline Clapham (1996)

5 Verbal Protocol Analysis in Language Testing Research: A handbook
   Alison Green (1998)

6 Multilingual Glossary of Language Testing Terms
   ALTE members (1998)

7 Dictionary of Language Testing
   Alan Davies, Annie Brown, Cathie Elder, Kathryn Hill, Tom Lumley, and Tim McNamara (1999)

8 Learner Strategy Use and Performance on Language Tests: A structural equation modeling approach
   James E Purpura (1999)

9 Fairness and Validation in Language Assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida
   Antony John Kunnan (Ed) (2000)

10 Issues in Computer-adaptive Testing of Reading Proficiency
    Micheline Chalhoub-Deville (Ed) (1999)

11 Experimenting with Uncertainty: Essays in honour of Alan Davies

12 An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes
    Cyril Weir, Yang Huizhong and Jin Yan (Eds) (2000)

13 The Equivalence of Direct and Semi-direct Speaking Tests
    Kieran O’Loughlin (2001)

14 A Qualitative Approach to the Validation of Oral Language Tests
    Anne Lazaraton (2002)

15 Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002.

16 A Modular Approach to Testing English Language Skills: The development of the Certificates in English Language Skills (CELS) examinations
    Roger Hawkey (2004)

17 Issues in Testing Business English: The revision of the Cambridge Business English Certificates
    Barry O’Sullivan (2006)


19 IELTS Collected Papers: Research in speaking and writing assessment
    Lynda Taylor and PeterFalvey (Eds) (2007)

20 Testing the Spoken English of Young Norwegians: A study of testing validity and the role of ‘smallwords’ in contributing to pupils’ fluency
    Angela Hasselgreen (2005)

21 Changing Language Teaching through Language Testing: A washback study
    Liying Cheng (2005)

22 The Impact of High-stakes Testing on Classroom Teaching: A case study using insights from testing and innovation theory
    Dianne Wall (2005)

23 Impact Theory and Practice: Studies of the IELTS test and Progetto Lingue 2000
    Roger Hawkey (2006)

    Alan Davies

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

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

Forthcoming volumes:

   Alan Davies

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

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