30th Annual Language Testing Research Colloquium

LTRC 2008

Zhejiang University, Hangzhou, China

23 – 28 June, 2008

Focusing on the Core:
Justifying the Use of Language Assessments to Stakeholders

Program and Abstracts

International Language Testing Association
Zhejiang University
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 30
Focus on the FCE and CAE Review Project

- The 2004-2008 FCE and CAE Review Project: historical context and perennial themes
- Using Structural Equation Modelling to facilitate the revision of high stakes testing: the case of CAE
- Introducing short themed texts into the CAE Reading paper
- Establishing the impact of reduced input and output length in FCE and CAE Writing
- Reviewing the FCE and CAE Listening and Speaking tests
- Developing revised assessment scales for Main Suite and BEC Speaking tests
- Overview of FCE and CAE Review Project research activity

Research Notes 31
Focus on testing reading

- A cognitive processing approach towards defining reading comprehension
- Applying a cognitive processing model to Main Suite Reading papers
- A corpus-informed study of specificity in Financial English: the case of ICFE Reading
- Exploring lexical differences in General English Reading papers
- Text organisation features in an FCE Reading gapped sentence task
- IELTS award news

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
# Table of Contents

LTRC 2008 Conference Organizers and Committee Members 2

LTRC 2008 Conference Sponsors 6

Message from the ILTA President 7

About ILTA 9
  - ILTA Goals 9
  - ILTA Executive Board 2008 9
  - ILTA Committees 2008 9
  - ILTA Institutional affiliates 10
  - ILTA Presidents 10

LTRC 2008 Program Overview 11

LTRC 2008 Full Program 14

2008 UCLES/ILTA Lifetime Achievement Award 27

Samuel J. Messick Memorial Lecture 29

Abstracts 31
  - Workshops 32
  - Symposia 38
  - Papers 45
  - Works-in-Progress 63
  - Posters 91

Presenters Index 118
LTRC 2008 Conference Organizers and Committee Members

LTRC 2008 Organizing Committee
Lyle Bachman, University of California, Los Angeles
Nathan Carr, California State University, Fullerton
Lianzhen He, Zhejiang University
Antony John Kunnan, California State University, Los Angeles
Huizhong Yang, Shanghai Jiaotong University
Yan Jin, Shanghai Jiaotong University

LTRC 2008 Program
Cover Design: Zhang Wang, Chinese University of Hong Kong
Program Design: Yasuhiro Imao, University of California, Los Angeles
Youngsoon So, University of California, Los Angeles
Huan Wang, University of California, Los Angeles

Proposal Reviewers

Priyanvada Abeywickrama
San Francisco State University
abeywick@sfsu.edu

Alison Bailey
University of California, Los Angeles
abailey@gseis.ucla.edu

Annie Brown
Ministry of Higher Education, UAE
a.brown@unimelb.edu.au

Carol Chapelle
Iowa State University
carloclastate.edu

Liying Cheng
Queen's University, Canada
liyinglecheng@queensu.ca

Andrew Cohen
University of Minnesota
adcohen@umn.edu

Alister Cumming
OISE, University of Toronto
acumming@oise.utoronto.ca

Fred Davidson
University of Illinois, Urbana-Champaign
fdg@uiuc.edu

Alan Davies
University of Edinburgh
adavies@ed.ac.uk

Chris Davison
University of Hong Kong
cdavison@hkuc.hku.hk

Dan Douglas
Iowa State University
dandouglas@iastate.edu

Catherine Elder
Monash University
caelder@unimelb.edu.au

Janna Fox
Carleton University
Janna_Fox@carleton.ca

Liz Hamp-Lyons
University of Hong Kong/University of Bedfordshire
ljhang@oise.utoronto.ca

Eunice Eunhee Jang
OISE, University of Toronto
dkenyon@cal.org

Dorry Kenyon
Center for Applied Linguistics
constant.leung@kcl.ac.uk

Constant Leung
King's College London
lorena.illosa@nyu.edu

Lorena Llosa
New York University
qiluxia@hotmail.com

Luxia Qi
Guangdong University of Foreign Studies
mmalone@cal.org

Margaret (Meg) Malone
Center for Applied Linguistics
ockey@umn.edu

Gary Ockey
University of New Mexico
jp248@columbia.edu

James Purpura
Teachers College, Columbia University
egdavid@polyu.edu.hk

David Qian
Hong Kong Polytechnic University
J.J.M.Schoonen@uva.nl

Rob Schoonen
University of Amsterdam
ysawaki@ets.org

Yasuyo Sawaki
Educational Testing Service
shin36@indiana.edu

Sang-Keun Shin
Ewha Woman's University
cstansfield@2Lti.com

Sunyoung Shin
Indiana University
Taylor.L@cambridgeesol.org

Charles W. Stansfield
Second Language Testing, Inc.
carolyn.turner@mcgill.ca

Lynda Taylor
University of Cambridge ESOL Examinations
miwolf@cse.ucla.edu

Carolyn Turner
McGill University
xxi@ets.org

Mikyung Wolf
CRESST, University of California, Los Angeles
yjin@mail.sjtu.edu.cn

Xiaoming Xi
Education Testing Service

Yan Jin
Shanghai Jiaotong University
Address of ILTA Business Office
ILTA Business Office
Prime-Management, Inc.
3416 Primm Lane
Birmingham, AL 35216 USA
(1) (205) 823-6106
ILTA@primemanagement.net (Robert Ranieri)

2008 UCLES/ILTA Lifetime Achievement Award Committee
Charles W Stansfield (Chair), Second Language Testing, Inc.
Jim Purpura, Teachers College, Columbia University
Nick Saville, University of Cambridge ESOL Examinations
Carolyn Turner, McGill University

2006 and 2007 ILTA Best Article Award Committees
2006 Committee
Kathryn Hill (chair), University of Melbourne
Christine Coombe, Dubai Men’s College, UAE
Ute Knoch, University of Auchland
Rob Schoonen, University of Amsterdam
Xiaoming Xi, Educational Testing Service

2007 Committee
Constant Leung (chair), King’s College London
Annie Brown, Ministry of Higher Education, UAE
Gary Buck, University of Michigan
Carol Chapelle, Iowa State University
Pauline Rea-Dickins, University of Bristol

2008 Selection Committee for the Robert Lado Award for Best Graduate Student Paper at LTRC
Yong-Won Lee (chair), Seoul National University
Liying Cheng, Queen’s University, Canada
Yasuyo Sawaki, Educational Testing Service

2008 ILTA Student Travel Award Committee
Glenn Fulcher (chair), University of Leicester
Lynda Taylor, University of Cambridge ESOL Examinations
Gary Buck, University of Michigan
2006, 2007 and 2008 Jacqueline A. Ross TOEFL Dissertation Award committees

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

2008 IELTS Masters Award Committee

The IELTS Joint Research Committee chaired by Nick Saville

The 2008 Selection Committee for the Spaan Fellowship for Studies in Second or Foreign Language Assessment

Jeff S. Johnson (chair), University of Michigan, ELI
Barbara Dobson, University of Michigan, ELI
India Plough, University of Michigan, ELI
Natalie Nordby Chen, University of Michigan, ELI
Sarah Van Bonn, University of Michigan, ELI
Eric Lagergren, University of Michigan, ELI

The 2008 ILTA Student Travel Award winners

Luke Harding, University of Melbourne
- Test-taker attitudes towards diverse-accented speakers in EAP listening assessment: The verbal guise approach

Jiyoung Kim, University of Illinois, Urbana-Champaign
- Justifying intended test effects to stakeholders by using effect arguments: A case of development and validation of an ESL writing test

The Jacqueline A. Ross TOEFL Dissertation Award winners and affiliations

2006

Eunice Eunhee Jang, University of Illinois at Urbana-Champaign
Dissertation Adviser – Dr Fred Davidson
Dissertation title: A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL

2007

Usman Erdosy, University of Toronto
Dissertation Adviser – Dr Alister Cumming
Dissertation title: Responding to non-native and native writers of English: A history professor’s indigenous criteria for grading and feedback in an undergraduate Sinology course
2008

Guoxing Yu, University of Bristol
Dissertation Adviser – Dr Pauline Rea-Dickins
Dissertation title: Towards a model of using summarization tasks as a measure of reading comprehension

The 2008 IELTS Masters Award winner and affiliation

Talia Isaacs, McGill University, Canada
Advisor – Dr Carolyn E Turner
Thesis title: Towards defining a valid assessment criterion on pronunciation proficiency in non-native English speaking graduate students

The 2008 Spaan Fellows

Beverly Baker, McGill University, Canada
The validity of the Ministère de l’éducation, du loisir et du sport (MELS)

Atta Gebril, United Arab Emirates University, United Arab Emirates, and Lia Plakans, University of Texas at Austin, USA
Investigating source use, discourse features, and process in integrated writing tests

Luke Harding and Kerry Ryan, University of Melbourne, Australia
Decision-making in making open-ended listening test items: Implications for the development of marking guides

Ute Knoch, University of Melbourne, Australia
Involving ESP stakeholders in rating scale validation: The case of the ICAO rating scale

Tian Song, Michigan State University, USA
Comparing dichotomous and polytomous IRT models in equating GCVR sections in ECPE
LTRC 2008 Conference Sponsors

The LTRC 2008 Organizing Committee would like to express its sincerest appreciation to the following publishers and organizations for their generous support of this year’s LTRC:

Cambridge ESOL Examinations
Cambridge University Press
Educational Testing Service
Encyclopedia of Language and Education (2nd Edition)
Foreign Language Teaching and Research Press
Language Assessment Quarterly
The National College English Testing Committee of China
Oxford University Press
Pearson Language Assessments
Second Language Testing International
Shanghai Foreign Language Education Press
The University of Michigan English Language Institute
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 30th Annual Language Testing Research Colloquium in Hangzhou, China. The theme of this year’s conference is: “Focusing on the core: Justifying the use of language assessments to stakeholders.” Some 228 proposals were submitted for review. Of these, 3 symposia, 24 papers, 23 posters and 23 works-in-progress were selected. This year’s LTRC offers 5 pre-conference workshops: one on language test development, one on learning-oriented language assessment in the schools, one on basic statistics for language assessment, and two more advanced statistics for language assessment (G-theory and latent growth modeling). Finally, this year the Samuel J. Messick Memorial Lecture, sponsored by Educational Testing Service, will be delivered by J.D. Brown. JD will question why stakeholders don’t seem to cooperate with language testers. In response to this question, he will argue that language testers need to create “stakeholder-friendly” tests if they wish to provide defensible testing rooted in current theories of validity.

As we have done in years past, ILTA will recognize the many achievements of our members at LTRC. Our most prestigious award is the Lifetime Achievement Award. This award acknowledges long-established leaders who have distinguished service and scholarship records in the field of language assessment. This year, ILTA, together with the Cambridge University ESOL Exams, is proud to honor the work of J. Charles Alderson, Professor of Linguistics and English Language Education at the University of Lancaster. In accepting this award, Charles will be speaking to us on the topic of “A lifetime of tests” in which he will share with us his own personal experience with tests and testing over the last five decades. At LTRC 2008 ILTA will also present its Best Paper in Language Testing Award—one for 2006 and one for 2007. Finally, to promote excellent scholarship and exemplary presentation skills from students in language assessment, ILTA will present the 2008 Robert Lado Award for Best Graduate Student Presentation at LTRC.

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 volunteers at LTRC. In this effort, ILTA is proud to provide support for two students to travel to LTRC. The ILTA 2008 Student Travel Award is given to those students with the most promising paper abstracts. The 2008 awardees are Luke Harding from the University of Melbourne and Jiyoung Kim from the University of Illinois.

I am delighted to inform you of the success of this year’s AAAL/ILTA Joint symposium held at AAAL in Washington, DC. The purpose of this symposium is for researchers from language testing and the broader field of applied linguistics to “explore areas of common interest, both substantive and methodological, from a variety of perspectives” (Bachman & Cohen, 1998, p. xi). This year, Tim McNamara of the University of Melbourne and Nick Saville of Cambridge ESOL organized the symposium entitled: “Issues of language acquisition and assessment as related to migration and citizenship.” Besides the organizers, presenters were Piet Van Avermaet, Tzahi Kanza, Antony Kunnan, Joe LoBianco, and Elana Shohamy.

I’m also pleased to inform you that LTRC 2009 will be held in Denver right before AAAL (March 21-24, 2009). We will have a day of LTRC/AAAL overlap, and the LTRC 2009 organizers have been working with the organizers of AAAL 2009 to coordinate efforts around testing and assessment.
ILTA is moving forward on many fronts. In addition to the Code of Ethics, we now have an organization-approved Code of Practice, thanks to Alan Davies and his committee. The Executive Board has spent much of this year putting together a set of Standard Operation Procedures (SOPs) for the organization. These SOPs aim to codify how we function as an organization—no more reinventing wheels. Also, ILTA will soon launch its newly designed, interactive website, thanks to the efforts of Dan Douglas and Erik Voss. One goal of the new website is to provide an archival documentation of ILTA as an organization and of LTRC as a venue for scholarly exchange from their beginnings. Finally, ILTA is now in its second year with Prime Management, our managing company. They have helped us maintain a stable membership of nearly 200 people from countries all over the world. On-line membership has made it easier to join and to renew. If you are at LTRC and not an ILTA member, I encourage you to join. If you are a member, please encourage your colleagues and, if relevant, 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 behalf of ILTA, I would first like to thank the many, many individuals who have volunteered to serve on ILTA committees this year—a special word of thanks to the chairs of the committees and to the Executive Board, who have worked tirelessly all year on a full range of ILTA projects.

Last but not least, I would like to express a sincere word of gratitude to the Organizers of this year’s LTRC in Hangzhou. For those of you who have not organized an LTRC, the time, effort, and patience needed to organize a conference of over 200 people from around the world is no small feat. In particular, I would like to thank Lyle Bachman, Antony Kunnan and Nathan Carr, who have organized many LTRCs including the present one in Hangzhou. I’d also like to thank the other members of the 2008 organizing committee: He Lianzhen, Yang Hiuzhong and Jin Yan. And I’d like to thank Xiaoming Xi, the ILTA secretary and liaison to LTRC 2008, for her work on the conference. Thank you all for your hard work in making LTRC Hangzhou a great success.

On that note, I’d like to wish all LTRC 2008 attendees a thoroughly enjoyable, culturally enriching and intellectually stimulating conference in Hangzhou.

With very best wishes,

Jim Purpura
President of ILTA 2008
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 2008

President: Jim Purpura, Teachers College, Columbia University (2007-2008)
Vice President: Carolyn Turner, McGill University, Canada (2007-2008)
Ex-officio member: Glenn Fulcher, University of Leicester, UK (2007-2008)
Secretary: Xiaoming Xi, Educational Testing Service (2008-2010)

Members-at-large 2008:
Lynda Taylor, Consultant to University of Cambridge ESOL Examinations (2007-2008)
Gary Buck, University of Michigan (2008-2009)
Rob Schoonen, University of Amsterdam, Netherlands (2008-2009)

ILTA Committees 2008

ILTA Newsletter 2008

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, Temple University

ILTA Archivists 2008

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

ILTA Task Force on Testing Standards Update 2008

Samira ELAtia, Chair, University of Alberta
Fred Davidson, University of Illinois, Champaign-Urbana
Alexis A. Lopez, Universidad de los Andes, Columbia
Ana Oscoz, University of Maryland, Baltimore County
Paul Jaquith, Ministry of Education, United Arab Emirates
ILTA Institutional affiliates

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

ILTA Presidents

1992-3  Charlie Stansfield
1993-4  Charles Alderson
1995    Bernard Spolsky
1996    Lyle Bachman
1997    John Clark
1998    Tim McNamara
1999    Elana Shohamy
2000    Alan Davies
2001    Caroline Clapham
2002    Fred Davidson
2004    Antony Kunnan
2003    Liz Hamp-Lyons
2005    Dan Douglas
2006    Glenn Fulcher
2007-8  James Purpura
## LTRC 2008 Program Overview

### WORKSHOPS - MONDAY 23 JUNE

<table>
<thead>
<tr>
<th>Time</th>
<th>Workshop</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 5:00</td>
<td>Workshop 1: An introduction to latent growth modeling techniques using SEM</td>
<td>Room 216, East 6</td>
</tr>
<tr>
<td>9:00 - 5:00</td>
<td>Workshop 2: Applying school-based assessment for learning in a formal examination system</td>
<td>Room 218, East 6</td>
</tr>
</tbody>
</table>

### WORKSHOPS - TUESDAY 24 JUNE

<table>
<thead>
<tr>
<th>Time</th>
<th>Workshop</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 5:00</td>
<td>Workshop 3: What teachers need to know about developing useful assessments for language learners</td>
<td>Room 216, East 6</td>
</tr>
<tr>
<td>9:00 - 5:00</td>
<td>Workshop 4: Basic statistics for language assessment</td>
<td>Room 218, East 6</td>
</tr>
<tr>
<td>9:00 - 5:00</td>
<td>Workshop 5: The application of Generalizability theory in L2 assessment</td>
<td>Room 220, East 6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:00 - 9:00</td>
<td>LTRC Opening Reception (Sponsored by Cambridge ESOL)</td>
</tr>
</tbody>
</table>

### LTRC DAY 1 - WEDNESDAY 25 JUNE

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 - 6:00</td>
<td>Registration</td>
</tr>
<tr>
<td>9:00 - 6:00</td>
<td>Publishers’ Exhibits</td>
</tr>
<tr>
<td>8:45 - 9:00</td>
<td>Welcome to LTRC 2008</td>
</tr>
<tr>
<td>9:00 - 9:15</td>
<td>Introduction of the Messick Memorial Lecturer</td>
</tr>
<tr>
<td>9:15 - 10:15</td>
<td>Samuel J. Messick Memorial Lecture</td>
</tr>
<tr>
<td>10:15 - 10:25</td>
<td>Presentation of the Messick Award</td>
</tr>
<tr>
<td>10:25 - 10:50</td>
<td>BREAK</td>
</tr>
<tr>
<td>10:50 - 12:20</td>
<td>Paper Session 1</td>
</tr>
<tr>
<td>12:20 - 1:50</td>
<td>LUNCH BREAK</td>
</tr>
<tr>
<td>1:50 - 3:50</td>
<td>Symposium 1: Interacting with stakeholders on language tests: Issues, problems and challenges</td>
</tr>
<tr>
<td>3:50 - 4:00</td>
<td>BREAK</td>
</tr>
<tr>
<td>4:00 - 5:00</td>
<td>POSTER SESSION 1</td>
</tr>
<tr>
<td>8:00 - 6:00</td>
<td>Registration</td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>9:00 - 6:00</td>
<td>Publishers’ Exhibits</td>
</tr>
<tr>
<td>8:30 - 10:00</td>
<td>Paper Session 2</td>
</tr>
<tr>
<td></td>
<td>Monitoring validity claims in a high-stakes setting</td>
</tr>
<tr>
<td></td>
<td>Whose standards is it anyway? Theoretical perspectives on defining test standards in language testing</td>
</tr>
<tr>
<td></td>
<td>From Code of Practice to Standards: Auditing the quality profile</td>
</tr>
<tr>
<td>10:00 - 10:20</td>
<td>BREAK</td>
</tr>
<tr>
<td>12:20 - 1:50</td>
<td>LUNCH BREAK</td>
</tr>
<tr>
<td>1:50 - 3:20</td>
<td>Paper Session 3</td>
</tr>
<tr>
<td></td>
<td>Justifying intended test effects to stakeholders by using effect arguments: A case of development and validation of an ESL writing test</td>
</tr>
<tr>
<td></td>
<td>Large-scale oral proficiency assessment: From the perspective of stakeholders</td>
</tr>
<tr>
<td></td>
<td>Engaging stakeholders in the development of a comprehension test for an indigenous language</td>
</tr>
<tr>
<td>3:20 - 3:30</td>
<td>BREAK</td>
</tr>
<tr>
<td>3:30 - 4:30</td>
<td>POSTER SESSION 2</td>
</tr>
<tr>
<td></td>
<td>Local opera (Sponsored by Foreign Language Teaching and Research Press)</td>
</tr>
<tr>
<td>LTRC DAY 3 - FRIDAY 27 JUNE</td>
<td></td>
</tr>
<tr>
<td>8:00 - 6:00</td>
<td>Registration</td>
</tr>
<tr>
<td>9:00 - 6:00</td>
<td>Publishers’ Exhibits</td>
</tr>
<tr>
<td>8:30 - 10:00</td>
<td>Paper Session 4</td>
</tr>
<tr>
<td></td>
<td>L2 test takers’ cognitive load as a potential rebuttal resource</td>
</tr>
<tr>
<td></td>
<td>An investigation of the relationships of cognitive and metacognitive strategy use to GEPT test performance</td>
</tr>
<tr>
<td></td>
<td>Language assessment and the professional integration of immigrant physicians: A successful model</td>
</tr>
<tr>
<td>10:00 - 10:20</td>
<td>Break</td>
</tr>
<tr>
<td>10:20 - 12:20</td>
<td>Symposium 3: The ‘teacher factor’ in washback contexts: Evidence from mixed methods research in Canada, China and Malaysia</td>
</tr>
<tr>
<td>12:20 - 1:50</td>
<td>LUNCH BREAK</td>
</tr>
<tr>
<td>1:50 - 3:20</td>
<td>Paper Session 5</td>
</tr>
<tr>
<td></td>
<td>Investigating secondary school ELL students’ literacy skill profiles using cognitive diagnosis modeling</td>
</tr>
<tr>
<td></td>
<td>Criterion-related validity of the TOEFL iBT Listening section</td>
</tr>
<tr>
<td></td>
<td>Validating the use of a single rating scale to assess heritage and non-heritage learners’ writing ability: A triangulation approach</td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>3:20 - 3:30</td>
<td>BREAK</td>
</tr>
<tr>
<td>3:30 - 4:30</td>
<td>Works-In-Progress SESSION 1</td>
</tr>
<tr>
<td>5:30 - 7:30</td>
<td>STUDENT ORGANIZED SESSION</td>
</tr>
<tr>
<td></td>
<td>Visit to the Tea Village (Sponsored by Pearson Language Tests)</td>
</tr>
<tr>
<td></td>
<td>LTRC DAY 4 - SATURDAY 28 JUNE</td>
</tr>
<tr>
<td>8:00 - 6:00</td>
<td>Registration</td>
</tr>
<tr>
<td>9:00 - 6:00</td>
<td>Publishers’ Exhibits</td>
</tr>
<tr>
<td>8:00 - 8:15</td>
<td>Introduction of the Lifetime Achievement Awardee</td>
</tr>
<tr>
<td>8:15 - 9:15</td>
<td>UCLES/ILTA Lifetime Achievement Awardee Lecture</td>
</tr>
<tr>
<td></td>
<td>A lifetime of testing J. Charles Alderson</td>
</tr>
<tr>
<td>9:15 - 10:45</td>
<td>Paper Session 6</td>
</tr>
<tr>
<td></td>
<td>Investigating the performance of Indian raters in scoring TOEFL® iBT</td>
</tr>
<tr>
<td></td>
<td>Speaking and the effectiveness of a special training designed for</td>
</tr>
<tr>
<td></td>
<td>scoring Indian speakers</td>
</tr>
<tr>
<td></td>
<td>Test-taker attitudes towards diverse-accented speakers in EAP</td>
</tr>
<tr>
<td></td>
<td>listening assessment: The verbal guise approach</td>
</tr>
<tr>
<td></td>
<td>Assessing the speaking proficiency of health professionals: The</td>
</tr>
<tr>
<td></td>
<td>nurses’ criteria</td>
</tr>
<tr>
<td>10:45 - 11:00</td>
<td>BREAK</td>
</tr>
<tr>
<td>11:00 - 12:30</td>
<td>Paper Session 7</td>
</tr>
<tr>
<td></td>
<td>Examining the effect of response format type on construct validity</td>
</tr>
<tr>
<td></td>
<td>to inform score uses of an EFL listening comprehension test</td>
</tr>
<tr>
<td></td>
<td>Effects of presentation mode and computer familiarity on performance</td>
</tr>
<tr>
<td></td>
<td>in reading-to-summarize extended English texts in two languages</td>
</tr>
<tr>
<td></td>
<td>A meta-analysis of effects of multiple-choice and open-ended formats</td>
</tr>
<tr>
<td></td>
<td>on reading and listening test performance</td>
</tr>
<tr>
<td>12:30 - 2:00</td>
<td>LUNCH BREAK</td>
</tr>
<tr>
<td>2:00 - 3:30</td>
<td>Paper Session 8</td>
</tr>
<tr>
<td></td>
<td>The characteristics of academic English language and DIF on items</td>
</tr>
<tr>
<td></td>
<td>in an English language proficiency test and standardized tests of</td>
</tr>
<tr>
<td></td>
<td>math and science</td>
</tr>
<tr>
<td></td>
<td>Learning to write essays or teaching to crack tests: An empirical</td>
</tr>
<tr>
<td></td>
<td>study investigating the washback and impact of TOEFL test-preparation</td>
</tr>
<tr>
<td></td>
<td>in China</td>
</tr>
<tr>
<td></td>
<td>Impact and consequences of school-based assessment in Hong Kong:</td>
</tr>
<tr>
<td></td>
<td>Views from students and their parents</td>
</tr>
<tr>
<td>3:30 - 3:40</td>
<td>BREAK</td>
</tr>
<tr>
<td>3:40 - 4:40</td>
<td>Works-In-Progress SESSION 2</td>
</tr>
<tr>
<td>6:30 – 7:00</td>
<td>LTRC 2008 GROUP PHOTOGRAPH</td>
</tr>
<tr>
<td>7:00 – 11:00</td>
<td>Banquet &amp; Awards Presentation</td>
</tr>
</tbody>
</table>
## LTRC 2008 Full Program

### WORKSHOPS
**MONDAY 23 JUNE**

<table>
<thead>
<tr>
<th>Time</th>
<th>Workshop</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 5:00</td>
<td><strong>WorkShop, Day 1</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 1: An introduction to latent growth modeling techniques using SEM</strong></td>
<td><strong>George A. Marcoulides, University of California, Riverside</strong>&lt;br&gt;<strong>Location:</strong> Room 216, East 6</td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 2: Applying school-based assessment for learning in a formal examination system</strong></td>
<td><strong>Chris Davison, University of Hong Kong</strong>&lt;br&gt;<strong>Liz Hamp-Lyons, Universities of Hong Kong</strong>&lt;br&gt;<strong>with the assistance of Gerry Davies and Nicole Tavares, University of Hong Kong</strong>&lt;br&gt;<strong>Location:</strong> Room 218, East 6</td>
</tr>
</tbody>
</table>

### WORKSHOPS
**TUESDAY 24 JUNE**

<table>
<thead>
<tr>
<th>Time</th>
<th>Workshop</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 5:00</td>
<td><strong>WorkShop, Day 2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 3: What teachers need to know about developing useful assessments for language learners</strong></td>
<td><strong>Adrian Palmer, University of Utah</strong>&lt;br&gt;<strong>MaryAnn Christenson, University of Utah</strong>&lt;br&gt;<strong>Location:</strong> Room 216, East 6</td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 4: Basic statistics for language assessment</strong></td>
<td><strong>Gary J. Ockey, University of New Mexico</strong>&lt;br&gt;<strong>Lorena Llosa, New York University</strong>&lt;br&gt;<strong>Location:</strong> Room 218, East 6</td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 5: The application of Generalizability theory in L2 assessment</strong></td>
<td><strong>Yasuyo Sawaki, Educational Testing Service, Princeton, New Jersey</strong>&lt;br&gt;<strong>Xiaoming Xi, Educational Testing Service, Princeton, New Jersey</strong>&lt;br&gt;<strong>Location:</strong> Room 220, East 6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:00 - 9:00</td>
<td><strong>LTRC Opening Reception (Sponsored by Cambridge ESOL)</strong></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>8:30 - 10:25</td>
<td>Opening Session at Zhejiang University Conference Center</td>
</tr>
<tr>
<td>8:45 - 9:00</td>
<td>Welcome on behalf of Zhejiang University</td>
</tr>
<tr>
<td>9:00 - 9:15</td>
<td>Introduction of the Messick Memorial Lecturer</td>
</tr>
<tr>
<td>9:15 - 10:15</td>
<td>Presentations by Toru Kinoshita, Nagoya University</td>
</tr>
<tr>
<td>10:15 - 10:25</td>
<td>Presentation of the Messick Award: Xiaoming Xi, Educational Testing Service, Princeton, New Jersey</td>
</tr>
<tr>
<td>10:25 - 10:50</td>
<td>BREAK</td>
</tr>
<tr>
<td>10:50 - 12:20</td>
<td>Paper Session 1, Chair: David Qian, at Zhejiang University Conference Center</td>
</tr>
<tr>
<td>8:30 - 10:25</td>
<td>Using summative assessments for formative purposes: The ultimate justification for learners and teachers</td>
</tr>
<tr>
<td>10:25 - 12:20</td>
<td>LUNCH BREAK</td>
</tr>
<tr>
<td>12:20 - 1:50</td>
<td>Symposium 1 at Room 227, East 6</td>
</tr>
<tr>
<td>1:50 - 3:50</td>
<td>Interacting with stakeholders on language tests: Issues, problems and challenges</td>
</tr>
<tr>
<td>3:50 - 4:00</td>
<td>BREAK</td>
</tr>
</tbody>
</table>

**LTRC DAY 1: WEDNESDAY 25 JUNE**

Registration: 8:00 a.m. - 6:00 p.m.
Publishers' Exhibits: 9:00 - 6:00 p.m.

**Day 1**
A construct validation research of the writing task in NMET (GD) — Text-based English writing for Chinese EFL learners
Xinling Zhang, School of Foreign Languages, Shanghai University; Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies

A different topic, a different test? Investigating stakeholder concerns about the fairness of a writing test
Barbara Dobson and Gad Lim, University of Michigan

A dual-testlet approach to student placement
Toshihiko Shiotzu, Kurume University

Adding flexibility to automated scoring of low-level ESL essays
Diane Strong-Krause, Brigham Young University

Assessing the degree of knowledge in vocabulary use
Yuko Morimoto, University of Tsukuba

Assessing young EFL learners in China: An introduction of a new test — Tsinghua Children’s English Proficiency Test
Wenxia Zhang, Meihua Liu and Xiangjun Deng, Tsinghua University

Automatic assessment of spoken English skills in aviation professionals
Masanori Suzuki, Alistair Van Moere and Ryan Downey, Ordinate Corporation

Cognitive processing and test validity in the IELTS academic reading test
Tony Green, University of Bedfordshire

Construction and validation of assessor-centered rating scale in Chinese
So-young Jang, Kadeessa Abdul-Kadir, Youngshin Chi, Hui-Lien Hsu, Jiyoung Kim and Chih-kai Lin, University of Illinois at Urbana-Champaign

Describing test taker performance in relation the Common European Framework of Reference: How should it be justified to stakeholders?
Spiros Papageorgiou, University of Michigan

Development of a large scale placement test
Yuji Nakamura, Keio University

Dimensionality and the construct of language proficiency
John H.A.L. DE JONG, Pearson Language Assessments

Feedback for feedforward: Understanding teacher feedback to student stakeholders in assessment for learning
Geraldine Davies, The University of Hong Kong

Informing ability grouping in EFL listening with a placement test
Hongwen Cai, Guangdong University of Foreign Studies

Investigating the effects of reading and writing on summaries written by L2 writers
Yasuhiro Imao, University of California, Los Angeles

Issues in developing a model for verbatim translation exams
Maria Brau and Rachel Brooks, Language Testing and Assessment Unit, Language Services Section, Federal Bureau of Investigation

Justifying test design decisions to stakeholders using a consensus log: A case of an argument-based approach to test development
Jiyoung Kim, So-young Jang, Kadeessa Abdul-Kadir, Youngshin Chi, Hui-Lien Hsu and Chih-kai Lin, University of Illinois at Urbana-Champaign

Language, culture and the California High School Exit Examination (CAHSEE)
Liberty Van Natten, University of California, Davis
Rater variation in web-based assessment of KFL learners’ pragmatics using FACETS analysis
   Soo Jung Youn, University of Hawai‘i

Relationships between lexical access and selection of choices in multiple-choice vocabulary tests
   Chikako Nakagawa, Graduate School, University of Tsukuba

The influence of partner proficiency in a paired oral assessment task.
   Larry Davis, University of Hawai‘i, Manoa

The role of alignment in the content validation of a large-scale, standards-based English language proficiency test: Background, methodology and applications to other contexts
   Carsten Wilmes and Gary H. Cook, University of Wisconsin-Madison

Variability and validity of automated essay scores for ESL writers
   Yong-Won Lee, Seoul National University

Washback or fairness? Insights from the criterion-irrelevant factors in the assessment of Cambridge YLE Speaking Test
   Hongtao Ren and Binbin Dai, The Chinese University of Hong Kong

Tour of the Westlake (Sponsored by Shanghai Foreign Language Education Press)
8:30 - 10:00  **Paper Session 2** at Room 227, East 6

4. **Monitoring validity claims in a high-stakes setting**  
   J. Charles Alderson, *Lancaster University* and Jayanti Banerjee, *University of Michigan*

5. **Whose standards is it anyway? Theoretical perspectives on defining test standards in language testing**  
   Samira El Atia, *University of Alberta*

6. **From Code of Practice to Standards: Auditing the quality profile**  
   Nick Saville, *University of Cambridge ESOL*

10:00 - 10:20  **BREAK**

10:20 - 12:20  **Symposium 2** at Room 227, East 6

**Canvassing stakeholder perspectives: The key to test usefulness?**  
Cathie Elder, Organizer

Cathie Elder, *University of Melbourne*, “Language standards: How are they viewed by the ELT industry?”


Ute Knoch, *University of Melbourne*, “An examination of stakeholder views of diagnostic feedback to EAP writing performance”


Discussant: David Nevo, *Tel Aviv University*  
Convenor: Cathie Elder, *University of Melbourne*

12:20 - 1:50  **LUNCH BREAK**  
Language Assessment Quarterly EAB meeting *(Location To Be Announced)*

1:50 - 3:20  **Paper Session 3** at Room 227, East 6

7. **Justifying intended test effects to stakeholders by using effect arguments: A case of development and validation of an ESL writing test**  
   Jiyoung Kim, *University of Illinois, Urbana-Champaign*

8. **Large-scale oral proficiency assessment: From the perspective of stakeholders**  
   Weiping Wu, *Chinese University of Hong Kong*

9. **Engaging stakeholders in the development of a comprehension test for an indigenous language**  
   Gillian Wigglesworth, *University of Melbourne*

3:20 - 3:30  **BREAK**
A discourse-analytic investigation of topic-related performance variation on a high-stakes writing test
Barbara Dobson, Aaron Ohlrogge and Sarah Van Bonn, University of Michigan

A comparability study on the old and the new CET-4
Qin Liu and Zhenhua Wang, University of Shanghai for Science and Technology

A critical inquiry to academic language constructs across U.S. states
Mark Nigolian, St. Michael’s College

A L2 fluency rating scale construction project - Challenges and suggestions
Huei-Lien Hsu, Kadeessa Kadir, So-young Jang and Youngshin Chi, University of Illinois at Urbana-Champaign

A quantitative study on subject-verb agreement in English
Chih Yuan Cho, California State University, Los Angeles

“Assessment is important because it helps me to learn”: How are attitudes and actions related among L1 and FL learners?
Ari Huhta, Mirja Tarnanen and Reeta Neittaanmäki, University of Jyväskylä

An application of multidimensionality model-based DBF/DIF procedures to the Ontario Secondary School Literacy Test
Youn-Hee Kim and Eunice Jang, OISE / University of Toronto

Are teachers teaching to the test? A case study of the College English Test (CET) in China
Hongli Li, Pennsylvania State University

Assessing reading comprehension with graphic organizers
Xiangying Jiang, West Virginia University

Assessment for learning- Providing diagnostic feedback and language enhancement guidance to test-takers
Eva Lui, City University of Hong Kong

Miyoko Kobayashi, Nobuko Hasegawa, and Nahomi Machida, Kanda University of International Studies

Beyond CTT and IRT: An interactional testing model for writing performance assessments
Xin Wang, University of Illinois at Urbana Champaign

Bridging the gender gap in computerized speaking test performance
Eunjyu Yu, The Ohio State University

Chinese EFL learners’ perceptions of their experiences with a portfolio-based writing assessment project: A case study
Qinghua Li, Shaoxing University

Developing online tests for advanced language assessment
Martyn Clark, Center for Applied Second Language Studies

Investigation of the Tunghai English Placement Examination using factor analysis
Jonathan Whittinghill, California State University, Los Angeles

Language assessment for the business processing outsourcing industry
Gail Forey and Jane Lockwood, Hong Kong Institute of Education

Perceptions on English academic writing rubrics: A comparison between Chinese teacher educators and Chinese pre-service teachers
Danbin Wang, Shanghai Normal University and Ye He, University of North Carolina at Greensboro

TBEST: A new on-line English speaking test in China
Jun Liu, University of Arizona
The assessment use argument: A web-based Spanish placement listening exam
   Cristina Pardo, Iowa State University

The washback effect of a high stakes assessment on EFL teachers’ teaching strategies and assessment practices in Taipei City
   Yu-Ching Chan, Graduate School of Children English Education, National Taipie University of Education

Using performance descriptors to develop a criterion-referenced test of language proficiency: An exploratory study
   Gary Buck, University of Michigan

Where do teachers go wrong in item writing?
   Nathan Carr, California State University, Fullerton

Local opera (Sponsored by Foreign Language Teaching and Research Press)
8:30 - 10:00  Paper Session 4 at Room 227, East 6

10. **L2 test takers’ cognitive load as a potential rebuttal resource**  
   Toru Kinoshita, *Nagoya University* and Harumi Oishi, *Gifu Shotoku Gakuen University*

11. **An investigation of the relationships of cognitive and metacognitive strategy use to GEPT test performance**  
   Jessica Wu, *Language Training and Testing Centre, Taiwan*

12. **Language assessment and the professional integration of immigrant physicians: A successful model**  
   David Watt, *University of Calgary*

10:00 - 10:20  BREAK

10:20 - 12:20  Symposium 3 at Room 227, East 6

The ‘teacher factor’ in washback contexts: Evidence from mixed methods research in Canada, China and Malaysia
   Carolyn Turner, Organizer

   Carolyn E. Turner, *McGill University*, “The specificity of the “research approach” in classroom studies: Probing the predictability of washback through teacher conceptual and instrumental evidence in Quebec high schools”

   Jing Wang, *McGill University*, “Identifying the main impact on teacher classroom behaviour: Examining the effects of the CET, teacher beliefs, teacher knowledge bases and teacher proficiency”

   May Tan, *McGill University*, “Bilingual high-stakes mathematics and science exams in Malaysia: Pedagogical and linguistic issues”

   Discussant: Liying Cheng, *Queen’s University*

12:20 - 1:50  LUNCH BREAK
   Language Testing EAB meeting *(Location To Be Announced)*

1:50 - 3:20  Paper Session 5 at Room 227, East 6

13. **Investigating secondary school ELL students’ literacy skill profiles using cognitive diagnosis modeling**  
   Eunice Jang and Youn-Hee Kim, *OISE/University of Toronto*

14. **Criterion-related validity of the TOEFL iBT Listening section**  
   Yasuyo Sawaki and Susan Nissan, *Educational Testing Service, Princeton, New Jersey*

15. **Validating the use of a single rating scale to assess heritage and non-heritage learners’ writing ability: A triangulation approach**  
   Nathan Carr and Jinghui Liu, *California State University, Fullerton*

3:20 - 3:30  BREAK
3:30 - 4:30  Works-In-Progress SESSION 1 at Room 118, East 6

A validation study of the use of STEP EIKEN Test as an alternative English proficiency screening test for college admission purposes
Yao Hill, University of Hawai’i at Manoa

Assessing EFL university students’ argumentative writing ability from a discourse perspective
Jingjing Qin, Northern Arizona University

Assessing productive collocational competence of advanced EFL learners
Su-mi Han, Seoul National University

Assessing vocabulary and lexical proficiency in EFL teacher education programmes
Melissa Baffi-Bonvino and Douglas Consolo, UNESP-Brazil

Building validity argument in an oral proficiency assessment project
Raii Hilden, University of Helsinki, Finland

Challenges in developing comparable tests in a multilingual India
Vyjayanthi Sankar, Educational Initiatives Pvt Ltd

Collaboration and student anxiety in the group interaction task of school-based assessment
Manman Gao, The University of Hong Kong

Creating an end-to-end assessment solution at the Hong Kong Institute of Education
Jane Lockwood and Michelle R. Raquel, Hong Kong Institute of Education

Developing reference level descriptors for higher proficiency learners of English
Tony Green, University of Bedfordshire

Development and initial validation of a questionnaire measuring students’ perception of the test impact
Hongli Li and Hoi K. Suen, Pennsylvania State University and Qi Zhong, Central University of Finance and Economics

Examining washback in the Taiwanese context: Voices from teachers and textbook writers
Ching-Ni Hsieh, Michigan State University

From placement test to post-test: Students’ views, teachers’/raters’ perspectives, test developers’/administrators’ rationale and institutional demands
Claudia Kunschak and Lily Zhang, Shantou University

Investigating learning in higher education through the lens of assessment
King Siong Lee, Universiti Kebangsaan Malaysia

Is an English language learner ready to take a test in English? The relationship between English proficiency and performance on content assessments
Youngsoon So, University of California, Los Angeles

Learning for assessment for learning: A case study
Ying Yu, University of Hong Kong

Looking beyond rater performance: Exploring the role of knowledge and beliefs in raters’ decision-making
Jie Zhang, Guangdong University of Foreign Studies

Mixed-method designs: Is it still early for language testers?
Heejeong Jeong, University of Illinois at Urbana-Champaign

Tests as productive learning opportunities
David Carless, University of Hong Kong

The effects of EFL teachers’ qualification examination on pre-service EFL teacher training in China
Luxin Yang and Baocheng Han, Beijing Foreign Studies University
Validating steps to English Proficiency (STEP) for assessing ELL students’ English language development
   Eunice Jang, Jim Cummins and Christian Chun, OISE / University of Toronto

Washback of CET-4 on students’ learning: Implications for test validity
   Qin Xie, University of Hong Kong

Which Englishes for English language tests? A conversation
   Liz Hamp-Lyons, University of Bedfordshire and Alan Davies, University of Edinburgh

5:30 - 7:30  STUDENT ORGANIZED SESSION (Location To Be Announced)

Visit to the Dragon Well Tea Village (Sponsored by Pearson Language Tests)
LTRC DAY 4: SATURDAY 28 JUNE
Registration: 8:00 a.m. - 6:00 p.m.
Publishers' Exhibits: 9:00 - 6:00 p.m.

8:00 - 9:15 UCLES/ILTA Lifetime Achievement Awardee Lecture (Location To Be Announced)

8:00 - 8:15
Introduction of UCLES/ILTA Lifetime Achievement Awardee
Lyle F. Bachman, University of California, Los Angeles

8:15 - 9:15
UCLES/ILTA Lifetime Achievement Awardee Lecture: A lifetime of testing
J. Charles Alderson, Lancaster University

9:15 - 10:45 Paper Session 6 at Room227, East 6

16. Investigating the performance of Indian raters in scoring TOEFL® iBT Speaking and the effectiveness of a special training designed for scoring Indian speakers
Xiaoming Xi, Educational Testing Service, Princeton, New Jersey

17. Test-taker attitudes towards diverse-accented speakers in EAP listening assessment: The verbal guise approach
Luke Harding, University of Melbourne

18. Assessing the speaking proficiency of health professionals: The nurses’ criteria
Kerry Ryan, University of Melbourne

10:45 - 11:00 BREAK

11:00 - 12:30 Paper Session 7 at Room227, East 6

19. Examining the effect of response format type on construct validity to inform score uses of an EFL listening comprehension test
Huan Wang, University of California, Los Angeles

20. Effects of presentation mode and computer familiarity on performance in reading-to-summarize extended English texts in two languages
Guoxing Yu, University of Bristol

21. A meta-analysis of effects of multiple-choice and open-ended formats on reading and listening test performance
Yo In'nami, Kanda University of International Studies and Rie Koizumi, Tokiwa University

12:30 - 2:00 LUNCH BREAK
LTRC/ILTA Annual Business Meeting (Lunch provided) (Location To Be Announced)

2:00 - 3:30 Paper Session 8 at Room227, East 6

22. The characteristics of academic English language and DIF on items in an English language proficiency test and standardized tests of math and science
Lyle Bachman, Applied Linguistics/University of California, Los Angeles and Tim Farnsworth, Mikyung Kim Wolf, Sandy Chang, Patina Bachman and Seth Leon, CRESST/University of California, Los Angeles

23. Learning to write essays or teaching to crack tests: An empirical study investigating the washback and impact of TOEFL test-preparation in China
Michael Chau, Hong Kong Polytechnic University
24. Impact and consequences of school-based assessment in Hong Kong: Views from students and their parents
   Liying Cheng, Queen's University, Ying Yu, University of Hong Kong, and Stephen J. Andrews, University of Hong Kong

3:30 - 3:40 BREAK

3:40 - 4:40 Works-In-Progress SESSION 2 at Room 118, East 6

Evaluation of the foreign language requirement at Duke University
   Carolyn Lee, Clare Tufts and Ingeborg Walther, Duke University

A test of cognitive academic language proficiency in English and Xhosa, an indigenous South African language: Justification for use in terms of the aims of bilingual education and international standards on test equivalence
   Elize Koch, University of the Western Cape and John Landon, University of Edinburgh

Bridging language tests and test-takers: A social cognitive investigation
   Ying Zheng, Queen's University

Can teachers become testers? And testers, teachers?
   Rama Mathew, Delhi University, Delhi

Development and evaluation of integrated language tasks
   Robb McCollum, Brigham Young University

Development of an automated oral reading fluency assessment
   Ryan Downey, Alistair Van Moere, Masanori Suzuki and David Rubin, Ordinate Corporation

Do changes of curriculums justify evolution of CET-Band 4? A retrospective textual analysis of CET-4
   Xu Wang, Nanjing University of Information Science and Technology

EFL teachers’ perceptions of a national English test
   Lisa Amdur, Tel Aviv University

Employers’ use of CET: From the test takers’ perspective
   Dayong Huang and Mark Garner, The University of Aberdeen

Evaluating aviation policy in relation to English language proficiency testing: The case of Korea
   Hyejeong Kim, The University of Melbourne

Integrating quantitative and qualitative approaches in L2 writing assessment
   Ling He, The University of British Columbia

Interactive assessment in an ESL context: Student learning, teacher constraints and the need for capacity-building
   Nicole Tavares, University of Hong Kong

Involving ESP stakeholders in rating scale validation: The case of the ICAO rating scale
   Ute Knoch, University of Melbourne

Justifying the use of classroom achievement assessment of reading for all stakeholders
   Kangzhou Peng, China West Normal University

Justifying the uses of a high-stakes EFL test: An application of the assessment use argument
   Huan Wang, University of California, Los Angeles and Ru Xing, Tsinghua University

Stakes, needs and washback: An investigation of the English graduation benchmark policy and EFL higher education at two technological universities in Taiwan
   Hsiu-yu Chu, Ming-chi University of Technology

Supporting teacher change in assessment practice through professional development
   Bordin Chinda, The University of Nottingham
Test interactivity as assessment of assessments in portfolio-based EFL writing program: The potentiality to incorporate test usability into test usefulness
Hongxia Zhang, Fanglin Zhou and Lihui Chen, Anhui University

Test-takers’ perceptions of and performances in a classroom-based oral assessment
Yongfei Wu, Anhui University

The effect of visual options of MCQ on listening comprehension
Zhi Li, Hunan University of Arts and Science

The impact of the language of examinations on students’ performance in national examinations: Individual, social and political perspectives
Guoxing Yu and Pauline Rea-Dickins, University of Bristol

What does the story retelling speaking test measure? Comparison with other types of speaking tests
Rie Koizumi, Tokiwa University and Akiyo Hirai, University of Tsukuba, Japan

6:30 - 7:00  LTRC 2008 GROUP PHOTOGRAPH (To be announced)

7:00 - 11:00  Banquet & Awards Presentation at Zhejiang Xizi Hotel

Music: To be announced

Awards
Spann Fellowship Awards for Second/Foreign Language Assessment and Evaluation, 2008
IELTS Masters Award, 2008
Robert Lado Memorial Award for Best Graduate Student Paper at LTRC 2008
ILTA Best Article Award, 2006, 2007
UCLES - ILTA Lifetime Achievement Award, 2008, Presented by Jim Purpura, Teachers College, Columbia University, and Nick Saville, University of Cambridge, ESOL

Closing thanks, farewells, passing the torch to organizers of LTRC 2009
2008 UCLES/ILTA Lifetime Achievement Award

J. Charles Alderson
Lancaster University

A lifetime of testing

Abstract

In this talk I shall reflect on a personal history of involvement in testing spanning over five decades. From early experiences of being a test-taker, to the difficulties of having to write tests, to doing PhD research in language testing, and on to delivering plenary presentations on testing, being inescapably branded as a language tester, then deciding I had to go with the flow and simply attempt to understand testing, tests and testers better. I shall end with an attempt to understand why anybody would ever get involved in testing, what their motivations, ambitions, hopes and fears might be, and what difficulties they face on the way.

Biography

J. Charles Alderson is Professor of Linguistics and English Language Education at the University of Lancaster. He was Scientific Coordinator of DIALANG 1999-2002 (www.dialang.org). He is internationally well known for his research and publications in language testing, including 17 books, 79 articles in refereed journals and chapters in books, 19 other publications, including research reports, 165 papers presented at professional conferences and seminars, and 197 seminars, workshops and consultancies outside Lancaster University, for numerous sponsors (The British Council, Overseas Development Administration, Council of Europe, overseas universities, etc). He was Director of a major international project to produce the
IELTS test, was Director of Testing and Certification at the University of Michigan and is currently adviser to a number of examination boards and agencies in Central Europe. He is former President of the International Language Testing Association, was co-editor of the main international journal in the field of language testing for five years (until 2002), and is currently co-editor (with Lyle Bachman of UCLA) of the Cambridge Language Assessment Series, published by Cambridge University Press, in which series his book Assessing Reading was published in 2000. He was the holder of a Leverhulme Research Fellowship in 2003, as a result of which his latest book, on Diagnosing foreign language proficiency: The interface between learning and assessment (Continuum Press) was published in 2005.
Samuel J. Messick Memorial Lecture

James Dean Brown
University of Hawai‘i at Manoa

Why don’t the stakeholders in language assessment just cooperate?

Abstract

The LTRC theme this year is “Focusing on the Core: Justifying the Use of Language Assessments to Stakeholders.” That theme lead to my title question, which I answer as follows: stakeholders don’t cooperate because, arrogantly or inadvertently, we fail to consider them, consult them, or properly inform them. Perhaps because we pursue the higher truth of validity, we fail to create stakeholder-friendly tests. In this paper, I define stakeholder-friendly tests, testing needs analysis, and defensible testing. I then go on to discuss the rational for this stakeholder-friendly approach and delineate the stages involved in any curriculum project. These stages (and substeps) are then translated into terms that relate directly to any language testing project as follows: (a) get ready to do testing project (define the purpose of the test, delimit the examinee population, decide upon approach(es) and syllabus(es), recognize constraints, and select data collection procedures); (b) do the testing project research (collect data, analyze data, and interpret results); and (c) use the testing project research results (determine impact of results on testing procedures, implement testing procedures and decision making, evaluate and report on the testing project, and decide on further information to gather). I also describe an example testing project (the ongoing evaluation of the Eiken tests and testing program) that illustrates how this framework can be applied. I end by considering the relationship between defensible testing and current notions of validity.
Biography

James Dean ("JD") Brown is Professor of Second Language Studies at the University of Hawai‘i at Manoa. He has spoken and taught courses in more than thirty countries ranging from Brazil to Yugoslavia. He has also published numerous journal articles and book chapters (on language testing, curriculum design, research methods, and program evaluation) and authored or co-authored numerous books (on reading statistical language studies, language curriculum, language testing, language testing in Japan, testing L2 pragmatics, performance testing, criterion-referenced language testing, using surveys in language programs, doing research, language test development, ideas for classroom assessment, connected speech, and heritage language curriculum).
Abstracts

Workshops 32
Symposia 38
Papers 45
Works-in-Progress 63
Posters 91
Workshops

Workshop 1

Time: June 23, Monday, 9:00 - 5:00  Location: Room 216, East 6

An introduction to latent growth modeling techniques using SEM

George A. Marcoulides, University of California, Riverside

Modeling change within a structural equation modeling (SEM) framework is a relatively recent and popular approach for studying developmental and longitudinal trends. This is because SEM offers a greater degree of flexibility in testing a variety of hypotheses concerning these trends than a number of other more traditional techniques like repeated measures analysis of variance. In simple terms, the approach resembles a basic confirmatory factor analysis model in which the latent variables are interpreted as chronometric variables representing individual differences over time. A variety of alternative names have been used in the literature to identify this approach, such as latent change analysis, latent growth curve, latent curve analysis, or just growth curve models.

This workshop will provide an introduction to the kinds of latent growth modeling strategies that can be used for the analysis of longitudinally obtained data. The workshop will include a “hands-on” computer component during which the popular Mplus software package will be illustrated with various data examples.

The level of presentation for the workshop presumes some familiarity with the basic principles of factor analysis. No prior familiarity with the Mplus software is assumed. Participants can review the basics of factor analysis by reading Chapter 4 in Raykov & Marcoulides (2006), A First Course in Structural Equation Modeling, Lawrence Erlbaum Associates, Inc.

George A. Marcoulides is Professor of Research Methods & Statistics at the University of California, Riverside. His research primarily focuses on methodologies used to study individual, group, and organizational phenomena, particularly those involving the use of structural equation modeling techniques and generalizability theory. He has served as a consultant to numerous government agencies, educational organizations and institutions, companies in the United States and abroad, and to various national and multi-national corporations. He is currently Editor of the Structural Equation Modeling, Editor of the Quantitative Methodology Book Series, and on the editorial board of numerous scholarly journals.
Applying school-based assessment for learning in a formal examination system

Chris Davison, University of Hong Kong
Liz Hamp-Lyons, Universities of Hong Kong, Bedfordshire and Nottingham
with the assistance of Gerry Davies and Nicole Tavares, University of Hong Kong

The Education Bureau (EDB) in Hong Kong has been encouraging a shift from assessment of learning to assessment for learning (AfL) in secondary schools, where assessment tasks, activities and criteria are strongly aligned with learning and teaching and where students are active participants in the assessment process. As part of this shift, the Hong Kong Examinations and Assessment Authority (HKEAA) has introduced a substantial school-based assessment (SBA) component into the high stakes senior secondary English examinations, which is now in its third year of staged implementation. This workshop will frame school-based assessment within assessment for learning, outline the key components of the English oral SBA, and provide participants with the opportunity to use the assessment criteria to assess a range of samples of oral language use. In the afternoon participants may choose one of four breakout sessions on various aspects of the assessment system, each of which highlights a key feature of SBA within the AfL paradigm: the design of learning-assessment tasks; peer assessment and self-evaluation; interactive assessment; and the role of feedback. The topics will be as follows:

Understanding SBA

Definitions, background, theoretical assumptions
Processes of “test” development and validation, research findings
Deriving assessment domains and judging levels; aligning established and new instruments. Working with performance samples. Standardisation and monitoring. Problems and issues.

Session 1: Preparing for the assessment: Task design and bias for the best
Chris Davison

Session 2: The role of the students in the assessment process: Self and peer assessment
Liz Hamp-Lyons

Session 3: The role of the teacher during the assessment process: Interactive assessment
Nicole Tavares

Session 4: After the assessment: The process of feedback
Gerri Davis

Implications for classroom practice, teacher preparation and policy support

Chris Davison is Associate Dean (Research) in the Faculty of Education at the University of Hong Kong. She has worked for many years in English language teacher education in the Asia-Pacific region, and has published widely in the areas of English language development and assessment in schools. Her most recent books are on English language teaching in China (Continuum Press, 2008, with Xinmin Zheng) and internationally (Springer, 2007, with Jim Cummins). With Liz Hamp-Lyons, she developed the HKCEE school-based assessment speaking component for the HKEAA and is directing a range of related school-based assessment research and development projects in the region.

Liz Hamp-Lyons, Honorary Professor in Education at The University of Hong Kong, has worked extensively in English language education, applied linguistics and language assessment. She is a Past President of the International Language Testing Association, she holds honorary positions at Shanghai Jiaotong University and the University of Nottingham, and is Professor of English Language Assessment at the University of Bedfordshire. She has published widely and has consulted to major testing bodies and examinations agencies, acting as the senior
advisor to the National College English Testing Committee of China on CET-4. With Chris Davison, she developed the HKCEE school-based assessment speaking component for the HKEAA, and is engaged in several related school-based assessment research projects, including a policy-level impact study.

**Geraldine Davies** has been teaching English language since 1979 and has taught secondary school students and adults in England, Australia and Hong Kong. She began teacher training in 1983 at the British Council in Hong Kong and subsequently ran teacher training programmes at the West Australian College of Advanced Education and then at an International College, also in Australia. She joined the University of Hong Kong in 2003. Her current research interest is in school based assessment and the provision of formative feedback in assessment for learning and she is a key researcher in a large scale action research project aimed at developing more effective assessment for learning among junior secondary students.

**Nicole Tavares** is a teacher educator in the Faculty of Education of The University of Hong Kong and runs courses for the Centre for the Advancement of University Teaching. Nicole holds major positions in the Hong Kong Examinations and Assessment Authority in the Territory-Wide System Assessment for primary students, Certificate Level and Advanced Level English examinations for secondary students and Language Proficiency Assessment for Teachers of English. Before joining the University in 2001, Nicole was English Panel Chairperson at a secondary school for many years where she taught second language learners across all levels and headed Quality Education Fund projects. Her main research interests are in dynamic/interactive assessment, collaborative learning, and teacher assessment readiness.

**Workshop 3**

**Time:** June 24, Tuesday, 9:00 - 5:00  **Location:** Room 216, East 6

**What teachers need to know about developing useful assessments for language learners**

Adrian Palmer, *University of Utah*

MaryAnn Christison, *University of Utah*

In the field of second and foreign language teaching, there has often been a disconnect between instruction and assessment. For instructional design specialists, assessment is often an afterthought. For assessment and language testing experts, assessment often starts from the point of view that instruction is excellent and principled. The presenters of this workshop believe that both points of view are problematic. For instructional design specialists, assessment should not be an afterthought and assessment and language testing specialists should not assume that all instruction is principled and suitable as a starting point for the development of assessment tasks. In order to develop useful assessment tasks for classroom purposes, there must be an integration of principled language instruction and language assessment.

In the first part of this workshop, we will first address principles that must govern the design of instruction. If instruction is to be used as a basis for the development of assessment tasks, key principles, such as defining content concepts, addressing demands on cognition, and identifying learning behaviors, must be clearly presented. Having established a principled framework for instructional design, we will then provide participants with a structure for developing assessment tasks based on specific task characteristics that can be identified and manipulated. Sample materials and examples will be provided.

Whether you are in a low stakes assessment situation in a classroom or whether you are in a high stakes assessment situation, you must be able to justify the decisions you make on the basis of the assessment to individual stakeholders such as parents. In the second part of this workshop, we will turn our attention from the alignment of instruction and assessment to the process of justifying the use of the assessment as a whole. This starts with identifying 1) the beneficial consequences of assessment use, 2) the specific decisions that need to be made to promote these consequences, 3) the information about the test takers language ability needed to make these decisions, and 4) the kinds of tasks used to obtain this information.
Audience: This workshop is for all classroom teachers and specifically for classroom teachers interested in assessment. Materials will be provided.

Adrian Palmer teaches a variety of courses in applied linguistics at the University of Utah. His areas of specialization include language assessment, teacher training and professional development, and language teaching methodology, and he regularly offers a wide variety of workshops and presentations on these topics. Recent publications include Seeking the Heart of Teaching (U. of Michigan Press), 2007.

Mary Ann Christison is a professor and Director of Graduate Studies in the Department of Linguistics and adjunct professor in the Department of Teaching and Learning at the University of Utah. She teaches courses in the MA and Ph.D. graduate programs and was TESOL President 1997-98. She has been a teacher educator for 23 years, working with teachers in the U.S. and in over 30 different countries throughout the world.

Workshop 4

Time: June 24, Tuesday, 9:00 - 5:00 Location: Room 218, East 6

Basic statistics for language assessment

Gary J. Ockey, University of New Mexico
Lorena Llosa, New York University

This workshop will introduce the statistical concepts and techniques needed to understand and conduct basic quantitative language assessment analyses. Participants will receive hands-on training for how to conduct these statistical analyses in Excel and SPSS as well as how to interpret the information after it has been produced. The workshop is designed to address the needs of a diverse group of members of the language testing community including: students in applied linguistics or related programs, language teachers, and individuals developing language tests in the public or private sectors. Specifically, the workshop is intended for individuals who have little or no experience with basic statistical techniques.

The workshop will cover three general topics: 1) introductory statistics concepts, 2) test analyses procedures, and 3) research techniques for language testers. Participants would benefit from reading an introductory textbook on language assessment statistical analyses, such as Bachman (2004) prior to participating in the workshop.


Gary J. Ockey is an assistant professor in the Educational Psychology program at the University of New Mexico. He teaches courses in applied statistics and measurement in education and psychology. His research interests include second language group oral assessment, validation of research instruments, and applications of structural equation modeling and item response theory.

Lorena Llosa is an Assistant Professor in the Department of Teaching and Learning at New York University. Prior to joining NYU she worked as a research analyst for the Los Angeles Unified School District where she directed a large scale evaluation of a computer-based literacy program. She also served as a research analyst at the Center for the Study of Evaluation/CRESST at UCLA where she worked on the development of performance assessments in English and Spanish. She received her Ph.D. in applied linguistics from the University of California, Los Angeles. Dr. Llosa’s interests include language assessment, second and foreign language teaching and learning, and program evaluation. Her current research projects address validity issues related to the use of standards-based classroom assessments of English proficiency in large urban school districts, the development of a diagnostic assessment of high school students academic writing, and the testing and placement of language minority students in California’s community colleges.
Workshop 5

Time: June 24, Tuesday, 9:00 - 5:00  Location: Room 220, East 6

The application of Generalizability theory in L2 assessment

Yasuyo Sawaki, Educational Testing Service, Princeton, New Jersey
Xiaoming Xi, Educational Testing Service, Princeton, New Jersey

This course provides an overview of generalizability theory (G theory) and its application in the context of second language (L2) assessment. G theory, developed by Lee Cronbach and his associates in the 1970s, is a conceptual framework and set of statistical procedures that inform assessment instrument design and validation. Participants in this course will be introduced to the theoretical bases of G theory and typical G study designs in L2 assessment, and will have opportunities to conduct hands-on data analyses of sample L2 assessment data sets using a computer program, GENOVA (Crick & Brennan, 2001). The course starts with a review of traditional approaches to the investigation of test reliability, followed by an overview of fundamentals of the univariate G theory framework and practical issues surrounding the application of G theory to analysis of L2 assessment data. Participants in this course are expected to be familiar with classical test theory (CTT) and analysis of variance (ANOVA). A brief description is provided below.

- Overview of methods to analyze score reliability
  - Analyzing test reliability in norm-referenced and criterion-referenced testing
  - Analyzing rater reliability & agreement in performance assessment
  - Classical test theory (CTT) vs. G theory and limitations of CTT
- Key concepts of G theory
- Demonstration of typical G study designs in L2 assessment
  - For multiple-choice items and for constructive-response items
- Practice session -- Turning verbal descriptions of assessment designs into G study designs
- Hands-on analyses of sample L2 assessment data sets using GENOVA

Prerequisites:
Participants in this course are expected to be familiar with classical test theory (CTT) and analysis of variance (ANOVA). Individuals who want to participate but who are not familiar with classical test theory are advised to read:


For participants who would like to obtain the software themselves, it is available at http://www.education.uiowa.edu/casma/computer_programs.htm#genova

Yasuyo Sawaki is a research scientist at Educational Testing Service, Princeton, New Jersey. She earned her Ph.D. in applied linguistics from the University of California, Los Angeles. Since joining ETS in 2003, she has been working on various test validation projects for TOEFL iBT and basic research projects on language ability of adult ESL learners. She has research interests in diverse topics, including second language reading development and assessment, dependability of second language assessment, and diagnostic language assessment. She has employed generalizability theory (G theory) in her own previous research projects and recently taught a graduate-level course on the application of G theory in second language assessment at Teachers College, Columbia University.

Xiaoming Xi is a research scientist in the Research & Development Division at Educational Testing Service, Princeton, New Jersey. She earned her Ph.D. in applied linguistics from the University of California, Los Angeles.
Her areas of interest include factors affecting performance on speaking tests, rating scales for speaking tests, rater bias issues in speech scoring, automated scoring of speech, and validity and fairness issues in the broader context of test use. She has published in leading journals and wrote a chapter on “Methods of Test Validation” for the second edition of the Encyclopedia of Language and Education. She also serves on the Editorial Boards of Language Testing and Language Assessment Quarterly. She is a recipient of the 2002 Lado Best Student Paper Award, the 2003 Spann Fellowship Award for Second/Foreign Language Assessment, the 2005 ILTA Best Paper Award, and the 2005 and 2006 ETS Presidential Award.
Symposium 1

Interacting with stakeholders on language tests: Issues, problems and challenges

Time: June 25, Wednesday, 1:50 - 3:50  Location: Room 227, East 6

Elana Shohamy and Bernard Spolsky, Organizers

Alan Davies, *University of Edinburgh*, “Should we listen to (all) stakeholders?”
Janna Fox, *Carleton University*, “Policy intent and consequence: When test purpose is at odds with test use”
Cathie Elder, *University of Melbourne*, “Comparing tests: What to tell the stakeholders?”
Elana Shohamy and Ofra Inbar, *Tel Aviv University* and *Beit Berl Academic College*, “Different patterns of negotiating testing policies and practices with stakeholders”

Discussant: Bernard Spolsky, *Bar Ilan University*

In this symposium we will focus on the main theme of the LTRC colloquium entitled “Focusing on the Core: Justifying the Use of Language Assessments to Stakeholders”. Four papers representing different contexts will be presented describing the complex relationship of language testers and stakeholders and showing where language testers can create constructive approaches to the topic, a double-pronged approach that considers what stakeholders can expect from a language assessment, i.e., how much error is left in, what is the risk of naive interpretation, and conversely how to design an assessment and say honestly what it can be expected to do, especially based on cases of our experiences with these interactions. More specifically, the papers will focus on what considerations and message we, researchers on language testing, can deliver to or receive from (different) stakeholders about language test uses and testing policy, so to contribute to ‘a better testing world’.

**Should we listen to (all) stakeholders?**

Alan Davies, *University of Edinburgh* (a1adavie@staffmail.ed.ac.uk)

The norm or model widely used in English language proficiency testing is that of the native speaker. This model is now under attack from three groups who, although not institutionalised, function as stakeholders. The first are the so-called *native users*, native-like speakers of English (in, eg India) whose professional and sometimes personal lives are lived largely or wholly in English but who are not labelled native speakers and may therefore be discriminated against in terms of employment. The second are speakers of English as a Lingua Franca (ELiF), those who use English with one another for specific commercial, professional etc purposes and do not see the need for their English to conform to the native speaker model. The third are those academics who take an ideological position on language variety and change and who maintain that the Standard English model(s) required in international English proficiency tests, such as TOEFL, IELTS, represents a post-colonial position. The paper will characterize the views of these three stakeholders (Canagarajah 2005, Elder and Davies 2006, Jenkins 2006, Lowenburg 1993, 2002, Poehner and Lantolf 2005, Seidlhofer 2001) along with data from recent studies of spoken and written performance of international test candidates comparing international and local judges (Davies, Hamp-Lyons and
Kemp 2003, Hamp-Lyons and Davies 2007). It will argue that while there is no reason why there should not be
different native speaker (or native user or ELiF) models, at present there is no evidential basis to support the use
of any new model or to support the claim of bias in international tests, if only because what is criticised as bias
is more properly considered to be legitimate difference. The paper concludes with a consideration of how best to
communicate this argument to these stakeholders who are concerned with the symbolic power of tests and cannot
accept the notion that explicit norms are essential in any test situation.

**Policy intent and consequence: When test purpose is at odds with test use**

Janna Fox, *Carleton University* (jfox@ccs.carleton.ca)

Policy makers in a Canadian university were intent on increasing the fairness of admission procedures for second
language (L2) international students who arrived with the expectation of beginning their degree programs and were
faced instead with unanticipated delays and mandatory English for Academic Purposes (EAP) courses as a result
of lower than anticipated scores on the university’s placement test. The following logic was applied in defining the
new policy: 1) students need to know in advance of their arrival if they will be required to take EAP during the first
term(s) of their university program; 2) proficiency test scores are used for full-time admission to degree programs
(i.e., based on cut-scores for IELTS; TOEFL iBT, CBT, PBT; or a locally-administered test, the CAEL Assessment); 3)
therefore, cut-offs can also be used to place students in levels (introductory, intermediate and advanced) of the
mandatory EAP program; 4) concordance tables are available for all of these tests; thus, 5) concordance tables can
be used to establish cut-score comparability across these tests for placement in different levels of the EAP program.
This post hoc study reports on the consequences of the new policy for key stakeholders. Applying a mixed method
approach to examine policy impact over time, data were drawn at admission from (n=386) test scores; and, during
the first academic year, from questionnaire responses, diagnostic test results, and academic outcome indicators
(e.g., grades, withdrawals, attendance, etc.). Concurrently, semi-structured interviews were conducted with 24 of
the students and 5 of their EAP teachers. Results highlight academic outcomes in relation to the proficiency test
used for placement, problems with stakeholder ‘uptake’ of proficiency test scores and concordance tables, and the
role that diagnostic assessment can play in language teaching, learning and curricular reform.

**Comparing tests: What to tell the stakeholders?**

Cathie Elder, *University of Melbourne* (caelder@unimelb.edu.au)

Accrediting bodies differ in the English language requirements for overseas-trained health professionals applic-
ing 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 one test is more appropriate than the other or 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. A key question for language testing experts is: How to demonstrate and
communicate to stakeholders the lack of parity between measures and the relative costs and benefits of taking a
general versus a specific-purpose test?

The paper reports on an attempt to undertake such a cost-benefit analysis, drawing on a) published information
about each test, b) a content/construct comparison, c) a small-scale comparison of scores from operational ad-
ministrations of each test involving the same group of overseas-trained health professionals d) a survey of these
test-takers’ perceptions about each test’s validity/fairness as a measure of their language ability. Results accord
with findings of previous language test comparisons, showing moderate correlations between the two measures in
spite of their very different content, format and scoring procedures, but also revealing (as expected) substantial dis-
crepancies in “pass/fail” classifications. And while survey results showed a clear preference for OET over IELTS,
test-taker perceptions are clearly insufficient evidence for claiming superiority of one test over another.

The paper concludes by discussing the problems experienced in communicating the clear message required by ac-
crediting authorities about score equivalence and test appropriateness, given the difficulties inherent in such com
parisons. It also poses the more general question of how testing experts can assist stakeholders in making informed and responsible test choices and to what extent our profession should be held responsible for decisions made.

**Different patterns of negotiating testing policies and practices with stakeholders**

Elana Shohamy and Ofra Inbar, *Tel Aviv University and Beit Berl Academic College* (elana@post.tau.ac.il, inbarofra@bezeqint.net)

In this paper we will describe the complex relationship between language testers and stakeholders, especially government agencies, as is the case in centralized educational systems. We will claim that this relationship is complex and unpredictable and should be viewed within the broader perspective of policy making whereby a large number of factors affect testing policy making. As such it falls within the set of factors that are part of educational and linguistic ideologies and policies. We will support our claims by introducing three different language testing patterns of interactions between stakeholders and testing researchers. These models differ in the roles taken by the ‘two parties’ the language testing experts and the stakeholders. Each of the three cases of interaction patterns exemplifies negotiations of testing policies and practices. In the first case, the recommendations for assessing immigrants by incorporating their L-1 through test accommodations resulted in actual change of policies in that direction. We attribute these changes to the availability of convincing empirical data to support the findings and Ministry of Education financial support of the project. In the second case, proposals for changes in language assessment were initiated by the Ministry for alternative, formative and continuous assessment tools. The interaction pattern was of complete cooperation of the two teams in all phases of the assessment - conceptualization, development, validation and implementation. Yet, the actual use of the product was very limited. In the third case a high stake test, initiated by the stakeholders is taking place whereby testing experts serve as advisers and are instrumental in attempting to contribute ‘from within’. We will discuss the unique aspects of each of these three patterns by demonstrating their advantages and disadvantages as we contextualize them within theories of decision making and language policy.

**Symposium 2**

**Canvassing stakeholder perspectives: The key to test usefulness?**

**Time:** June 26, Thursday, 10:20 - 12:20  **Location:** Room 227, East 6

Cathie Elder, Organizer

Cathie Elder and Kieran O’Loughlin, *University of Melbourne*, “Language standards: How are they viewed by the ELT industry?”


Ute Knoch, *University of Melbourne*, “An examination of stakeholder views of diagnostic feedback to EAP writing performance”


Discussant: David Nevo, *Tel Aviv University*

Convenor: Cathie Elder, *University of Melbourne*

It is generally accepted that language testers know more about assessment than lay people and that the tests they develop are rooted in scientific and professional knowledge. Indeed the process of professionalisation requires the tester to discard previously held lay views and accept professional ideas about the what and how of valid measurement. Doing so at the expense of listening to stakeholders’ views and feelings may however cause communication problems and/or policy failure, leading to inappropriate test interpretation and use. This 2-hour symposium presents the outcomes of empirical research in a number of different contexts exploring stakeholder perspectives on
the design and use of tests and assessment frameworks. Stakeholders include government decision-makers, school administrators, ESL teachers, ESL assessors, test candidates and other users of tests, scales and score reports. Each paper will identify divergent stakeholder views and consider the implications of these for decisions about assessment design and delivery and also for communications about test use. Titles and presenters are listed below, together with a brief outline of each paper. At the end of the session a discussant (David Nevo, Tel Aviv University) will link the insights from this symposium to the notion of utilization-focussed evaluation (Patton, 1986; 1997). He will draw on principles from the program evaluation literature to propose strategies for canvassing and responding to stakeholder feedback with a view to promoting more appropriate and ethical test use.

**Language standards: How are they viewed by the ELT industry?**

Cathie Elder and Kieran O’Loughlin, *University of Melbourne* (caelder@unimelb.edu.au, kjo@unimelb.edu.au)

Debates about the value of common language standards or frameworks are prevalent in the educational community, and while these clearly have wide symbolic appeal for accountability purposes, their usefulness as means of raising the quality of language teaching learning and assessment practices depends at least partly on how they are viewed and implemented by the teaching force (Brindley, 1999; 2001; McKay, 2000; Moore, 2001).

The study used a combination of focus group, survey and interview techniques to elicit the opinions of a representative sample of 50 ELT practitioners on the perceived feasibility and utility of introducing the CEFR for the ELT industry in Australia. The study was motivated by regulatory bodies’ concerns about the lack of uniformity and portability of teaching and assessment practices within the sector. Findings revealed divergent views about the usefulness of a common framework. While some welcomed the idea of adopting the CEFR (in preference to other locally developed schemes) as means of giving “international currency” to Australian courses, others were concerned about the conflict between the universalising tendency of the framework and the need to tailor programs to particular client needs in what is a highly competitive industry.

The paper argues that this perceived tension between notions of regulation and ELT practitioners’ desire for local autonomy is based on misunderstandings about the intended purpose and uses of the CEFR. It proposes a professional development program to assist teachers and other stakeholders in selecting, interpreting and using such frameworks in ways which are consonant with their local needs and purposes.

**Decision-making in marking open-ended listening test items: The story of a marking guide, from designers to users**

Luke Harding and Kerry Ryan, *The University of Melbourne* (lharding@unimelb.edu.au, ryank@unimelb.edu)
John Pill, *The OET Centre, Melbourne* (john.pill@oet.com.au)

The creation and application of marking guides for open-ended listening and reading responses have attracted little attention in the language testing literature. A marking guide is intended to stipulate precisely what kind of response test designers accept as evidence of the ability under test and as such is a key component of the test construct. However, in a test containing open-ended questions there remains scope for markers, key stakeholders in the implementation of the test, to apply their own interpretations of the construct in judging responses which fall outside the information provided in the marking guide.

This paper charts the progress of the development, and eventual use, of a marking guide for the listening component of the Occupational English Test (OET) – a language test for overseas-qualified health professionals wishing to work in Australia. Three questions are posed: what decisions do markers make while working independently with a marking guide to score answer papers? Do these decisions accord with the test designers’ interpretation of the construct? And how sufficiently are the needs and perceptions of the users of the marking guide addressed by test developers in the design of the marking guide?
To draw out the types of decisions that are made regarding the appropriateness of responses it records and analyses discussions at a) a designer paneling session b) a subsequent rater-training session and c) a follow up focus group discussion on completion of the marking process.

Findings, which show notable points of divergence between designers and markers in their interpretation of the marking scheme, underscore the importance of carefully eliciting user perspectives at the design stage in order to clarify interpretations of the test construct and produce more valid and useable marking guides.

**An examination of stakeholder views of diagnostic feedback to EAP writing performance**

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

Recently, renewed interest has surfaced in the area of diagnostic assessment (e.g. Alderson, 2005). Very few studies, however, have investigated diagnostic assessment of writing, especially in large-scale EAP settings like universities. Providing large numbers of students and staff members with individualized diagnostic feedback on their writing could prove useful in identifying areas that need to be worked on either individually or in the context of language support classes or discipline specific tutorials. Because the feedback profile generated on the basis of the assessment is the most important outcome for the stakeholders, this study set out to establish what type of feedback is considered feasible and useful to staff, students and raters at a university with a large population of international students.

Ten ESL raters and 10 university staff (including language support staff and a range of faculty lecturers) were interviewed to elicit their ideas about what level of detail should be included in the rating scale descriptors so that these can be used as the basis of the feedback protocols to inform teaching and learning. Finally, 10 students, also from a variety of departments, were asked to produce a piece of writing based on a diagnostic writing task currently used to identify support needs at the university and interviewed about their preferences for different types of diagnostic feedback. The findings suggest that while students prefer as much detail in their feedback profiles as possible, there is some anxiety among both staff and raters about the useability of such detailed feedback. Ways to reconcile these different stakeholder views are discussed.

Finally, ten students, from a variety of university departments, were asked to produce a piece of writing based on a diagnostic writing task currently in use at the university for the purposes of identifying students’ language support needs. The students were then interviewed about their preferences for different types of diagnostic feedback.

**The old TOEFL: Students’ perspectives**

Liz Hamp-Lyons, *Universities of Nottingham* (lizhl@hkucc.hku.hk) and Annie Brown, *Ministry of Higher Education & Scientific Research, United Arab Emirates* (abrown@hct.ac.ae)

The concept of washback is situated within the focus on ‘test use’, and when used in educational contexts, refers to behavioural changes or influences in areas such as curriculum, teaching methods, learning strategies, materials and courseware, assessment practices, and the content of instruction. The related concept of impact is concerned with the values, motivations, and consequences of tests, and impact moves beyond the immediate learning environment to consider changes or influences in societal domains such as selection, gatekeeping, ideology, ethics, and fairness. A focus on impact necessitates a close awareness of the perspectives of stakeholders, particularly those most closely affected by, as in this case, changes to a highly influential test.

In this paper we draw on data collected as part of a study of the attitudes of students in “old TOEFL” test preparation classes in three contexts before the introduction of “new TOEFL” (the TOEFL iBT) in China, heterogeneous nationalities in the USA, and Egypt. The data consist of students’ survey responses (N=505) and transcripts of interview and focus group sessions (N=27). The data show that students in all countries were generally positively oriented to the test, but that while they valued narrowly-focused test preparation activities, they also valued more
communicatively oriented activities which allowed them to develop their proficiency more generally. However, significant differences in attitudes towards the test and test preparation were found across the countries.

The paper considers the potential implications of these findings for the uptake of the iBT TOEFL, arguing that acceptance of this more communicatively oriented test may depend less on its intrinsic qualities than on students’ existing perceptions and test attitudes.

**Symposium 3**

**The ‘teacher factor’ in washback contexts: Evidence from mixed methods research in Canada, China and Malaysia**

**Time:** June 27, Friday, 10:20 - 12:20  **Location:** Room 227, East 6

Carolyn Turner, Organizer

Carolyn E. Turner, *McGill University*, “The specificity of the “research approach” in classroom studies: Probing the predictability of washback through teacher conceptual and instrumental evidence in Quebec high schools”

Jing Wang, *McGill University*, “Identifying the main impact on teacher classroom behaviour: Examining the effects of the CET, teacher beliefs, teacher knowledge bases and teacher proficiency”

May Tan, *McGill University*, “Bilingual high-stakes mathematics and science exams in Malaysia: Pedagogical and linguistic issues”

Discussant: Liying Cheng, *Queen’s University*

**Paper 1 (Canada) The specificity of the “research approach” in classroom studies: Probing the predictability of washback through teacher conceptual and instrumental evidence in Quebec high schools**

Carolyn E. Turner, *McGill University* (carolyn.turner@mcgill.ca)

This paper examines teacher behavior and beliefs when dealing with a new speaking exam. Participants were 153 high school teachers, in addition to 5 case study high school teachers and their classrooms. The study’s intent was to explore the (un)predictability of washback by targeting the specific characteristics of the new exam in each of its instruments. Results provide evidence of features that appear differentially important in predicting washback on teaching (in this case, positive washback), as opposed to evidence for developing generic models of teacher behavior.

**Paper 2 (China) Identifying the main impact on teacher classroom behaviour: Examining the effects of the CET, teacher beliefs, teacher knowledge bases and teacher proficiency**

Jing Wang, *McGill University* (lily_wang_99@yahoo.com)

This paper addresses the research question: What role does the ‘teacher factor’ play in washback in the Chinese university context? Participants were 195 tertiary-level EFL teachers of the non-English programs. The intent of this study was to investigate whether tests constitute a major constraint on College English instructional innovation in China. Results demonstrate that it is not the tests themselves that determine teacher practice.
This paper examines a change in language of instruction, from Bahasa Malaysia (BM) to English, for mathematics and science in Malaysian high school classrooms. To ease transition, high-stakes exit exams are in both BM and English. The research questions are: “How do teachers perceive the bilingual Mathematics and Science exams?” and “How do their perceptions affect their classroom practices?” Participants were 362 secondary school teachers as well as 6 case studies of teachers and their classroom practices. The results indicate that teachers see BM and English questions as being equivalent, and tend to teach using a mix of direct translation and code switching techniques. These practices have implications for students who are average and weak in English proficiency: their exam performance can be negatively affected.

Discussant: Liying Cheng, Queen's University (liying.cheng@queensu.ca)
Session 1

Time: June 25, Wednesday, 10:50 - 12:20  Location: Zhejiang University Conference Center

Using summative assessments for formative purposes: The ultimate justification for learners and teachers

Chris Davidson, University of Hong Kong (cdavidson@hkucc.hku.hk)

This paper reports on the findings of an investigation into the extent to which summative assessments can be used for formative purposes in secondary Hong Kong English as a second language contexts. Formative assessment (the gathering of information about students and their language learning needs while they are still learning, so that decisions can be made about what is to be done next) requires both constructive teacher feedback and active student involvement, understanding and action. However, few researchers have investigated the formative use of summative assessments in actual classroom contexts, perhaps because in traditional examination-dominated cultures, formative and summative assessment are seen as distinctly different in both form and function. However, recently a number of researchers, especially those working in the Asian region (eg. Biggs, 1998; Carless, 2007; Davison & Hamp-Lyons, 2007; Harlen, 2005; Kennedy, 2006; Taras, 2005) have questioned this sharp distinction between ‘formative’ and ‘summative’ assessment. They argue that provided summative assessment is undertaken while students are still learning (and teachers are still teaching), such assessments can also be used for formative purposes, that is, to improve learning and teaching – thus building a more coherent and stronger assessment for learning culture, at the same time providing stronger justification for the investment of time, resources and effort into designing and implementing high quality assessments. In recent high stakes assessment reforms to the senior secondary curriculum in Hong Kong, summative assessments of the students’ speaking skills have been designed to be used formatively to ensure constructive student feedback and enhance learning. Adopting a Vygotskian sociocultural perspective, this paper draws on interview, observational and questionnaire data collected over an 18 month period of working collaboratively with a group of secondary English language teachers and their students from a range of 8 schools in Hong Kong. Building on a detailed analysis of classroom interactions, this paper presents a multidimensional framework mapping the ways in which oral summative assessments in school-based assessment can be used formatively to improve teaching and learning. The implications for professional development and assessment theory and practice will also be discussed.

ESL teachers learning to assess interactively: The challenges and teacher development issues

Nicole Tavares and Liz Hamp-Lyons, University of Hong Kong (tavaresn@hkucc.hku.hk, lizhl@hkucc.hku.hk)

This paper examines Hong Kong English teachers’ exploration of the meaning and practice of ‘interactive assessment’ (IA) and its implications for teacher education. Grounded in Vygotsky’s Zone of Proximal Development (1978) that knowledge is constructed through dialogue, in line withValsiner’s (2001) future-in-the-making model and developed based on the theories of educative assessment (Stiggins 1998), assessment as feedback (Wiggins 2004) and dynamic assessment (Lantolf and Poehner 2004; Poehner and Lantolf 2005), IA is an initiative in the most recent school-based assessment reform in Hong Kong which advocates an expansion of the interactive and mediational role of the teacher in the process of assessing students’ oral language.
IA requires that teacher-assessors intervene during students’ spoken performance to provide the most appropriate ‘mediated guidance’ – (i) to scaffold and support students in producing an assessable performance that they know students are capable of producing in an everyday environment and (ii) to stimulate and challenge them to reach a higher-level performance than that being produced. IA is nonetheless counter to prevailing conceptions in most educational cultures of how assessment ‘should’ be done. This calls for re-conceptualisation of the teacher-assessor’s role and strategy use, and thus presents ideological, linguistic and pedagogical challenges.

Adopting a Vygotskian sociocultural approach to the analysis of teacher-student while-assessment discourse and teachers’ mediational strategy use, this paper reports on the results of an action-research study of seven Hong Kong junior secondary schools seeking to help teachers improve their IA skills. The paper will discuss how local English teachers interpret IA and the major constraints and difficulties they face in its implementation. Drawing on pre- and post-assessment interview data, teacher reflections, and the observation and analysis of actual IA episodes in the classroom, the paper explains aspects of students’ performance teachers (fail to) focus on, reasons for their ‘strategic’ choice and the factors limiting their ability to assess interactively. Pedagogical implications for teachers’ lack of ability to practise IA will also be examined and recommendations for teacher professional development made.

Stakeholder involvement in the language assessment evaluation process: A Hong Kong case study

Alan Urmston and Felicia Fang, Hong Kong Polytechnic University (ecalanu@polyu.edu.hk, ecfang@inet.polyu.edu.hk)
Tom Lumley, Australian Council for Educational Research (lumley@acer.edu.au)

The benefits of consulting stakeholders during the process of test development and the potential roles that stakeholders can play in test evaluations have been discussed frequently in recent language assessment forums.

The presentation of this paper focuses on the Language Proficiency Assessment for Teachers (English Language) (LPATE), a high-stakes language assessment, which was introduced by the Hong Kong government in 2001. A deadline of September 2006 was set for all English teachers to attain the required English language proficiency and one way they could do this was by taking the Assessment. In 2006, instead of scrapping the Assessment, the government moved to extend the life of the LPATE and to commission a test revision project to review the original test design. A major part of the review entailed taking the views of stakeholders.

Stakeholders are defined as those who make decisions and those who are affected by those decisions (Rea-Dickens, 1997). Consultations and dialogues with stakeholders are considered a crucial stage of this assessment revision process. After briefly discussing the controversial nature of the implementation of the LPATE, the presentation details how the revision project team targeted seven stakeholder groups and then conducted both focus group discussions and an evaluation questionnaire survey to make sure that adequate relevant information was collected. The presentation further describes how this initial input was analysed, compiled and applied to inform the revision of test specifications and the writing of new tests.

The first administration of the revised LPATE will be launched in 2008. It will be interesting to see whether the stakeholder evaluation approach to assessment (re)development enhances the validity and overall fairness of the Assessment.
Session 2

Time: June 26, Thursday, 8:30 - 10:00  Location: Room 227, East 6

Monitoring validity claims in a high-stakes setting

J. Charles Alderson, Lancaster University (c.alderson@lancaster.ac.uk)
Jayanti Banerjee, University of Michigan (j.v.banerjee@googlemail.com)

By 5 March 2008, air traffic controllers and pilots will need to have a certificate attesting to their proficiency in the language(s) used for aeronautical communication. Aviation language tests are extremely high stakes, not just for the test-takers but for every potential airline passenger, not to mention the airline companies, insurance companies, etc. Evidence of aviation test quality would certainly reassure travellers, particularly evidence that adheres to codes of practice or guidelines for good practice in language testing. Several such codes of practice and guidelines for good practice have been developed in language testing in recent years (for example, the ILTA draft Code of Practice, the EALTA Guidelines for Good Practice in Language Testing and the Japanese Language Testing Association Code of Practice). However, it is not yet clear how adherence to codes of practice for language tests is being, or should be, monitored.

Serious concerns have been expressed on LTEST-L and at LTRC 2007 about the process of test design and the lack of testing expertise engaged in producing at least some tests of aviation English. Additionally, as part of a project to validate the development of one test of English for air-traffic controllers, we surveyed the websites of a number of competitor tests and were extremely surprised and concerned to find a lack of evidence for validity and reliability of almost all of these tests. We therefore decided to conduct a survey based on the EALTA Guidelines for Good Practice in Language Testing and Assessment (http://www.ealta.eu.org/guidelines.htm) in order to establish what evidence might be available to substantiate the claims made on the websites. This paper will report on the results of that survey, and we will discuss the implications for the implementation of codes of practice in language testing.

Whose standards is it anyway? Theoretical perspectives on defining test standards in language testing

Samira El Atia, University of Alberta (selatia@ualberta.ca)

In 1995, the International Language Testing Association (ILTA) published the Report of the Task Force on Testing Standards (TFTS). There were two objectives at the time:

“first, it provides a general resource for scholars of educational and language assessment standards, and second, it serves as a specific resource for later ILTA efforts to draft its own code of practice, in that ILTA can consult this report to obtain information about extant documents of that nature” (TFTS report 1995, p.10)

In this report, the issue of defining testing standards for languages was raised but never answered. Given the changes and the advances in both the field of language testing and language acquisition and teaching, agreeing on a common definition for Language Testing Standards (LTS) appears as a matter of utmost importance in the work of the second TFTS that was reconstituted in December 2006 and whose mandate is to update the first report. The second TFTS is due to present its report in 2009 and at the core of its work is defining test standards in language testing. This theoretical paper addresses the complex issue of defining standards for language testing as is being discussed in the second TFTS.

To define testing standards, it is important first to have a closer look at the very notion of testing standards and their purpose. The adherence to a given testing standard can serve primarily as an indication of good practice in test development and utilization. It is also a strong validation argument for, or in some cases against, a test or its use in a particular application. Test developers use standards in language testing to justify the use of a test to various stakeholders. Through the standards, the purposes of a test can also become clearer to the users.

A problem in defining standards stems from the complex nature of the field of language testing which can be studied...
ied from various perspectives: that of linguistics, testing and measurement, language acquisition and teaching. The nature and use of the language itself is a source of complexity: language is more than its mechanics; it is cultural, social, economic, ethnical, historical and political.

Initially two definitions could be given to LTS. As guidelines for good practice, LTS may serve as a reference for test developers, users, and anyone closely or remotely linked to a language test. In this situation, it would serve the same purpose as the Standards for Educational and Psychological Testing (APA, NCME, AERA 1999), but for language testing. LTS can also be a statement of expected ability. In this case, the standards are indicators of language competence. They serve as scales for establishing item difficulty and establishing language constructs and criteria to be assessed at different levels. Thus, the standards are linked closely to the language performance indicators such as “the Canadian Language Benchmarks,” “the Common European Framework of Reference for Languages,” or the “ACTFL guidelines.”

In effect, a standard as guideline for good practice may serve exactly the same political control function as an indicator for language competence. They may say things very differently, but the impact on the world could be equivalent. This is particularly true in the modern standards-driven era in language test development, in which the second meaning often guides test development practice more than the first. For example, in the case of Canada, test developers strictly adhere to the Canadian Language Benchmarks for the national immigration language tests.

The two definitions of the LTS are intertwined and are complicated by sociolinguistic, social, cultural, and political factors. These factors play a major role in defining test standards. A test standard should be specific to its context: it has a political side since most prominent standards are government initiatives. Standards are also social and cultural because they are meant to serve the community where they were developed and thus they must serve their specific needs. This paper raises critical perspectives on using one set of standards outside of their context. The second TFTS must define LTS as part of updating the TSTF report not only in terms of guidelines or language performance indicators but also in taking into consideration the political, social and cultural context of the standards.

**From Code of Practice to Standards: Auditing the quality profile**

Nick Saville, *University of Cambridge ESOL* (saville.n@cambridgeesol.org)

This paper reports on the ALTE Procedures for Auditing and the outcomes of audits conducted in 2007. It summarises the background to the approach, presents data collected so far and reflects on the lessons to be learnt in implementing this approach. The need to justify the use of language assessments to stakeholders using codes of practice and quality management systems forms part of this discussion.

The ALTE Procedures for Auditing derive from ongoing work on the ALTE Code of Practice and draws on approaches to auditing adopted by other organisations (e.g. ISO – the International Standards Organisation). The formal scrutiny of standards is the culmination of a long process of working towards establishing audited “quality profiles” across the ALTE examinations. The aim of the process is to allow ALTE members to make a formal, ratified claim that a particular test or suite of tests has a quality profile appropriate to the context and use of the test based on 17 minimum standards. Ultimately this is to ensure that the assessment is fair and meets the needs of the stakeholders in appropriate ways.

The early work took place in the 1990s and led to the publication of the ALTE Code of Practice in 1994, and at about the same time, the adoption of the ALTE Principles of Good Practice. In many ways the ALTE Code of Practice bears similarities to the ILTA Code of Ethics which was extensively discussed by the members of ILTA in the late 1990s (first published in Language Testing Update in 2000). The ALTE Code of Practice and the ILTA Code of Ethics are similar in that both approaches show a concern for the broad principles of professional behaviour by those involved in language assessment and both address the kinds of action that should be taken in order to achieve good practice, social justice and fairness. The ALTE Principles of Good Practice are also congruent with the recent work on the IILTA Code of Practice (Version 3, 2005).

Having established a Code and the appropriate Principles, the issue of how to put the principles into practice was
addressed within ALTE towards the end of the 1990s. The appropriate paradigm for this activity was thought to be that of Quality Management Systems (QMS) e.g. as represented by the ISO 9000 series. In such systems the aim is to improve the products and/or services of an organisation in order to meet the requirements of the clients and stakeholders. As part of this phase of work, the ALTE Code was redesigned and expanded into practical checklists.

In line with the QMS practice, acceptable standards are established through the implementation of the QM approach itself and the procedures for monitoring standards are initially based on self-assessment and peer-monitoring. External auditing is then introduced to confirm that the minimum standards are being met and to highlight opportunities for improvement. In the ALTE case, it is not the aim to make all 31 members conform to the same models of assessment, and it is important to recognise the varied linguistic, educational and cultural contexts within which the examinations are being developed and used around the world. An appropriate balance is therefore required between the need to guarantee the professional standards to users and the need to take into account the differing organisational features and contexts in which the exams are used.

Taking the Code of Practice and QMS checklists into account, the 17 minimum standards are now used to establish the quality profile for an exam or suite of exams. In 2006 Auditors were appointed by the ALTE membership and a standing committee on Code of Practice and Quality Management issues was also established. In 2007 the auditing system was put into practice and by the end of the year all members had been audited or were in the process of being audited. The data from these audits will be used in this talk.

(For further information on ALTE see www.alte.org)

Session 3

Time: June 26, Thursday, 1:50 - 3:20  Location: Room 227, East 6

Justifying intended test effects to stakeholders by using effect arguments: A case of development and validation of an ESL writing test

Jiyoung Kim, University of Illinois, Urbana-Champaign (jykim@uiuc.edu)

To justify the use of an assessment to stakeholders, an interpretive argument needs to include a chain of inferences from observation to decisions (Kane, 2001; Bachman, 2005). Additionally, an argument linking intended effects and effect-driven test design needs to be articulated and investigated (Fulcher & Davidson, 2007), especially when an assessment aims for test-driven effects, for example, positive washback. Until now, however, an argument-based approach has not been extensively employed in positive washback studies.

This study reports on the development and the validation of a diagnostic reading-to-write (integrated writing) test for an ESL writing program of a large state university. Observing problems or issues related to integrate writing tasks and diagnostic tests, it projected four intended effects. Following Toulmin’s (2003) argument structure, arguments on the four effects were formulated, containing an intended effect (claim), effect-driven test design (warrant), rationales for the design (backing), observation on the effect (data), and counterclaims for the intended effect (rebuttal).

Four intended effects are: (a) with a process-oriented test design, the writing test provides cognitive scaffolding to test-takers (Effect Hypothesis 1) and also assists ESL instructors’ and researchers’ understandings of skill dependency in reading-to-write process (Effect Hypothesis 2), and (b) with on-going assessment and feedback, the diagnostic test helps students’ understandings of their own skill profiles (Effect Hypothesis 3) and also assists them to monitor and guide their own learning (Effect Hypothesis 4). As an assessment framework, the study used the Evidence-Centered Design (ECD) and the ECD design served as backing and warrant.

For validation on the effects, the study adopted a mixed-method approach for data collection and analyses to
increase defensibility of data and to reflect different stakeholders’ views. Statistical analyses on test scores and student questionnaires, qualitative analyses on student/teacher interviews and think-aloud protocols, and textual analyses on student writing were conducted to validate the intended effects. This presentation will report the findings of the study, focusing on (a) how four effect claims were supported by the multiple data, and (b) what kinds of counterclaim were observed. Then, it will discuss suggestions and implications of the study.

Large-scale oral proficiency assessment: From the perspective of stakeholders

Weiping Wu, Chinese University of Hong Kong (wwpclc@cuhk.edu.hk)

As developers of high-stakes assessment tools, we need to know what the test takers would say as stakeholders about these tools from their perspective. Reported in this paper is a large-scale investigation involving over 8000 voluntary participants to find out how test takers view an assessment tool that measures the proficiency of dialect speakers in communication using standard Chinese, which is known as Putonghua in China and carries the official status of the standardized language of the country. Based on information provided by the feedback forms collected during the investigation, several analyses were conducted and findings from these analyses are discussed in this paper.

Design of the study: A number of selected speaking tasks similar in principle to those used in the Simulated Oral Proficiency Interviews (SOPI by the Center for Applied Linguistics in Washington D.C.) were created to form the database. A sample test with three tasks (instead of 12 tasks in the real test which permits simultaneous large-scale administration) was then randomly generated by the computer from this database each time a volunteer took the test. Every test taker, upon completion of the sample test with one task from one proficiency level (i.e. Intermediate, Advanced and Superior according to the ACTFL Guidelines), would be required to provide feedback, on a 1-5 scale indicating their degree of agreement (from Strongly Agree to Strongly Disagree) to the four statements below:

Q1. The format of this test is stimulating and interesting.
Q2. The content areas of this test are closely related to language use in daily life.
Q3. This test can help me increase my interests and understanding of Putonghua.
Q4. My responses during the test indicate my ability to speak Putonghua.

Findings from the study: Of the 7904 valid feedback forms received from all the participants, well over 85 percent of them, regardless of their performance ratings in the sample test, expressed agreement (including Strongly Agree, Quite Agree and Agree) to the four statements above. Further analyses include a closer look at the responses by various groups within the population, which was categorized by gender, age (children vs grown ups) and occupations (students, teachers, government employees, taxi drivers, sales persons, and so on), and comparisons of their feedback in various combinations.

Discussion and concluding remarks: For most test takers in dialect areas who are eager to prove their speaking proficiency in Putonghua, it is common (though not yet verifiable by large scale experiments) for them to accept a test as a fair measurement tool if they perform well. For those who find themselves below the cut off line, however, it is much easier to complaint about the test, from form to content, and much harder to willingly recognize its validity. Many findings from this investigation offer some interesting insights that will help us understand why a test works, or does not work as a measurement tool, especially when no administrative or political support is available. If acceptability of an assessment tool by test takers as stakeholders is important to test developers and administrators, it is high time we tired to learn more about how a high-stake test is viewed by those who need to take them and, based on findings from these studies, ask ourselves what we should do to develop a test that test takers are willing, and not forced, to take.

Engaging stakeholders in the development of a comprehension test for an indigenous language
Gillian Wigglesworth, *University of Melbourne* (gillianw@unimelb.edu.au)

Assessing the health of indigenous languages, in both the community and the educational context, presents particular problems for language testers which are quite different from those encountered in more conventional testing situations (McGroarty 1995). In particular, to ensure fairness and validity, the inclusion of indigenous community members in the process is critical.

This paper discusses a variety of issues which impact on the ways in which we can assess the degree to which indigenous languages are being both learned and spoken by children in Australia, and proposes a range of alternative assessment approaches which can offer a means of measuring the receptive skills children are acquiring their local indigenous languages in community and/or classroom contexts. The paper also reports on the process of developing one particular test for assessing the receptive language skills of young children (5-8) in the local indigenous language, Walmajarri. The test development process involved working with indigenous stakeholders to ensure that items were appropriate plausible and feasible in the remote community contexts in which these children live.

A substantial number of the world’s languages are endangered, although the extent to which they are endangered varies across different communities. Children’s comprehension of indigenous languages in particular is a crucial indicator of the health of the languages, and although number of typologies have been identified which are indicative of the state of the health of a particular language, they do not involve the detailed evidence of the child’s understanding that can be elicited through more formal testing. The appropriate assessment of users’ proficiency in these languages has implications for language revitalisation, education and for the identification of potential language delays in children. The implications of appropriate, fair and accurate assessment of children’s comprehension skills in these languages for both education and language revitalisation are discussed.

Session 4

**Time**: June 27, Friday, 8:30 - 10:00  **Location**: Room 227, East 6

**L2 test takers’ cognitive load as a potential rebuttal resource**

Toru Kinoshita, *Nagoya University* (kinoshita@gsid.nagoya-u.ac.jp)
Harumi Oishi, *Gifu Shotoku Gakuen University* (oishi@gifu.shotoku.ac.jp)

Like many fields of research, Language Assessment has increasingly become interested in its social accountability, of which various stakeholders are regarded as its primary audience. Reflecting this tendency, many studies have been published concerning this issue in language testing. Relatively recent examples include Shohamy (1998), applying the concept of critical applied linguistics to language testing; Lynch (2001), emphasizing the ethical aspects of consequences of test use; Kunnan (2003), introducing his “test fairness framework”; Luxia (2005), reporting the absence of expected positive side effects of test washback in China; NcNamara (2006), revising Messick’s legacy in language testing validity; and Ockey (2007), reporting no existence of suspected different item functioning (DIF) in English language learners with regard to mathematical tasks.

Incorporating related studies, Bachman (2005) proposes a comprehensive theoretical framework for the validation procedure of language testing which affects stakeholders. Bachman introduces a unified model which consists of a “data-warrant-claim”chain with a possible “rebuttal unit” for use in the proper interpretation of observed test outcomes and appropriate decisions based on that interpretation.

The study here addresses the role of test takers’ cognitive load as a potential resource of the rebuttal part of Bachman’s (2005) theoretical framework. Using a brain imaging technique called functional near infra-red spectroscopy, we compared brain activation patterns of Indo-European (IEG, N=14) and non Indo-European (NIEG, N=15) groups on four ESL/EFL listening tasks. Results revealed that although the groups were measured as equivalent...
(averaging approximately 800) by TOEIC, the NIEG required significantly more blood to the cortex, indicating a need for significantly larger amounts of cognitive resources for the same tasks.

If decisions such as placement are made mainly based upon TOEIC scores, stakeholders—including the two learner groups, their instructors, and the school—could be negatively affected due to overlooking different levels of L2 abilities despite TOEIC-based resemblances.

**An investigation of the relationships of cognitive and metacognitive strategy use to GEPT test performance**

Jessica Wu, *Language Training and Testing Centre, Taiwan* (jwj@lttc.org.tw)

This study investigated language learner strategy use reported by 956 Taiwanese EFL learners through a questionnaire and the relationships of strategy use to language performance on a high-stakes English proficiency test: the General English Proficiency Test (GEPT) Listening and Reading, at both Intermediate and High Intermediate levels. Regression analysis and structural equation modelling (SEM) were employed to explore the complex relationships between test-takers’ cognitive and metacognitive strategy use and L2 test performance. In addition, the results of intra- and inter-GEPT levels were compared. Major findings include

1. Learners used more metacognitive strategies than cognitive strategies in general;
2. Some strategies were found to be significant predictors of test performance, accounting for 14% of the variance;
3. Not every type of strategy use helped test performance;
4. Variations in the effect of learner strategy use may be associated with L2 proficiency and language skill (listening vs. reading) assessed in the test being investigated;
5. The higher achievers at both GEPT levels reported more use of cognitive and metacognitive strategies than the lower achievers in general. Some strategies were used more frequently by the higher achievers, which seemed to suggest that particular strategies are useful for EFL learners.
6. The model showing the relationships between learner strategy use and test performance was similar to Purpura’s (1997), albeit some differences were noted.

Despite the limitations, the findings have a number of implications for pedagogy and future research.

**Language assessment and the professional integration of immigrant physicians: A successful model**

David Watt, *University of Calgary* (dwatt@ucalgary.ca)

The professional integration of immigrant physicians is an essential component of many nations’ long term plans for addressing their national physician supply. In Canada, 25% of all practicing physicians are International Medical Graduates (IMG) and about 40% of the rural physician population consists of IMGs. There is a recognized need to find ways to accountably evaluate the medical knowledge, clinical skills and professional language proficiency (PLP) of immigrant physicians in order to accelerate the professional integration process. Some of the challenges of professional language assessment are establishing the importance of PLP to stakeholders and designing appropriate means of assessing it for the purpose of determining readiness for medical practice in Canada.

This presentation reports on a research project that created an assessment process for measuring the PLP of immigrant physicians and subsequently designed an instructional program to accelerate its development. The assessment process was integrated into the Objective Structured Clinical Examination (OSCE) that is part of the path to licensure. Language performance was assessed in multiple authentic interactions using actors as patients and practicing physicians as raters. Language assessment scales were adapted from the Canadian Language Benchmarks (CLB) and the Calgary-Cambridge Guide to Medical Communication. Raters were trained in both clinical skills assessment and the assessment of language proficiency. The instructional preparation program used a performance
The paper will report on the results of approximately 60 immigrant physicians across four different language assessment points culminating in the language assessment protocol used in the OSCE.

The paper will close with a discussion of the importance distinguishing a knowledge of medical language, from an academic language proficiency required for the study of medicine, from the professional language proficiency needed to meet the professional communication expectations of Canadian medical practice.

Session 5

Time: June 27, Friday, 1:50 - 3:20  Location: Room 227, East 6

Investigating secondary school ELL students’ literacy skill profiles using cognitive diagnosis modeling

Eunice Jang and Youn-Hee Kim, OISE/University of Toronto (ejang@oise.utoronto.ca, younkim@oise.utoronto.ca)

A rapid increase in the English Language Learning (ELL) student populations in Canadian urban schools has become a serious challenge to educators. They are charged with daunting tasks to ensure success in the literacy development of students who have limited access to the language of instruction. The Ontario Secondary School Literacy Test (OSSLT) is designed to determine whether students have the necessary literacy skills for success in high school. The test items are developed by educators based on specifications of the curriculum and learning objectives. Students who do not pass the OSSLT can retake it at least twice, and those who fail the test twice must fulfill the requirement through the Ontario Secondary School Literacy Course (OSSLC). Successful completion of the OSSLT or OSSLC is a graduation requirement. The test score reports not only include the level of attainment with reference to the curricular standards at four levels but also provide detailed diagnostic feedback on weak skills. These current diagnostic reports are based on average subscores of items associated with six literacy skills identified by the test developers. The present study was motivated by a need to develop dependable diagnostic skill profiles with detailed information about both weaknesses and strengths in literacy skills. A cognitive diagnosis modeling (CDM) approach was applied to develop diagnostic skill profiles of ELL students based on their 2006 OSSLT performance. The study focused specifically on the following questions: (1) What are the characteristics of literacy skill mastery profiles of ELL secondary school students; (2) How do the ELL students’ skill profiles characteristically differ from those of English speaking students?; (3) How are the students’ literacy skill profiles characterized by students’ self-reported literacy learning behaviours?; and (4) How similar or dissimilar are the skill profiles developed from the CDM approach to the diagnostic reports currently in use? A total of 7,611 students comprised of 3,675 ELL and 3,936 English speaking (L1) students were randomly sampled from the grade 10 OSSLT test taker population (N=208,000). Two Q matrices based on 54 test items (48 multiple-choice and 6 construct-response items) were used for cognitive diagnosis modeling. The first Q matrix, which included 6 skills used in the current diagnostic reports, featured one skill per item and varying numbers of items across the skills. The research team developed an alternative Q matrix through the content analyses of the test items and curriculum expectations. This alternative Q matrix included six skills with minimum 7 items per skill. The sample test takers’ skill profiles were estimated using Hartz et al.’s Fusion Model (2002) and were further triangulated by von Davier’s General Diagnostic Model (2004). Differences in the estimated skill profiles between L1 and ELL students
were tested using MANOVA. Structural Equation Modeling was used to examine the relationship between the estimated skill profiles and students’ self-reported literacy learning behaviours. The preliminary study results indicate marked differences in literacy skill profiles between L1 and ELL student groups. These test taker groups also exhibit different learning behaviour patterns in terms of the amount of time spent for reading and writing, the use of a computer, and kinds of reading materials. The study results can add considerably to improving ELL secondary school students’ literacy skills and are of interest to policy makers and educators who need explicit information about skills where ELL students need the most improvement. The paper discusses the implications of the study results for various stake holders.

Criterion-related validity of the TOEFL iBT Listening section

Yasuyo Sawaki and Susan Nissan, Educational Testing Service, Princeton, New Jersey (ysawaki@ets.org, snis-san@ets.org)

Previous criterion-related validity studies of EAP (English for academic purposes) tests have often employed a summative measure such as grade point average (GPA) as the criterion of academic success. Despite the importance of such measures on various academic decisions made about students, the aggregated nature of such measures makes it difficult to make inferences as to how nonnative speakers of English actually perform in the classroom. Given that the interpretability of results of criterion-related validity studies hinges on the construct representativeness of the criterion measure (AERA/APA/NCME, 1999), it is imperative to seek an alternative measure of academic success that better reflects what nonnative speakers of English actually have to do with the language in the academic domain. As an attempt to address this issue, the present study investigated the criterion-related validity of the TOEFL iBT Listening section by examining its relationship to a criterion measure designed to reflect language use tasks that nonnative speakers of English would often encounter in everyday academic life: academic lecture listening.

The design of the criterion measure was informed by 145 undergraduate and graduate students’ survey responses on the frequency and the importance of various classroom tasks that require academic listening as well as the relationship between various class activities and assignments and the successful completion of academic courses. The criterion measure consisted of three video-taped lectures, each covering a different content area (physics, history, and psychology), and included tasks created by (former) professors of the relevant content area. These tasks reflected what the professors expected students to have comprehended during the lecture.

The criterion measure and the TOEFL iBT Listening section were administered to nonnative speakers of English who were enrolled in undergraduate and graduate programs at eight institutions in the U.S. and Canada. Data from 221 participants were analyzed. Substantial correlations were observed between the criterion measure and the TOEFL iBT Listening section score for the entire sample and for subgroups based on gender, academic level and field of study. The obtained Pearson correlation coefficients, ranging from .56 to .74, and the disattenuated correlations, ranging from .62 to .82, were generally higher than the validity coefficients typically reported in previous criterion-related validity studies of English language measures. Moreover, the mean scores on the criterion measure were compared across the High-, Intermediate- and Low-scoring groups (as reported in the TOEFL iBT Examinee Score Report). The Intermediate-scoring group (i.e., those scoring between 14 and 21 on the TOEFL iBT Listening section) represented the listening ability level required for admission to many undergraduate and graduate programs based on the results of previous TOEFL iBT standard setting studies and cut scores reported by TOEFL iBT score users. The results suggested that those participants who were classified in the High- or Intermediate-scoring groups scored, on average, nearly 50% or more on the criterion measure, demonstrating reasonable comprehension of the academic lectures in the video sets.

Validating the use of a single rating scale to assess heritage and non-heritage learners’ writing ability: A
triangulation approach

Nathan Carr and Jinghui Liu, California State University, Fullerton (ncarr@fullerton.edu, jinghuiliu@fullerton.edu)

Although there is research on the differences between heritage and non-heritage learners (e.g., Iwashita & Liem, 2005; Byon, 2005; Kondo-Brown, 2005), the appropriacy of using the same assessments for both groups has received little attention to date. This is an important question, however, for the increasing number of institutions that provide separate instructional tracks for heritage and non-heritage learners. This study investigates the adequacy of using a single rating scale to assess both heritage and non-heritage students who wish to demonstrate intermediate-level writing proficiency in Chinese at a large public university in the United States. To that end, three research questions were posed:

1. How easily differentiable are writing samples by heritage and non-heritage learners?
2. To what extent do ratings of the two groups’ writing reveal differences in terms of content, organization, cohesion, grammar, and vocabulary?
3. How adequate did raters find each scale for describing heritage and non-heritage learners’ writing?

Students in two intermediate-level Chinese courses, one for heritage learners and one for non-heritage learners, wrote in response to the same two prompts. Their responses were scored using the same analytic rating scale. To explore the adequacy of this rating scale to describe the performance of these two groups, multiple sources of information were collected: test takers’ scores, qualitative descriptions of test takers’ writing, the opinions of the raters who scored the compositions, and a survey of native speakers’ impressions of the writing samples. It is hoped that this use of diverse types of data and analyses will lead to a more principled justification of the use of the same rating scale for both heritage and non-heritage learners.

Session 6

Time: June 28, Saturday, 9:15 - 10:45  Location: Room 227, East 6

Investigating the performance of Indian raters in scoring TOEFL® iBT Speaking and the effectiveness of a special training designed for scoring Indian speakers

Xiaoming Xi, Educational Testing Service, Princeton, New Jersey (xxi@ets.org)

Although non-native raters are frequently used in many large-scale speaking assessments, there has been inconclusive evidence regarding the impact of raters’ familiarity with examinees’ L1 on their evaluations of examinees’ speaking proficiency. Some research has shown that familiarity with the speaker’s accent facilitates comprehension and may thus lead to more lenient evaluations of the overall speech quality (Brodkey, 1972; Smith & Rafiqzad, 1979; Smith & Bisazza, 1982) whereas other studies have supported lower tolerance of their peers’ speech by non-native speakers of the target L2 than native speakers (Fayer & Krasinski, 1987; Sheorey, 1985). These conflicting findings may be partially due to differences in whether naïve or trained raters are used, how adequately the raters are trained, and whether the raters are certified. In particular, few investigations have looked into whether rigorous training and certification procedures could minimize potential bias introduced by greater exposure to examinees’ L1. Also, no previous research has examined what kind of training may mitigate the potential negative effects of raters’ familiarity with examinees’ L1.

This study investigated the scoring of the TOEFL iBT Speaking Section by bilingual speakers of English and Indian dialects. It attempted to explore whether Indian raters, after being trained and certified, were able to score TOEFL examinees with mixed L1 backgrounds, especially those speaking an Indian dialect, accurately and consistently despite their greater familiarity with the Indian accents than regular raters. The effectiveness of a special training package designed for scoring Indian speakers was examined as well.

The Indian raters were tested for their speaking proficiency, went through an extensive on-line training program...
and completed a rater certification test. The 26 selected raters were randomly divided into two groups and participated in two on-site scoring sessions in Mumbai. In the first session, both groups received regular training for scoring the TOEFL iBT Speaking Section, which was largely similar to that received by raters in North America. After being trained on each of three items, they rated 100 responses of Indian and non-Indian languages on each item. In the second scoring session, the first group continued to receive the regular training while the second group was trained using a special training package. This special training involved using benchmarks and calibration samples of Indian speakers, along with those of mixed-L1 speakers commonly used in operational rater training.

The correlations and weighted kappas between the scores assigned by the Indian raters and by operational raters were computed for different groups across the two scoring sessions. In addition, Generalizability (G) studies were conducted to examine the reliability of the scores on Indian and non-Indian responses for the two rater groups across the two scoring sessions. A rater feedback survey was given to all the raters at the end of the first scoring session to elicit their feedback on the training and their scoring experiences. Another survey, focusing on the effectiveness of the special training, was given to the second group after their completion of the second scoring session.

It was found that with training similar to what operational raters receive, the Indian raters performed as well as the operational raters in scoring both Indian and non-Indian speakers. In addition, the special training helped the raters score Indian speakers more consistently, leading to increased score reliability estimates. It also boosted raters’ confidence scoring Indian speakers. The implications of the findings for expanding the rater pool for TOEFL iBT Speaking Section will be discussed.

**Test-taker attitudes towards diverse-accented speakers in EAP listening assessment: The verbal guise approach**

Luke Harding, *University of Melbourne* (lharding@unimelb.edu.au)

Recent literature that has considered the issue of accented speech in listening assessment has focused on the rationale for greater accent diversity (Jenkins, 2006), the practical constraints of the testing process (Taylor, 2006), or the empirical question of test bias (Major et al., 2002; 2005). Little is known, however, about what attitudes test-takers themselves hold towards speakers of varieties of English in the context of a listening test; yet these stakeholder views are important sources of evidence in a validity argument. Furthermore, the relationship between the attitudes that test-takers hold towards a speaker, and their subsequent test performance when listening to that speaker, is unknown.

This paper reports on findings from a larger investigation into accent and L2 listening assessment that addressed these issues. 210 participants completed a verbal guise task (see Giles & Billings, 2004) in which they evaluated the speech of an Australian English-accented speaker, a Japanese-accented speaker and a Mandarin Chinese-accented speaker delivering the same excerpt from a lecture. The verbal guise technique requires participants to provide ratings of each speaker on a range of adjectival attributes, from which listeners’ attitudes towards particular speakers may be inferred. Participants also took three versions of the Diagnostic English Language Assessment (DELA) listening sub-test delivered by the same three speakers.

Findings showed that listeners’ attitudes towards speakers on a linguistic competence trait were most positive for the Australian English speaker. However, listeners’ attitudes towards speakers on a social attractiveness trait were least positive for the Australian English speaker. On further investigation, it was found that a strong gender effect existed between female listeners and the (male) Australian English speaker. It was also found that L1 groups do not follow predictable patterns in their attitudes towards speakers of their own variety. In terms of test scores, there was little or no relationship between attitudes towards speakers and subsequent performance on a listening test delivered by that speaker.

This research has implications for the assessment of listening in contexts where there may be a strong theoretical rationale for diverse accented speech. In highlighting the heterogeneity of attitudes among test users, and the disjunction between attitudes towards diverse accented speakers and the actual comprehensibility of those same
speakers, this study serves as evidence to justify, in part, the use of accented speakers on a listening test.

Assessing the speaking proficiency of health professionals: The nurses’ criteria

Kerry Ryan, University of Melbourne (ryank@unimelb.edu)

It is a well established practice in the field of LSP testing to seek the involvement of specialist informants from the relevant occupational area in the process of language test development. This kind of specialist input is most commonly elicited at the test design phase for judgements of the authenticity and appropriateness of test materials and task design. Specialists, however, are less often involved in the rating process or in the formulation of rating criteria (Jacoby and McNamara, 1999; Douglas, 2000; 2001), although their input in these areas is arguably critical for test validation purposes.

This study investigated the ‘real world’ relevance of the linguistically based rating criteria currently used to assess performance on the speaking subtest of the Occupational English Test (OET), a test used to assess the English language competency of overseas-trained health professionals wishing to practice in Australia.

A group of native English speaking nurses/midwives first completed a pre-study questionnaire designed to elicit their beliefs and attitudes to oral communication in the health workplace and their expectations of colleagues from non-English speaking backgrounds. The nurses/midwives then listened to authentic speech samples from actual past candidates of the OET and were asked to assign proficiency ratings for each candidate on a 6-point scale, so designed as to mirror the 6-point scale currently used by the trained ESL professional raters who currently assess candidates. Each performance was then followed by a focus group discussion between the nurses/midwives where opinions of the preceding performance(s) were exchanged.

The focus group data were subjected to qualitative analysis and were described and compared with pre-study questionnaire data as well as with ratings given by (linguistic) professional raters.

The study demonstrates that nurses/midwives use a range of criteria similar in many ways to (linguistic) professional raters but markedly different in others. Findings suggest that the existing linguistic criteria may penalise, and perhaps confound, features of candidate performances that health professionals find desirable for workplace communication lending weight to the argument that language ability and background knowledge are intertwined (Douglas, 2001) and that separating them for assessment is problematic. The study also provides insights into the suitability of focus groups as a tool for eliciting subject specialist views and incorporating them in the process of developing LSP assessment criteria.

Session 7

Time: June 28, Saturday, 11:00 - 12:30  Location: Room 227, East 6

Examining the effect of response format type on construct validity to inform score uses of an EFL listening comprehension test

Huan Wang, University of California, Los Angeles (wanghuan@ucla.edu)

Many studies have been done to identify construct-irrelevant factors, and some of them have estimated the effect sizes of the identified factors with empirical test data; few studies, however, have used the estimated effect sizes to inform actual uses of the test scores. The present study tried to bridge the three areas under the guidance of the Assessment Use Argument (AUA, see Bachman, 2005). More specifically, the study addressed a particular concern raised by stakeholders that the response format type may be a construct-irrelevant factor that would question the appropriateness of the intended uses of the test scores. The study performed structural equation modeling (SEM)
analyses to examine if the factor of the response format type exists for the test under concern and then estimated the effect sizes of the observed format factors. The estimates were then used to help evaluate the appropriateness of the intended test uses in terms of relevance and intended consequences. The study results indicate the existence of the factor of the response format type, and the observed format type factors do not have much negative impact on the quality of relevance, although the quality of relevance is not satisfactory regarding either type of the test uses. With regard to consequences, however, the format type factors could have considerable negative impact. The study has practical values to the test stakeholders by addressing their concerns and informing their uses of the test scores with statistical evidence. The study also sets up an example of evaluating the effects of potential construct-irrelevant factors with empirical data and in light of the actual test uses.

Effects of presentation mode and computer familiarity on performance in reading-to-summarize extended English texts in two languages

Guoxing Yu, University of Bristol (Guoxing.Yu@bristol.ac.uk)

This paper reports the effects of presentation mode of source texts and students’ computer familiarity on their test performances in reading-to-summarize extended English texts in two languages. It is part of a larger study which investigated the premises, promises, problems and the compromises of using reading-to-summarize tasks as a measure of reading comprehension (Yu 2005, 2007, forthcoming). One hundred and fifty-seven Chinese undergraduates were asked to read one of the three extended English texts (around 2200 words each), presented either on computer screen or in print. All participants then wrote their summaries on paper in both English and Chinese in randomized order under examination condition. The effects of presentation mode and computer familiarity were investigated through two parallel datasets: students’ actual summarization performance (in terms of the key information included in the summaries, the holistic quality and the length of the summaries) and their perception of such effects as demonstrated in the post-summarization questionnaire and interviews. It was found that both text presentation mode and computer familiarity had some significant main effects on student actual summarization performances and that the effects were more pronounced on Chinese than English summarization. In particular, the only significant main effect of presentation mode was on the length of Chinese summaries. In addition, the Chinese summaries written by students of lower computer familiarity were unexpectedly of a statistically significantly higher quality than those written by students of higher computer familiarity. Some interactive effects of text presentation mode and computer familiarity with text type were also noted. However, these statistical findings from the actual performance data were only partly supported by the student perception data. The students did not think text presentation mode and computer familiarity would affect their summarization performance to a great extent. Those who were interviewed also acknowledged that there was not much difference between the two presentation modes, but they also articulated some minute physical and psychological differences (e.g. the historical friendliness, better tangibility, security and visibility of reading on paper) between the two modes and their potential effects on summarization process. Nevertheless it was also agreed that (lack of) familiarity with using computers did not help or hinder summarization performance because basic computer manipulation skills that they all had were sufficient for reading the source texts on screen in this research. The interview data also indicated that the effects of computer familiarity might be more as anticipated than actually experienced by the participants. Implications of these findings are discussed with specific reference to (1) the design of language test tasks which involve reading an extended text over several pages on computer, in contrast to many computer-delivered language tests using multiple choice formats which very often focus on understanding of details rather than main idea comprehension such as summary writing, and (2) the voice and the value of test takers – arguably the most important key stakeholders – in language test design and validation (see also Yu 2007).

A meta-analysis of effects of multiple-choice and open-ended formats on reading and listening test
In order to more accurately assess learners’ language ability, according to the purpose in each context, it is significant to examine how learners’ scores are affected by the method of measurement. Among a wide spectrum of variables that are considered to affect language test performance and must be taken into consideration in order to draw more valid inferences from test scores (e.g., Bachman, Davidson, & Milanovic, 1996; Kobayashi, 2002), one central issue is the effect of test formats on test performance (e.g., Alderson, 2000; Berne, 1992; Bernhardt, 1991; Brantmeier, 2005; Buck, 2001; Wolf, 1991). Although this topic has been researched in the fields of language learning and educational measurement, it appears previous studies have suffered from taking a narrative approach to accumulate their findings. This approach has been criticized because (a) it has less objectivity and replicability due to individual differences among reviewers, (b) it mainly builds on conclusions drawn by the authors of original studies without reanalysis or reinterpretation, and (c) it has difficulty handling the rich volume of information extracted from studies (e.g., Cook et al., 1992; Cooper & Hedges, 1994; Norris & Ortega, 2000). To address these limitations, the current study uses meta-analysis to quantitatively synthesize effects of multiple-choice and open-ended formats on L1 reading, L2 reading, and L2 listening test performance.

Fifty-six data sources located in a thorough search of the literature were the basis for the estimates of the mean effect sizes (g; Hedges & Olkin, 1985) of reading and listening test format effects. The effect sizes were computed for each format comparison in a study. Then the average effect size and its 95% confidence interval were computed and interpreted. The results using the mixed effects model of meta-analysis indicated that multiple-choice formats were easier than open-ended formats in L1 reading and L2 listening, with the degree of format effect ranging from small to large in L1 reading and medium to large in L2 listening. Overall, format effects in L2 reading were not found, although multiple-choice formats were found to be easier than open-ended formats when any one of the following four conditions was met: the studies involved between-subjects designs, random assignment, stem-equivalent items, or learners with a high L2 proficiency level. Format effects favoring multiple-choice formats across the three domains were consistently observed when studies employed between-subjects designs, random assignment, or stem-equivalent items.

Session 8

Time: June 28, Saturday, 2:00 - 3:30  Location: Room 227, East 6

The characteristics of academic English language and DIF on items in an English language proficiency test and standardized tests of math and science

Lyle Bachman, Applied Linguistics/ University of California, Los Angeles (lfb@humnet.ucla.edu)
Tim Farnsworth, Mihyung Kim Wolf, Sandy Chang, Patina Bachman and Seth Leon, CRESST/University of California, Los Angeles (tim.farnsworth@gmail.com, miwolf@cse.ucla.edu, sandychang@ucla.edu, bachman@cse.ucla.edu, leon@cse.ucla.edu)

When students take tests of academic achievement that are administered in a language that is not their mother tongue, there is always the question of the extent to which their test performance reflects their academic achievement as opposed to their proficiency in the language of the test. In the U.S., as part of the accountability requirements of the No Child Left Behind Act, millions of non-native English speaking students, or English Language Learners (ELLs), are required to take both an English language proficiency (ELP) test and several academic achievement tests. Since these high-stakes standardized tests are administered in English, it is critical to understand the extent to which the Academic English language (AEL) “demands” of items on these tests may lead to differential item performance between ELLs and non-ELLs. In addition, since high-stakes decisions are commonly based on the results of an ELP test, it is important to determine the extent to which the AEL demands of these tests
are comparable to those of academic achievement tests.

The present study aims to investigate the characteristics of AEL on both ELP and academic achievement tests in math and science, as a way of determining the AEL demands that these place on ELLs, and to relate this to items on academic achievement tests that show differential item functioning (DIF) between ELLs and non-ELLs. To analyze the AEL content of test items, a content analysis protocol was developed based on the research into the nature of AEL (Bailey, 2007; Chamot & O’Malley, 1994; Cummins, 2000; Scarcella, 2003; Schleppegrell, 2001). The actual math, science, and ELP test items that two states implemented at two grade levels were analyzed. Simultaneously, DIF analyses of these same items were conducted using ELL and non-ELL as focal and reference groups respectively. Items found to have DIF were compared on their linguistic features against items without DIF.

The results suggest that academic vocabulary was the most salient feature in determining linguistic complexity for both ELP and content-area tests, and that this academic vocabulary may be a source of DIF. The linguistic analysis showed that the AEL demands were more varied between states than between subject matters. The result of the DIF analysis was also congruent with this result. On the other hand, there was a clear mismatch between the amount and type of AEL on the ELP and the content tests. Finally, we discuss the methodological challenge of searching for language related DIF given that nearly all of the test items involved contained at least some level of language demand.

Learning to write essays or teaching to crack tests: An empirical study investigating the washback and impact of TOEFL test-preparation in China

Michael Chau, Hong Kong Polytechnic University (michael.engl@polyu.edu.hk)

There have been claims of the negative effect of “teaching to the test” of the TOEFL and the “ceiling effect” of the consistently high TOEFL scores from Asian students. On the other hand, some researchers (see, for example, Hughes, 1988) assert that tests can exert positive effects. However, there is little in the research literature about what actually happens in test preparation classrooms. This gap is particularly troubling in relation to large international and high-stakes tests such as TOEFL and IELTS, where there are many commercial providers of test preparation courses, especially in China.

The present study focuses specifically on TOEFL preparation courses and collects empirical evidence concerning the washback and impact of test-preparation instruction on students’ English language learning, especially the learning of academic writing. This study adopts a case study methodology and collects data through a field study at a well-established test-preparation school in China. By conducting participant observation in an intensive TOEFL preparation course, observing classroom teaching activities and student behaviours, and by interviewing TOEFL teachers and students, a detailed and vivid picture of how teachers and students prepare for the TOEFL in China is presented.

This study investigates the test preparation for the written component of the TOEFL (formerly the Test of Written English [TWE] now built in to the TOEFL iBT), in order to understand to what extent the course helps student improve their English writing. The research findings show that most of the instruction was centred on the official TWE prompts, which were used too extensively and were probably “abused” in the class. The teacher tried to prepare the students for TWE in a “standardized” way, which did not aid students’ learning of academic writing. Students’ learning of academic writing was largely restricted to the genre of the TWE, a short snapshot of writing performance that tends to discourage critical thinking and expanded argument. In addition, it is found that most of the beliefs about good teaching of writing are overlooked in the test-preparation classroom. Writing is regarded more as a standardized product utilizing the “five-paragraph essay” approach than a recursive process, which is also a tool of thinking. This empirical study concludes that the method of teaching “writing” in the TOEFL test preparation course is likely to have a negative washback effect on students’ attitudes to writing in English as well as on their ability to write well. Finally, the above issues are discussed in connection with the pedagogic and social implications arising from the findings of the study.

Impact and consequences of school-based assessment in Hong Kong: Views from students and their
As part of the worldwide movement to combine assessment of learning with assessment for learning in motivating students’ learning, standards-referenced school-based assessment (SBA) has recently been introduced into the Hong Kong Certificate of Education Examinations (HKCEE) in English, taken at the end of Secondary 5. The HKCEE SBA component in English focuses specifically on the assessment of students’ spoken English. This innovation, which was introduced into Secondary 4 classrooms in 2005-2006, has already caused reverberations within the Hong Kong education system (Davison, 2007). This SBA component is expected to induce a rethinking of what teaching and learning entails and a redefinition of key components of assessment such as validity, reliability, and fairness at the philosophical, theoretical, and practical levels. It is within the context of this assessment change that we are currently conducting an impact study, focusing in depth on the impact of SBA in three contrasting secondary schools.

Given that students are most directly affected by this SBA assessment change, yet have been researched the least among all stakeholders in previous impact studies, we designed our study to include a major focus on students. Previous research has also demonstrated that parents in Hong Kong often play an important role in shaping teaching and learning. We therefore took care to include parents in our study. The proposed presentation focuses on the impact of SBA as perceived by the students and parents in our impact study.

The data discussed in this presentation were gathered from large-scale surveys of students from the three case-study schools and their parents. Both surveys explored students’ and parents’ perceptions of SBA and their perceptions of the impact of SBA on learning. The students’ survey comprised 68 items and the parents’ survey 30 items, with both sets of items using a 6-point Likert-scale. Data were analyzed in three stages: first, descriptively on the item level; second, using exploratory factor analyses to plot the major constructs of the surveys; and third, using t-tests and multiple regressions to explore the relationships within and across the surveys.

Briefly, the results demonstrated a relationship between students’ perceptions of SBA-related learning activities and their perceptions of their own language competence. Students’ views of external examinations (which still form the majority of the components of the HKCEE) and of SBA showed significant difference. Unexpectedly, students were less favourably disposed towards SBA than towards external examinations at the point when the data were collected. The results also showed significant differences between students’ perceptions of the learning activities they had taken part in during the previous school year and those they were currently engaging in. This result indicated a change in students’ learning activities. In addition, we found that parents’ perceptions of SBA and the opportunities for them to know about SBA significantly and positively predicted their support for their children’s SBA learning (38% and 40% of the variance). Parents’ education level and the amount of time they spent with their child daily also predicted their support for their children’s SBA learning, though to a lesser extent (22%). Taking the results from both students’ and parents’ surveys together, we have gained a better understanding of the complexity of the impact of SBA as perceived by students and their parents. The results of this impact study will help to inform the on-going process of assessment change within the Hong Kong educational context.
Session 1

**Time:** June 27, Friday, 3:30 - 4:30  **Location:** Room 118, East 6

A validation study of the use of STEP EIKEN Test as an alternative English proficiency screening test for college admission purposes

Yao Hill, *University of Hawai‘i at Manoa* (yaohill@gmail.com)

Over years, TOEFL test scores have been used as English proficiency indicators of international students for admission purposes in North American universities and colleges. However, limited testing sites and relatively high test fees restrict students’ access to the TOEFL test. In response to this problem, and in keeping with its institutional commitment to open access, a North American community college initiated an investigation into alternative English proficiency admissions tests. EIKEN’s STEP grade level test, the most widely used English proficiency test in Japan, is the first alternative that has been studied in the past two years as part of KCC’s initiative. In order to establish concurrent validity evidence, both TOEFL and STEP tests were given to 398 students from seven English training programs in Hawai‘i. The results showed overall high correlation between the two tests ($r = .79$, $p < .05$). Logistic regressions showed that the probability for a student who passed STEP Pre-2 Grade Level to score 400 on the institutional TOEFL test is about 80%, and the probability for the students who passed STEP 2A to score 500 on the TOEFL test is 72%. The probability will be higher in reality due to the fact that only about 75% the students who pass the Grade 2 paper-and-pencil test will succeed on the additional oral interview test, and about 83% students who pass Grade Pre-2 paper-and-pencil test will pass the oral test to get the grade certificate.

The college in question has taken the “protecting-students position” and decided to use STEP pre-2 certificate as English proficiency evidence for the students to be admitted into non-credit courses, and to use STEP 2A certificate to admit students into credit courses.

The current stage of the project aims to gather the predicative validity evidence of using STEP as an alternative English proficiency test by investigating potential language and academic performance differences between the EIKEN and TOEFL admitted students. The intended uses of the study are for the college admission committee to evaluate and possible adjust their decision-making in using these two EIKEN grade levels for admission purposes, and possibly to establish some academic assistance program if negative evidence were to be found showing the adversary effect of the admission decision by allowing students in the college, whose language ability has not prepared them well enough to handle academic study at North American colleges and universities.

The students’ English and academic performance were investigated both quantitatively and qualitatively. Quantitative study looked at the differences in EIKEN and TOEFL admitted students’ performance on campus English placement tests, including writing and speaking tests, their GPA, dropout rate, and performance on non-credit English and credit content courses offered by the college across two years.

The preliminary findings based on descriptive statistics showed that EIKEN population performed equal or better than their TOEFL counterparts on average, and therefore, provided positive evidence justifying stakeholders’ decision-making.
The next step of the research is to interview both EIKEN and TOEFL population to gain deeper understanding of the similarities and differences in their academic experience in the community college and provide suggestions to the admission committee to provide assistance service to account for potential difficulty that either or both group of students have been experiencing.

Assessing EFL university students’ argumentative writing ability from a discourse perspective

Jingjing Qin, Northern Arizona University (jingjing.qin@nau.edu)

The ability to write an argumentative essay is assessed in both internationally recognized tests, such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing Systems (IELTS), and China’s national English tests, such as College English Test (CET) for non-English majors and Test for English majors (TEM). However, there have been relatively few studies assessing second language (L2) university students’ argumentative writing ability from a discourse structure perspective (Connor & Mbaye, 2002, Crammond, 1998).

To fill in the gap, the study is intended to examine how the use of elements of argument structures, based on Toulmin’s model (1958) (i.e., claim, data, warrant, counter-argument, and rebuttal), in EFL university students’ argumentative essays is related to the overall quality of essays. After reading two short editorial pieces (approximately 650 words each) with opposing views on the same controversial topic, 106 EFL Chinese sophomores majoring in English were asked to write an argumentative essay. Half of the participants wrote on a familiar topic, and the other half wrote on a less familiar one. The familiarity of topics is defined as the distance of the topic to writers’ personal experiences (Stapleton, 1998). From a discourse perspective, an analysis was conducted to tally the frequency of the different elements of Toulmin’s model presented in the essays (Crammond, 1998; McCann, 1989). The essays were also scored holistically based on a 5-scale rubric adapted from McCann (1989) and Nussbaum and Kardash (2005).

A standard multiple-regression is performed between the holistic rating of essays as the dependent variable and the frequencies of claim, data, warrant, counter-argument, and rebuttal as independent variables. Two independent t-tests are conducted to examine whether the holistic rating of essays written on the two topics and the total frequency of the elements of Toulmin’s model in the essays written on the two topics differ. Based on the results of the study, implications for large-scale L2 writing assessment and instruction of L2 argumentative writing are discussed.

Assessing productive collocational competence of advanced EFL learners

Su-mi Han, Seoul National University (sumihan8@snu.ac.kr)

Recently, there is a growing recognition in the English language teaching community that advanced learners of English need to have a good command of a wide range of complex lexical units, which are often processed as prefabricated chunks, fixed, or semi-fixed expressions by a native speaker (Peter Howarth, 1998). Although many studies have examined the collocational competence of EFL learners at various proficiency levels so far, few have attempted to systematically investigate developmental patterns of collocational competence, particularly for college-level, advanced EFL learners, who are usually required to produce extended pieces of writing as part of their course requirements. At present, there seems to be an urgent need to investigate not only EFL learners’ collocational knowledge at different proficiency levels but also the relationships between the learners’ productive collocational competence and their English writing ability in academic settings.

The main purposes of the current study are to: (a) investigate the collocational knowledge of advanced Korean EFL learners, from 12th graders in high school to senior-year students at university and (b) examine the relationships between the productive collocational competence and writing ability for advanced learners of English. Participants of the study are: about sixty 12th graders attending a foreign language high school located in a large city in South Korea, another group of 60 non-English majors enrolled at a four-year university in Seoul, Korea, and finally 60
English majors of all levels studying at the same university. Based on Benson et al.’s (1997) framework, a 50-item gap-filling collocation test will be constructed and piloted for the current study. The final version of the test will be used to measure the participants’ productive knowledge of both ‘grammatical’ and ‘lexical’ collocations. In addition, the participants will be required to take an off-the-shelf general English proficiency test consisting of multiple-choice (MC) vocabulary, grammar, and reading items to demonstrate their overall proficiency of English and also to write a 30-minute essay in response to a TOEFL-type essay topic for independent writing tasks.

Descriptive statistics will be computed for the collocation, essay, and general proficiency tests for the three different groups of test-takers. Mean collocation and essay scores at different proficiency levels will be compared and examined. A matrix of correlations will be obtained among the three different measures of English proficiency for the three different groups. The correlation between participants’ collocation test score and MC vocabulary test scores will also be inspected to examine the relationship between the students’ receptive knowledge of vocabulary and productive collocational competence. Moreover, various types of word-choice errors in a smaller, representative sample of the students’ writing samples will be analyzed to identify some meaningful patterns of collocational errors for subgroups of students at different proficiency levels.

It is expected that results of the study will reveal the developmental pattern of productive collocational competence of advanced EFL Korean students. The study results are also expected to provide very valuable information that can be used to identify the types of collocations with which advanced EFL learners have difficulties and to design more effective writing classes responsive to the students’ needs. This could have important implications for the teaching of collocations and writing, especially in college-level composition courses.

Assessing vocabulary and lexical proficiency in EFL teacher education programmes

Melissa Baffi-Bonvino and Douglas Consolo, UNESP-Brazil (melissabb7@hotmail.com, dconsolo@terra.com.br)

In this presentation we report on a study conducted within a larger research project about the assessment of vocabulary and the lexical proficiency in the oral production of graduating students of English Language and Literature BA courses (henceforth Letters courses) in Brazil, who are preparing to enter the field of ELT.

The first phase of the study started in 2005 and focused on three aspects: (a) the process of oral production in EFL lessons and the assessment of lexical proficiency in classroom activities, based on the teacher’s and the researcher’s observations and recorded data; (b) the students’ beliefs and experiences concerning the teaching and learning of vocabulary in foreign language lessons, especially in spoken language; and (c) the assessment of lexical proficiency by means of two oral tests – the TEPOLI (Consolo, 2004; Consolo & Teixeira da Silva, 2007), designed for the larger project, and the speaking paper from the Cambridge FCE examination. Students took two mock speaking FCE tests during the academic year and the TEPOLI at the end of the course. The major data analysis in the first phase, carried out by means of qualitative and quantitative procedures, was primarily based on four students who participated in all the occasions for data collecting. Information from the other students and test results from the whole class were also considered for a panorama of the context investigated. The analysis of language data produced in the two tests was conducted with the aid of the RANGE programme, and the overall results from phase one indicate that students displayed better lexical competence in the oral tests than in classroom activities. On the other hand, test results indicate that the teaching of vocabulary in speaking tasks reflected in the quality of students’ oral performance.

A second phase of the study has been conducted with another class of students graduating from the same course and another research instrument – the speaking paper from IELTS, was added in this phase to collect data under test conditions. Eight students have contributed with data. The study will proceed by analysing classroom and oral testing data, and the three tests will be compared as well, mainly concerning their criteria to assess lexical proficiency.

It is expected that results from phases one and two (the ongoing study) will contribute for the definition of more objective criteria to assess lexical proficiency in English spoken by EFL teachers-to-be, for a validity study of
TEPOLI as a test of English for professional purposes and to shed some light into the teaching of specialized vocabulary in teacher development courses. It can also contribute to provide data from those courses with regards to the characteristics of those (future) teachers’ oral proficiency. These data can indicate, to stakeholders in the ELT profession, the need of formal assessment of speaking proficiency in EFL at the end of Letters courses, as well as of clear criteria on which to base such assessment.

**Building validity argument in an oral proficiency assessment project**

Raili Hilden, *University of Helsinki, Finland* (raili.hilden@helsinki.fi)

A 3-year research project, called HY-Talk, was initiated at the University of Helsinki, Finland, in 2007, with a focus on the assessment of oral proficiency in foreign languages. The financial support from the Research Grants Committee was allocated specifically for the validation of five illustrative subscales of oral proficiency included in the new national core curricula for general language education in Finland (National Core Curriculum 2003; 2004). This scale is empirically related to relevant scales of the Common European Framework of Reference. The subscales adopted to the core curricula address overall task management in terms of themes, texts and purposes, fluency, pronunciation, linguistic range and accuracy. Each of these is related to different competences underlying speaking performance.

Theoretically the design draws on the validity argumentation scheme proposed by Kane (2001), further elaborated by Bachman 2005, and most recently by Fulcher & Davidson (2007). The major claim to be probed is that the oral proficiency scales currently included in the core curricula for languages in general school education are reliable and valid tools for assessing the communicative oral proficiency of students in general language education. The claim needs to be supported by a set of grounding evidence and warrant statements derived from the data. These address relevance, usefulness, the intended consequences, and the sufficiency of information. On the other hand, the claim will be confronted with counterclaims and rebuttal data to challenge the conclusions. These count for various sources of construct-irrelevant variance and construct under-representation. Specific research tasks assigned to individual researchers is generated from the overall argumentation frame. In addition, the research activities will pay attention to language specific cultural determinants of the evolving oral proficiency. The dynamics of test-taking and student interpretation of the test tasks conducted in pairs will also be explored.

The research consortium consists of 10 experts in English, French, German and Swedish languages at the Faculty of Arts, along with experts in language education and assessment from the Faculty of Behavioral Sciences. Three sets of test tasks have been designed, targeted to CEFR levels A2, B1 and B2 as an average. The data (about 160 speaking samples) are collected from schools and university institutions and rated by a sufficient number of experienced judges. The samples, student feedback, and rating protocols, as well as records from rating sessions will be investigated from various aspects in cooperation with professional researchers and students. The methodology is many faceted including both well-established quantitative and qualitative techniques of language testing research, but in addition to these, some novel approaches are applied to gain an in-depth insight into student interpretation and perception of the tasks and the process of test taking.

**Challenges in developing comparable tests in a multilingual India**

Vyjayanthi Sankar, *Educational Initiatives Pvt Ltd* (vs@ei-india.com)

Language learning at the school level in India, a natural bi/multilingual environment, is different from compared to that of monolingual communities (CAL, Washington D.C., 2007). Children here start school with basic vocabulary and syntax acquired in more than one language (National Focus Group, New Delhi, 2007). Given this scenario of multiple languages, dialects, and the predominant percolating phenomenon of English and Anglicization, this paper attempts to share some of the challenges we at Educational Initiatives face in developing comparable and valid tests in 10 different languages (namely, Assamese, Bengali, Gujarati, Hindi, Kannada, Marathi, Oriya, Tamil, Telugu and Urdu) belonging to different language families (Indo-Aryan, Dravidian, Tibeto-Burman and Austro-Asiatic) resulting in many variations such as script and script orientation (left-right or right-left) and structure of
the language. This paper will highlight these and other issues that we confront in developing comparable tests in India; we grouped the challenges with a few examples each as follows:

Variation among the Indian languages:

- In a vocabulary test, a simple two-syllable word ullu for ‘owl’ in Hindi with no consonant clusters in simple script becomes gudlagubba in Telugu a 4-syllable word that has conjoint letters and has a difficult script. Therefore such items are not comparable items.
- In a word identification task in a Telugu language test, a picture of a ‘kite’ galipatamu in Telugu may result in multiple answers written in Telugu which are actually words belonging to other languages such as patang in Hindi or ‘kite’ in English. This type of response will need complex scoring procedures.
- In agglutinating languages such as Telugu, a simple sentence such as ‘he started running’ in English is represented by a single word such as parugul-ankinchukunnadu (started + to run). This presents a problem for test development in grammar sections across languages.

Variation as a result of use of English:

- In a mathematics test, the item ‘Subtract 43 from 7’ written in English with the subtrahend appearing before minuend that requires reordering of the sequence before computation when translated into Hindi and other Indian Languages, the item will appear without order reversal and with simple words that provides clues to the computational process as 75 se 43 ghatao (meaning From 75 remove 43). This results in variation in item difficulty across languages.
- Wide use of English has made words such as ‘traffic signal’ more easily understood compared to the more uncommon Telugu term gamana-agamana-ghanta-suchika. In other words, the English equivalent is more widely understood than the word in the local language. This results in accepting equivalent words requiring familiarity and prevalence to be accounted for in item construction.

Variation due to social, economic, geographical or cultural diversity:

- The word for ‘school’ although correctly translated into shaala in Hindi, maybe unfamiliar in different parts of the same state as other correct variations such as pathashaala may be prevalent in the different areas. Such dialectal differences have to be taken into account in test development.
- Children from the north-west part of the country such as Rajasthan may not be able to relate to the idea of sea or coconut trees. Similarly, names of persons, festivals, food, etc. referred to in stories may be familiar or alien in different communities due to the cultural diversity in the country. Topics for tasks should take this variation into account.

Collaboration and student anxiety in the group interaction task of school-based assessment

Manman Gao, The University of Hong Kong (gaomm03@gmail.com)

The Hong Kong Examinations and Assessment Authority (HKEAA) has incorporated a school-based summative oral assessment component (SBA) into the compulsory English language subject in the Hong Kong Certificate of Education Examination (HKCEE). Though SBA places great emphasis on the mediation of teachers, less attention has been paid to another group of stakeholders, i.e. students. This study focuses on the impact of SBA on students’ anxiety.

“SBA is carried out in ordinary classrooms, not a special examination hall; it is conducted by the students’ own teacher, not a stranger; and it involves students more actively in the assessment process” (SBA Consultancy Team, 2005). SBA allows for and encourages collaboration between students in a group discussion since the students are classmates or even friends (Gan, Davison & Hamp-Lyons, 2006). The use of a group discussion task assumes high authenticity, and that students’ opportunities for collaboration in such tasks can have a positive effect by lowering students’ anxiety levels. This study explores empirically students’ collaboration in the group interaction task by examining verbal interaction, gestures and eye contact, and the impact of collaborative interactions on students’ anxiety in the SBA.
Since there is little previously established data-derived theory to guide research in school-based assessment in language teaching and learning, a qualitative case study approach is utilized to inductively explore the data in search of an explanatory theory. In the present research, students’ performances in school-based group interaction tasks will be videotaped and analyzed using conversational analysis to examine their facets of interaction, particularly collaboration. Students will be verbally debriefed immediately after the SBA group discussion, reporting their felt sense of anxiety or nervousness and the effect of collaboration on them. Teachers will also be interviewed preceding and following the assessment.

In this work-in-progress session participants will be invited to view at least one sample video of a group interaction, accompanied by preliminary conversation analysis and an analysis of gesture and eye contact, and to discuss the interpretations of the data and further appropriate methods for data-mining.

Creating an end-to-end assessment solution at the Hong Kong Institute of Education

Jane Lockwood and Michelle R. Raquel, Hong Kong Institute of Education (lockwood@ied.edu.hk, mrraquel@ied.edu.hk)

Substantial funding is currently provided to HK Higher Education Institutions (HEIs) who are required to provide more and better language proficiency outcomes reports to a range of funding stakeholders. They also recruit unprecedented numbers of non-local students in an effort to develop Hong Kong as an ‘education hub’ in Asia. Added to these external contextual factors, the Centre for Language in Education at the Hong Kong Institute of Education has recently embarked on a curriculum renewal process for its undergraduate English language programmes that entail providing tailored proficiency pathways for the students. These major contextual changes have resulted in a need for the CLE to have an institutional ‘end-to-end’ English language assessment solution to provide: (1) Language screening and placement; (2) Language formative and summative assessment processes and tasks for the new curriculum; and (3) Language reporting at year end. Such an ‘end-to-end’ solution will provide a systematic approach to screening and placement assessment, formative and summative assessment and language assessment reporting for all English enhancement programmes. Hitherto, CLE (and other HEI language centres) has relied on ad hoc language assessment solutions to these institutional requirements and/or an over reliance on international certification that is expensive and limited as an end-to-end solution.

The Tertiary English Language Test (TELT) being developed at HKIEd, has received seeding funding from the Institute, to develop assessment processes for the range of purposes outlined above. This paper reports on the work achieved so far with particular reference to the screening and placement test. It will outline the TELT test specifications, stages of piloting, process of item analysis using Rasch Measurement Analysis, scales and descriptors for speaking and writing and calibration issues in speaking and writing. The paper will also report on senior management’s response to the new tool and immediate challenges for TELT in further development.

Developing reference level descriptors for higher proficiency learners of English

Tony Green, University of Bedfordshire (tony.green@beds.ac.uk)

The Common European Framework of Reference for Languages (CEFR) has, since its publication in 2001, rapidly established itself as a valuable tool for comparing assessment results from diverse systems. However, the CEFR is designed to operate across languages and acknowledges that more detailed specifications or ‘reference level descriptions’ (RLDs) will be required to adequately demarcate levels for specific languages. To date RLDs have been developed for the lower levels of the CEFR for English at A (Basic User) and B (Independent User) levels through the Council of Europe draft Breakthrough (Trim, forthcoming) and Threshold (van Ek, 1977), Waystage (van Ek and Trim, 1991) and Vantage (van Ek and Trim, 1997) publications, but not for the C levels associated with higher education and professional contexts.

This paper will report on one strand of a large-scale project designed to develop RLDs for the C levels. This strand of the project identifies language functions that appear for the first time at the C levels. A database of materials
Development and initial validation of a questionnaire measuring students’ perception of the test impact

Hongli Li and Hoi K. Suen, Pennsylvania State University (hul151@psu.edu, HoiSuen@psu.edu)
Qi Zhong, Central University of Finance and Economics (sunnyzq@hotmail.com)

In the past two decades, the impact of testing on teaching and learning has gained much attention from educators and researchers. Alderson & Wall (1993) proposed fifteen washback hypotheses, which gave clear pointers to areas that might be influenced by a test. Hughes (1994) further claimed that a test could influence the participants, the process and the products of an educational system. While many studies have been done to investigate how testing influences classroom teaching, more is needed on how testing influences students’ behaviors, not only in terms of their academic behaviors but also their affective conditions. Among the instruments that have been adopted to study test impacts, questionnaire survey is probably the most widely used one. However, there is no validated questionnaire available to assess students’ perception of test impact. Therefore, the purpose of this study is to develop and validate a questionnaire measuring students’ perception of test impact, so as to enable teachers, parents, test designers, and policy makers to receive more information about how tests influence students.

In this study, the test being targeted is the College English Test (CET) in China, which is a high-stakes national test to assess college student’s English ability. Driven by the prevailing theories on the impacts of high-stakes testing, we first generate a list of questionnaire items to measure students’ perception of the impacts of the CET. The major aspect to be addressed is how the test influences students’ study, including study method, study content, motivation for study, etc. Another aspect is how the test influences students’ affective conditions, including self-efficacy, aspiration, and anxiety level, etc. Based on feedback from expert judgment, we revise the questionnaire items and have a pilot study with a group of students enrolled in Chinese universities. Based on the findings of the pilot study, some irrelevant items are deleted and a shorter questionnaire draft is finalized. Then a larger group of Chinese students are surveyed by the questionnaire, and more statistical analyses, e.g. factor analysis, are adopted to further validate the questionnaire. It is expected that this questionnaire can be generalized to the assessment of impact of other high-stakes test on students.

Examining washback in the Taiwanese context: Voices from teachers and textbook writers

Ching-Ni Hsieh, Michigan State University (hsiehc12@msu.edu)

This study explores the “washback effect” (Alderson & Wall, 1993) of a high-stakes national exam, the Basic Competence Test in English, which takes place in junior high schools in Taiwan. The study focuses on 1) teachers’ perceptions of the test and its washback effect on teaching and 2) textbook writers’ perceptions of the test and its washback effect on textbook design.

To explore the washback phenomenon, this study employed various methodological techniques, including questionnaire surveys, in-depth interviews, and an analysis of textbooks used. On-line survey was sent out to hundreds of English teachers across Taiwan to examine their perceptions of the Basic Competence Test in English. Semi-structured interviews were conducted with three textbook writers and three textbook editors. Five major textbooks were analyzed.
The preliminary survey results revealed that most of the teachers perceived the exam to have both positive and negative washback effects. The exam had motivated teachers to work hard on their teaching and learning. Many teachers also reported that preparing for the exam was a strong motive for their students and could sustain students’ motivation. However, the exam had caused extensive use of test preparation materials and mock exams. Additionally, since the test only tests reading skill, productive skills were less focused in the language classroom.

The interviews with the textbook writers and editors and the analysis of the textbooks showed that the exam had a tremendous influence on the design and content of the textbooks. Nevertheless, the nature of the washback effect differs across textbooks and features. Other factors, such as textbook writers’ perceptions and understandings of the exam and publishers’ marketing concerns, have added another layer of washback effect in relation to the Basic Competence Test in English.

The preliminary results of this project have shown a multi-layered account of the washback phenomenon from a nationwide exam in Taiwan. It is hoped that a deeper understanding of the washback mechanism in the Taiwanese context could be gained through the research findings.

From placement test to post-test: Students’ views, teachers’/raters’ perspectives, test developers’/administrators’ rationale and institutional demands

Claudia Kunschak and Lily Zhang, Shantou University (claudia.kunschak@azalumni.com, lily@stu.edu.cn)

With the reform of English language teaching at Chinese universities, and particularly at the English Language Center (ELC) of Shantou University, new tests are being implemented across the board. At ELC, a placement test covering all four skills and including a computerized oral component was instituted to increase entrance level homogeneity from a performance rather than a solely competence-based perspective. At the same time, a post-test was introduced to measure student gains after four years of attending English language classes and participating in campus-wide extra-curricular activities pertaining to the so called English Enhancement Program. These two measurements constitute the gateway and goalpost of an integrative, communicative, skill-based English language program that is not geared towards Chinese standardized tests but focuses on the development of communicative competence, learner autonomy and critical thinking. Both tests, though with a different immediate purpose, are large-scale, rating-intensive, complex constructs of students’ performance potential.

The present study is examining multiple perspectives on these particular instances of assessment in order to gauge the acceptance level among the various stake-holders. For students, the placement test represents quite a challenge since particularly the oral component greatly differs from their previous examination experience at high school. The post-test, on the other hand, does not entail any immediate consequence and is thus not regarded as a high priority. For teachers/raters, the challenge consists in rating a large number of student samples within a relatively short amount of time and, due to a changing teacher population, after an intense norming session previous to the rating activity. Rater training and rater fatigue are key issues here. For test-developers/administrators, the placement test needs to serve as diagnostic tool with the purpose of keeping individual changes in placement and splits in the rating process to a minimum. Both tests also need to work as reliable measures, ensuring consistency across the years and student groups. From the institutional perspective, placement and post-test have to demonstrate face validity, i.e., student acceptance of the instrument, and be able to measure actual progress of students from the first semester to the last.

This paper will present interview data from students, teachers, administrators and institutional representatives in an effort to describe the two instruments from the various stakeholders’ points of view. At the same time, concrete scores will be included to provide a backdrop against which the subjective statements can be understood. Particular emphasis will be placed on the differences in reactions of students to the placement test and post-test scenarios. A brief history of test development and piloting as well as an outlook for the future will complete the presentation. It is hoped that some discussion about the feasibility and challenges of pre-test/post-test design spanning four years of English teaching at universities in China will develop and an exchange of experiences with other institutional testing practices with a similar design can be established.
Investigating learning in higher education through the lens of assessment

King Siong Lee, Universiti Kebangsaan Malaysia (kingsiong_lee@yahoo.com)

This paper reports on part of my research for my PhD. My study is situated in the particular context of the BA English Language Studies (BAELS) programme in the School of Language Studies and Linguistics, Universiti Kebangsaan Malaysia (UKM) and focuses on the undergraduate experience of assessment. While they are the majority stakeholders, students are also the most disenfranchised. My study aims to address this imbalance by using their perspectives to interrogate the assessment praxis in higher education. Most of the students admitted to the programme may be described, according to their performance on the Malaysian University English Test (MUET) as ‘Modest Users, with a fair command of the language’. While all have gone through the public school system which uses Bahasa Malaysia as the medium of instruction, they are now pursuing an undergraduate programme taught in English, which may be a second or third language to different students. The BAELS cohort also may be described as possessing multicultural literacies, as is typical of the Malaysian populace.

Research in language assessment in Malaysia has hitherto been confined to technical interests of test construction and issues of validity. While of late, a number of academic dissertations have investigated these aspects using test-takers’ perspectives, there seems to have been little interest in investigating assessment as social and socially-constructed practice encompassing issues of power relations. The current view in the literature is that assessment plays a key role in the social restructuring of modern societies (Barnett 1992, Knight 1995, Broadfoot 1996, Filer 2000), that it is a social product of values and traditions of particular cultures and the interests of specific groups within them, and also that assessment praxis reveals “the truth about an educational system” (Rowntree, 1987:1). Using Bourdieu’s concepts and theory of practice (see Bourdieu 1992, Bourdieu & Passeron 1997, Grenfell & James 1998, Grenfell & Kelly 1999) as my broad framework, I use assessment as a lens for examining the Malaysian student experience of learning in the university. My findings will either provide further empirical and theoretical evidence to support the (Western) perspective, or contest it. Either way, my investigation will reveal the values inherent in assessment, specifically the values attached to language competence and academic literacy. The findings will have serious implications for the programme as they relate to issues of quality that are high on the university’s agenda.

In this presentation, I shall discuss my theoretical perspective, research design and the methods that I propose to use. I shall also present the main findings from a pilot study.

My approach can be described as institutional ethnographic, and since the focus is on a specific group of people in a specific context, it can be considered a case study. Data gathering activities include focus group discussions and individual interviews with final-year BAELS students, interviews with their lecturers, observations of lectures and classroom assessments, perusal of artifacts and documents related to the course descriptions and assessment, as well as a study of the administrative and procedural aspects of assessment which are part of the assessment praxis.

Is an English language learner ready to take a test in English? The relationship between English proficiency and performance on content assessments

Youngsoon So, University of California, Los Angeles (ysso@ucla.edu)

The importance of a valid measure of English proficiency of students whose first language is not English cannot be overemphasized in the United States where a growing number of English language learners (ELLs) enter the school system. Moreover, measuring an ELL’s readiness to receive instructions and take content assessments in English has become more important since the legislation of the No Child Left Behind (NCLB) act, under which schools should be accountable for the academic progress of all subgroups of students including ELLs.

In this context, it is worth investigating i) the relative importance of sub-skills of English (i.e., listening, reading, speaking, and writing) in designating and re-designating a student’s ELL status and ii) the influence of a student’s English proficiency on his/her performance on various content area assessments (i.e., language arts, math, and sci-
ence). Particularly, examining the relationship between English proficiency and performance on different content-area assessments is expected to provide an insight into the appropriate accommodations, if needed, to students with different levels of English proficiency.

A multi-group structural equation modeling (SEM) analysis will be conducted on two groups of ELLs: current ELLs and former ELLs, who have been re-designated as fluent in English. In the analysis, a student’s Free or Reduced-priced Lunch (FRL) status, which has been found to influence students’ academic performance/achievement, will be controlled for as a covariate. From this analysis, the factorial invariance of the relationship between English proficiency and performance on content assessments can be tested at different levels of English proficiency. The results of this study are expected to shed some light on more valid interpretation and use of results of content assessments for ELLs.

Learning for assessment for learning: A case study

Ying Yu, University of Hong Kong (yushadow@yahoo.com)

A school-based assessment (SBA) component was recently introduced into the Hong Kong Certificate of Education Examination (HKCEE) in English language, which assesses students’ speaking performance over the secondary 4 and secondary 5 periods. This SBA component is intended to be integrated into the normal teaching and learning, with both teachers and students making good use of the assessment information to improve learning. It is hoped that this formative use of summative assessment will bring positive washback to Hong Kong secondary schools. The SBA started in 2005 and is being implemented gradually in Hong Kong secondary schools so that by 2009, all schools will have to submit their students’ marks for inclusion in the HKCEE English with a weighting of 15%.

As Hong Kong has a long history of high-stakes external examinations, this assessment innovation imposes great challenges to the beliefs and practices of various stakeholders in Hong Kong. This work in progress paper describes the impact of the SBA innovation in one Hong Kong secondary school. The study adopts a case study strategy to examine how SBA influences the attitudes and behaviors of various stakeholders in one Hong Kong secondary school, including the principal, teachers, students, and parents.

To capture a fuller picture of the influence of SBA in the school, the data collection period of the study lasted for 18 months, a cycle of the secondary 4 and 5 teaching. First of all, questionnaires were administered to 17 English teachers, 65 secondary 4 students, and 139 parents in order to obtain a general picture of their understanding and practices in relation to SBA. Based on the findings from the survey, focus group interviews were conducted with these three groups of stakeholders to explore further the extent to which SBA influences what they think and what they do. Interviews with the school administration were also conducted to understand the school policy concerning English teaching in response to the introduction of SBA. Two secondary 4 English teachers were chosen as two sub-cases of the study. Their classroom teaching was observed in four different periods of time during the data collection period, the purpose of which was to examine to what extent the intended washback of SBA occurs in the normal teaching and learning processes.

From the data collected thus far, it has been found that the school administration and most of the English teachers are positive about the idea of the SBA. However, some student factors were reported as inhibiting the implementation processes, such as motivation and language proficiency. The students, on the other hand, did not have a good understanding of the rationale for the SBA innovation and seemed to see SBA as an extra exam paper rather than an effective way of improving their learning. The SBA innovation seemed to be lost on the parents as they showed very little knowledge about SBA, probably due to a lack of communication between the school, the students, and the parents.

The uniqueness of the study lies in the case study strategy adopted to examine the SBA impact within a school community. Such an approach has seldom been attempted in the washback literature. Examining the key stakeholders in a school setting helps us to achieve a better understanding of the complexity of the SBA impact and identify the key links among these stakeholders that can promote the intended positive washback of SBA. Findings of the study will have some practical implications for the further implementation of this assessment innovation.
This paper presents findings from a larger project that is supported by the Research Grants Council, Hong Kong (HKU7483/06H). The presenters gratefully acknowledge the RGC’s support.

**Looking beyond rater performance: Exploring the role of knowledge and beliefs in raters’ decision-making**

Jie Zhang, **Guangdong University of Foreign Studies** (zhangjie617@gmail.com)

Previous studies have found that raters would exhibit some general patterns as well as considerable idiosyncrasies in their performance. Few, if any, efforts have been devoted to exploring the underlying reasons why raters would behave the way they do. It has been well established that beliefs are strong predictors of behaviors and strategies in language teaching and learning process. It is thus hypothesized that different content and structures of knowledge and beliefs raters have with regard to scoring would explain, at least to some extent, raters’ performance in their decision-making. This Work in Progress therefore intends to probe into raters’ inherent knowledge and beliefs about essay scoring in CET4 (College English Test, Band 4, for non-English majors in China). Interviews will be conducted to elicit raters’ understanding and interpretation of the rating rubrics, their internalized knowledge and description of typical essay features at each score level, their perception about the writing ability measured by CET4 writing task as well as the responsibilities raters hold in the rating session and their willingness to conform with the common rating criteria, etc. Rater performance is operationalized in the present study as both the scores awarded and raters’ rating behaviors and strategies in the rating process. Many-Facet Rasch Model would be used to assess the accuracy and consistency of scores raters assign and concurrent think-aloud verbal protocols would be collected to examine raters’ content focus and decision-making behaviors when they rate a common set of essays. By interpreting raters’ performance against their elicited knowledge and beliefs about essay scoring, the present study seeks to explore the ideological basis of the observed difference and similarity in rating and thus to investigate the underlying factors leading to better performance and effective strategies of raters. Experts in the quality control panel of CET4 rating session are also to be consulted to see if beliefs of more proficient raters are more in conformity with those anticipated by the domain experts who are well informed in terms of the requirements and purposes of the test designers. Research findings are expected to shed more light on how and why some raters could outperform others in following the given rating criteria to rate candidates’ essays and what aspect of rater beliefs the rater training session should try to foster or transform so as to bring raters into a more uniform “judgment community” in the context of a particular test.

**Mixed-method designs: Is it still early for language testers?**

Heejeong Jeong, **University of Illinois at Urbana-Champaign** (hjeong5@uiuc.edu)

Recently in the area of language testing, there are an increasing number of researches done using both qualitative and quantitative methods. Such mixed method practices have expanded the breadth and depth of research; however the framework for mixed methods design that reflects the unique characteristics of language testing has not been developed. This study attempts to develop a conceptual framework for using mixed methods designs through a comprehensive review of empirical studies in the field of language testing. The study will first identify the purpose of mixing using Greene, Caracelli, and Graham’s (1989) definition (i.e, triangulation, complementarity, development, initiation, and expansion) and its match to the research design. Preliminary findings show, in many cases the purpose of mixing was not clearly stated and few presented appropriate research designs that aligned with the researcher’s initial intentions of mixing. Some studies did not state they were using a mixed methods design even though they were using both quantitative and qualitative methods. In addition, relatively few studies succeeded in integrating the concepts and methods from both qualitative and quantitative approach. This resulted into presenting two different studies with little connection to each other. The study will end by suggesting a mixed methods framework that will be appropriate to field of language testing.
Tests as productive learning opportunities

David Carless, University of Hong Kong (dcarless@hku.hk)

This presentation reports on work in progress derived from a study funded by the Research Grants Council in Hong Kong, entitled Primary English Assessment for Learning (PEAL). This project has as one of its main themes the development of instructional power from teaching processes which precede and follow tests.

The main aims of the paper are threefold:

• To analyze English teachers’ instructional strategies which bridge formative and summative functions of assessment;
• To examine under what circumstances internal school tests act as a barrier to formative assessment and under what circumstances they can be supportive;
• To review the potential of tests to provide useful information to key stakeholders: teachers, students and parents.

The paper is exploratory in nature and draws on qualitative data collected in Hong Kong primary schools. The research method involves: semi-structured interviews with teachers to identify their views and beliefs about assessment and its relationship with classroom practice; informal interactions with students to gauge their response to tests; classroom observations, mainly of lessons preceding and following tests in order to understand the interface between teaching and testing; and documentary analysis of test papers and student samples of work. Standard qualitative data analysis procedures are utilized, involving data reduction, summarizing, coding and categorizing.

The main interim findings relate to the potential of ‘test paper review’ and ‘consolidation days’ to enhance learning after a test; the risk of testing in the early years de-motivating lower achieving students; and tensions between performance or achievement goals and learning or mastery goals.

The significance of the paper relates to its contribution to assessment theory in terms of the exploration of synergies between formative and summative assessment. The paper also carries implications for the development of a context-sensitive version of formative assessment. In other words, where testing is frequent and pervasive negative impacts may be reduced if appropriate follow-up learning interventions are carried out after a test is taken.

The effects EFL teachers’ qualification examination on pre-service EFL teacher training in China

Luxin Yang and Baocheng Han, Beijing Foreign Studies University (yangluxin@bfsu.edu.cn, hanbaocheng@bfsu.edu.cn)

This paper reports on a survey research study on the examinees’ perceptions of the effects of EFL Teachers’ Qualification Examination (ETQE) on Pre-service EFL Teacher Training in China. ETQE is established in July, 2007 by the National Research Center for Foreign Language Education of Beijing Foreign Studies University based on the center’s projects on qualifications of excellent English teachers and language abilities of Chinese EFL over the past four years. The purpose of the ETQE is to help the university students who aim to become English teachers after graduation better prepare for their English language teaching profession. The first ETQE will be given in December 2007 over four cities of China. As ETQE is quite different from popular English language proficiency examinations in China such as College English Band-four/six Examination and Public English Tests, the study intends to obtain the examinees’ perceptions of the testing contents and methods on their understanding of the qualification of being English language teachers in China. The examinees will be invited to complete a survey questionnaire at the end of the exam within twenty minutes. SPSS will be used for the data analysis. The findings of this study will help us find the washback effect on the examinees’ preparation for being English teachers and detect any aspect of the test that could be improved in order to generate positive washback effects on pre-service English teacher training in China.
The study of EFL Chinese learners’ CET-band 4 test preparation strategy: A diachronic study

Xu Wang, Nanjing University of Information Science and Technology (dachengruoque@163.com)

Before sitting for high-stake English examinations here in China, such as NMET, CET Chinese EFL learners often spend a considerable amount of time on preparation for higher scores, enrolling in crash courses and taking model tests. Practice makes perfect as the English saying goes. The current study would like to find out the following research questions:

1. what differences of their test preparation strategies might be between the high score and the low score group;
2. what role the test preparation strategies play in their test performance;
3. how much the test strategies might contribute to their eventual score;
4. whether the strategies will make up their differences in their English proficiency;
5. what strategies do often the EFL Chinese adopt to approach different types of questions in CET?

In the study 4 months before CET-4, students have taken a model test provided by CET test authority, which is reliable and valid. Then given the test results, all the 210 subjects are divided into high score group and low group. During the four months, most of the students have started to make preparations for the examination. The researcher has started to collect their study journal and had interviews with them.

Validating steps to English proficiency (STEP) for assessing ELL students’ English language development

Eunice Jang, Jim Cummins, and Christian Chun, OISE / University of Toronto (ejang@oise.utoronto.ca, jcummins32@gmail.com, cchun@oise.utoronto.ca)

Canadian K-12 schools are increasingly characterized by linguistic and cultural diversity of English Language Learning (ELL) students born outside Canada as well as second-generation students born in Canada, whose home language is other than English. These students’ academic difficulties and a lack of systematic assessment of their language development are increasing concerns for educators. Steps to English proficiency (STEP) was developed by educators at the Ontario Ministry of Education to help teachers to measure ELL students’ English proficiency on a regular basis and track their language development through the grades. STEP consists of language development continua in three different language skills (writing, oral communication, and responding and reading) and assessment tools. The purpose of the proposed research is to seek validity evidence for teachers’ use of the STEP continua for assessing their ELL students in classroom settings. The following research questions are addressed:

1. Do the descriptors identify distinguishable performance features at six different steps?
2. Can the STEP continua be reliably calibrated against assessment tools for English-speaking students, such as provincial tests and common classroom-based tools such as DRA, CASI, PM Benchmarks?
3. Are the STEP continua equally valid for ELLs from different linguistic and cultural communities?
4. Does each STEP on the continua represent empirically distinguishable progress in English language development and take approximately the same amount of time for a student to complete?
5. Are the STEP ratings of teachers who have been appropriately trained internally consistent and reliable across teachers?
6. Are the descriptors and the samples of performance clearly embedded in the context of the Ontario Curriculum for the different grade clusters?

Approximately 40 ESL teachers including ESL coordinators, and 160 ELL students in Grades 3, 6, and 10 were recruited to participate in the validation study. The participating teachers will receive a two-day training workshop in the fall of 2007. Each student will be assessed by two ESL teachers on three different occasions in 2008. Teacher ratings combined with in-depth interviews and classroom observations will be used to address the main research questions. We will report on the results of a pilot study and seek input on the design of the validation study and data analytic approaches.
Washback of CET-4 on students’ learning: Implications for test validity

Qin Xie, University of Hong Kong (Xieqin-2006@hku.hk)

CET-4 is the most commonly used English test in China; it has been widely asserted to exert negative impact on teaching and learning. However, empirical studies of CET-4 washback remain scanty. In addition, in the language testing literature, no study has been done to relate test validity with test use through washback. This study investigates washback effects of the new CET-4 and their implication on test validity.

Focus group interviews were conducted and these were followed up with individual interviews. Four groups of students were purposed selected to cover the major types of university students. A panel survey design was used in that the interviews were conducted with the same groups of students 6 month and 1 month before these students took the test. Issues examined in the focus group interviews were (a) views about structure, content, question types, response formats, and scoring of CET-4; (b) strategies actually adopted in preparing for CET-4; and (c) changes in strategies before and during test preparation. The interviews were transcribed and analyzed according to Glaser and Strauss grounded approach.

The results of the study indicated that students adopted different strategies in preparing for CET4 and these strategies changed as they approached nearer to the testing time. Furthermore, as the test time approached, students increasingly adopted strategies that improved their test taking and test management skills rather than their reading, writing, or listening skills. In this study, a number of test claims were assessed based on the Toulmin model of warranties and rebuttals. The applicability of this approach was discussed in relation to Messick’s notion of consequential validity of test use.

Besides its implications for practical purposes in the design and development of a CET-4 with more potential to generate positive washback, the interest of this study also rests on its contribution to the language testing theory a washback study that links test validity with test use through a useful employment of the Toulmin model.

Which Englishes for English language tests? A conversation

Liz Hamp-Lyons, University of Bedfordshire (lizhl@hkucc.hku.hk)
Alan Davies, University of Edinburgh (a1adavie@staffmail.ed.ac.uk)

In previous work (Authors 2003; 2006; 2007) we have interrogated issues surrounding the question of which English(es) are appropriate for use in language tests. In this conversation we raise questions about what linguistic contexts (if any) necessitate the deliberate use of ‘Standard English’; whether certain SESD, ESL or EFL contexts or certain dominant monolingual, bilingual or pluri-lingual contexts might indicate the appropriacy of the conscious use of local or regional varieties of English in tests; and the role of international ‘standard’ English when a country or region possesses a recognized or emerging variety of English. We further ask what aspects of test design, implementation, administration, scoring and score reporting should be conducted in ‘Standard’ versus other Englishes.

1. Our empirical research using written performance data has led to indeterminate results, and we seek direction in proceeding with further research, specifically with reference to:
   2. Salient theoretical constructs such as ‘bias’, ‘language distance’ and ‘contrastive rhetoric’
   3. Research design that accounts for language-specific, context-specific and micro variables given the necessarily small scale of analysis possible with performance samples
   4. Research design and methods of analysis that account for rater variables, found to be a strong confounding factor in our previous research.
   5. Selection of judges for speaking and writing who are themselves not biased towards an international Standard English

We have many more questions and unresolved issues, including whether language tests should reform in response to society’s changing language(s) needs and uses. We invite others interested in these questions to join us in discus-
sion and we hope to find some colleagues in a range of countries willing to engage with us in study design and data collection as well as analysis in order to develop a robust data set and appropriate analytic processes.

Session 2

Time: June 28, Saturday, 3:40 - 4:40  Location: Room 118, East 6

Evaluation of the foreign language requirement at Duke University

Carolyn Lee, Clare Tufts and Ingeborg Walther, Duke University (kslee@duke.edu, ctufts@duke.edu, waltheri@duke.edu)

In the year 2000, Duke University instituted a new curriculum which includes a foreign language requirement for undergraduates. The uniqueness of Duke’s language requirement is that it allows no exceptions: No student can place out of the language requirement; and students are able to enter at different levels which imply different exiting points. Some students will need to complete three semesters of a language, others, just two and some would only take one semester in a language program (at the advanced level). We are now in the process of evaluating the effectiveness of the curriculum. In this conference we would like to present the implementation and first results of the evaluation project at Duke.

We propose the use of various mechanisms or methods to assess student learning outcomes for the purpose of understanding the extent to which students are achieving the goals articulated by the foreign language requirement. Thus, one of the main goals of the study is to chart student progress at various stages of their language acquisition in order to assess effectiveness of the program and ultimately improve foreign language programs and curricula.

The language study aims to answer the following questions: 1) What levels of language proficiency are students achieving in speaking, listening, reading, and writing? 2) What cultural knowledge, understanding, and perspectives are students gaining at each level of the curriculum? 3) To what degree do factors such as language program curricula, study abroad, heritage, and previous language learning experiences play a role in students’ learning outcomes? 4) What are student perceptions of their own language acquisition?

The multiple methods we will use to collect data include course embedded assessments, student portfolios, online questionnaires, individual interviews, and focus groups, as well as the Simulated Oral Proficiency Interview (SOPI).

The results of the pilot stage examine data from four different language programs: Chinese, German, French and Spanish. The primary stakeholders of the evaluation project are university administrators who have a particular interest in demonstrating the merit and value of the foreign language requirement. Also, foreign language departments have an interest in more clearly identifying and understanding how individual language programs prepare students to meet the foreign language requirement goals established by the university. The information gathered will be used by individual language program faculty to assess the extent to which their own programs and curricula are preparing students to meet these goals, for the purposes of engendering programmatic and curricular improvement. At this conference we hope to receive feedback and comments from the audience as well as to share our experience with anyone who is in the process of developing an evaluation project of such a dimension.
A test of cognitive academic language proficiency in English and Xhosa, an indigenous South African language: Justification for use in terms of the aims of bilingual education and international standards on test equivalence

Elize Koch, *University of the Western Cape* (elize.koch@nmmu.ac.za)
John Landon, *University of Edinburgh* (john.landon@ed.ac.uk)

In this presentation it will be argued that both the aims of bilingual education, as well as evidence that the international standards for tests that are available in two or more languages are met, need to be used in a justification for the uses of language tests in education in multilingual and multicultural societies.

In multilingual and multicultural societies, especially those where one dominant language, such as English, poses a potential threat to educational inclusiveness and other languages and diversity, language tests can play an important role in the implementation and monitoring of additive bilingual education, or developmental maintenance bilingual programmes (MBE). MBE is contrasted with transitional bilingual education (TBE), in terms of two categories of aims, namely societal/educational, and language outcomes (Baker 2001). An example of a societal aim in terms of TBE is ‘assimilation’ into the ways of talking, acting and thinking that are represented by the dominant language. With developmental MBE, the societal/educational aim is ‘language maintenance, pluralism and enrichment’. This represents a different kind of society and form of education – one in which diversity is acknowledged and valued. Aims in language outcomes for TBE are monolingualism or limited bilingualism; for developmental MBE they are bilingualism and biliteracy. In MBE programmes, language tests thus serve the purpose of:

- Monitoring the readiness of children to start with instruction in the additional language of the school, and
- Monitoring the continued development of the children in both languages of instruction.

One such programme of MBE for Xhosa speaking children in the rural areas of South Africa is being implemented at the moment. As a result of colonial and apartheid practices in education, this context is still typified by TBE for the majority of first language speakers of the African languages, even though the official language policy for education is one of additive bilingualism. The Additive Bi-Lingual Education (ABLE) project is a twofold research project that that investigates the practical implementation of MBE in the classroom, and aims to longitudinally research the outcomes of curriculum delivery in the mother tongue on language development in both isiXhosa and English, on cognitive processing skills and on academic achievement.

A test of cognitive academic language proficiency, the Woodcock Muñoz Language Survey, was selected to research the language outcomes of the project. This test is used widely in the USA for the evaluation of bilingual programmes, and is available in Spanish and English. After receiving permission for the test to be adapted into Xhosa, a team of translators, educators, linguists and a psychometrist, started the process of adapting the test into Xhosa. The challenge is, in the language of the International Testing Commission’s Guidelines for the adaptation of tests, to ensure the linguistic equivalence of the two versions of the tests, and to provide evidence of the statistical equivalence of the two versions of the test. The English version was also adapted for the South African context.

Data was collected on 200 English speaking grade 6 and 7 children (on the English version) and on 200 Xhosa speaking grade 6 and 7 children (on the Xhosa version). This study will report on the processes that were followed during the workshops to translate and adapt the test, a DIF analysis of the items of the two versions of the test, some of the steps that were taken to address problems in the initial adaptation of the test into Xhosa, and the results of an evaluation of the construct equivalence of the two versions. The results of a logistic regression analysis of the items and a weighted multidimensional analysis on the structures of the test will be reported. Further research will be discussed.
Bridging language tests and test-takers: A social cognitive investigation

Ying Zheng, Queen’s University (2YZ13@QUEENSU.CA)

With China joining more actively in the global economy, coupled with the explosion in commercial, technological, and cultural exchanges with Western society, there emerges a significant surge of English learning and English language testing in China. This study intends to examine the influence of test-taker characteristics on China’s nationwide high-stake College English Test (CET) performance. Identifying prominent test-taker characteristics and how they influence the CET test performance can contribute to the evidence of test fairness and construct validity of the test.

Language tests play a crucial role as instruments in political, educational, and social contexts (Shohamy, 2007). In many circumstances, however, there is not sufficient information on the relationship between language test-taker characteristics and test performance or on whether or not language tests provide adequate validity evidence (Kunnan, 1995). Disagreements rest in how the test-taker characteristics/variables are implicated, the relative dominance of the variables, the context in which models of relationships are most relevant, and the nature of the causal sequences (Clement & Gardner, 2001).

This study adopts Bandura’s (1986) social cognitive model in examining two major social psychological factors associated with test taking: motivation and anxiety. In education, it is generally agreed that motivation and performance bear a positive reciprocal relationship (Pintrich & Schunk, 2002), while anxiety and performance bear a negative one (Lim, 2004). Although previous research has examined extensively these two constructs in achievement contexts, it is of great importance to investigate them simultaneously and build a model of relationships in different research settings. In this study, motivation is approached from a macro-level socio-educational perspective (Gardner, 2001), as well as a personal causal attribution perspective from attribution theory (Weiner, 2005). Constructs to be examined in the socio-educational model include integrativeness, motivation, attitude toward the target-language group, and instrumental motivation. Constructs to be examined from attribution theory include locus of causality, stability, external controllability, and internal controllability. Anxiety is analyzed from general language anxiety theory (MacIntyre, 1995) and specific test anxiety theory (Horwitz, Horwitz, & Cope, 1986) to account for the anxiety effects, such as communication apprehension and fear of negative evaluation, arising from language classrooms and testing environments.

This study will be conducted employing a mixed-method research framework of both quantitative and qualitative approaches. At this stage, all the theoretical reviews have been carried out. Questionnaires about demographic information, motivation constructs, and anxiety constructs will be issued to over 500 CET test-takers from different regions around the summer CET administration of 2008. The data will then be factor analyzed to inform the major categories of test-taker characteristics. Structural Equation Modeling (SEM) will be employed to build a measurement model as well as a structural model to further understand the causal relationships between the investigated factors and test scores. Retrospective interviews will be followed with twelve test-takers to elicit their in-depth accounts of the test from a qualitative perspective (evenly split by gender, levels of proficiency, and majors in universities). Retrospective interviews can provide insights into improving the reliability and validity of the CET by obtaining test-takers’ accounts of their cognitive processing in writing the test. It is hoped that this work in progress will help better understand the relationships between test-taker characteristics and language test performance in broader contexts.

Can teachers become testers? And testers, teachers?

Rama Mathew, Delhi University, Delhi (mathewrama@gmail.com)

Indian school leaving exams in English are paper-pencil tests and focus largely on knowledge and understanding of prescribed texts. Therefore scores on these tests do not reflect one’s language ability and do not indicate how well students can read, write, speak or listen in different situations for different purposes. Ironically all educationists and parents are seriously concerned with this issue but nothing is done to circumvent the problem. The situation is grave as students after 6-8 years of English education do not show adequate competence in the language; the
problem is going to assume greater proportions in view of the fact that English has been introduced from Class I in mother tongue medium schools in a majority of the States.

A project in test development has been initiated in this context therefore, which aims to develop proficiency tests at different levels, for example, at class V, VIII, X and XII levels. The specific objectives are the following:

• Articulate proficiency levels relevant at different levels in the multilingual Indian context.
• Construct tasks in different skill areas based on these proficiency levels.
• Trial them in different school contexts and modify them accordingly.

In the first phase, at a workshop that was conducted for teachers and administrators we addressed questions such as (i) What is language proficiency and how is it different from scores on achievement tests that are based on prescribed texts? (ii) How many levels of proficiency can we conceive of that would be beneficial for language planning, teaching, testing purposes?

The next phases of the project will involve arriving at can-do statements at different levels and constructing tasks to match the levels, and piloting them. This would be an iterative process in which testing ‘experts’ and teachers will together work on fine tuning tasks, scoring criteria, marking and redesigning tasks. This stakeholder approach to test development, it is hoped, will not only help them to claim ownership of the ‘new’ proficiency test, but also enable teachers to be trained in the different processes of test development. It would also help to bridge the gap that exists between teaching and testing, which is crucial to the enterprise of language education.

**Development and evaluation of integrated language tasks**

Robb McCollum, *Brigham Young University* (robb@byu.edu)

This presentation will describe the process used to develop and evaluate integrated reading-listening-speaking and reading-listening-writing tasks for an English as a second language final exam at a multi-level intensive English language program in the USA.

Although integrated tasks, such as those in the new version of the the Test of English as a Foreign Language (iBT), are gaining exposure in the language testing community, this work in progress makes a valuable contribution in that it investigates how integrated tasks function in relation to novice and low intermediate level English language learners whose proficiency falls below the TOEFL target group.

First, the presentation will provide a rationale for the adoption of new integrated skills tasks. Then the process of developing and refining the tasks will be discussed. I will explain the investigation and analyses that are intended to evaluate the implementation of the tasks including teacher and student verifications, a linguistics features analysis, and scoring-related validity evidence.

The presentation will also detail preliminary data about washback from the integrated tasks in order to investigate consequential validity. Suggestions will be both offered and solicited for others looking to implement integrated tasks into their programs.

**Development of an automated oral reading fluency assessment**

Ryan Downey, Alistair Van Moere, Masanori Suzuki and David Rubin, *Ordinate Corporation* (rdowney@ordinate.com, avanmoere@ordinate.com, msuzuki@ordinate.com, drubin@ordinate.com)

This presentation describes steps in the development of an automated speech processing system for the assessment of oral reading fluency (ORF) in children aged 5 to 13 years. The resulting platform represents a novel application of speech processing technology whereby passages may be read into the telephone and scored automatically and rapidly.
Oral reading fluency, defined as “the ability to read a text quickly, accurately, and with proper expression” (National Reading Panel, 2000; p. 3.5), reflects the ability of fluent readers to decode the written word quickly and accurately and to recognize whole words rapidly and effortlessly (Rasinski & Hoffman, 2003). Three measurable aspects of oral reading ability that contribute to ORF are therefore reading rate (words correct per minute, or WCPM), reading accuracy (words correct/words attempted), and expressiveness (appropriate pausing, intonation, and phrasing). Traditional ORF assessment requires individualized testing using “pencil and paper” in which the tester follows along with a student reading a passage. In order to assess students’ reading progress regularly, the development of an efficient formative assessment tool is becoming increasingly relevant in the current educational climate where accountability is mandated. Because advanced speech processing technology takes advantage of fine-grained temporal parameters of the speech signal that can be precisely analyzed, it presents an ideal means for automating scoring of oral reading fluency.

For the present assessment, the automated scoring models were originally created for use in measuring oral reading fluency in adults for the 2003 National Assessment of Adult Literacy (NAAL). Over 181,000 individual oral reading samples were digitally recorded from adults reading into the telephone. Human transcriptions made from the recordings were used to develop the language models for the speech recognizer. Expert human raters listened to the passages and scored them manually for accuracy, number of words read correctly, and expressiveness; these data were then used to train and validate the automated scoring models. Machine scores were aligned with human scores in a series of training studies to maximize the correlation between scores and minimize outliers.

The automated scoring models were further refined during a second phase of data collection using 350 children. Over 2,200 recorded passages were collected and analyzed in a series of validation experiments. These recordings, and human ratings of the recordings, were used to train the automated scoring mechanism to accommodate for attributes specific to children’s voices. Finally, a training set of 136 passages from five- and six-year-olds was used for further refinement of the models to accommodate characteristics of young children’s speech. Inclusion of this final training set brought the correlation between human and computer scoring on WCPM to 0.97.

Challenges facing the system include difficulty in processing particularly “young” (cf. Higgins & Raskind, 2000), as well as accurately scoring passages read in particularly noisy environments. Another particular challenge underlying the scoring of passages read into the telephone relates to the differing quality of the telephone systems used by the readers, particularly loss of signal and line “noise”. Overcoming these challenges is currently the focus of a rigorous line of research.

Despite these challenges, the results of two large-scale studies (n ≈ 400) indicate that the automatic scoring of ORF is accurate and reliable (> .90 correlation between human and machine scores). These results indicate that automatic ORF scoring can reduce the time teachers spend administering and scoring oral reading test forms. The system also provides meaningful feedback to students and teachers concerning reading rate, accuracy, and phrasing/expressiveness of reading. Audio samples collected from students reading over the telephone are analyzed, scored, and available almost immediately for feedback on a website, providing the feature that students may hear their own recent reading performance.

Do changes of curriculums justify evolution of CET-Band 4? A retrospective textual analysis of CET-4

Xu Wang, Nanjing University of Information Science and Technology (dachengruoque@163.com)

The undergoing study is based on all the part of reading comprehension of CET-4 available. By taking advantage of corpus tools and analysis of the four versions of the English Language teaching curriculum in Chinese tertiary education, the researcher would like to find out the following research questions:
1. In the past 18 years’ does the reading comprehension really justify the reading comprehension requirement targeted by the four curriculums?
2. Do the changes of requirements of reading comprehension in curriculum justify and guide the changes in the reading comprehension of CET-band 4?
3. Does the recent change of ELT curriculum work in tandem with the changes of the part of reading comprehension?

The experts’ comments and suggestions are welcome on the research project.

**EFL teachers’ perceptions of a national English test**

Lisa Amdur, *Tel Aviv University* (lisa-a@013.net.il)

A research study is presently underway examining the interrelationship between policymakers, school administrators and teachers within the context of state-mandated national tests in Israel. The study focuses on the relationship between the intentions of policymakers regarding the tests and the perceptions of principals and teachers with regards the test implementation.

One component of the study focused on EFL teachers’ perceptions of the national English test administered to students in grades five and eight. Previous research has shown that teachers respond to external testing by narrowing the curriculum (Shepard, 2000, Smith & Rottenberg, 1991), emphasizing test-like instructional methods, teaching test items (Brown, 1993, Smith, 1991, Shohamy, Donitsa-Schmidt, & Ferman, 1996) and increasing time devoted to test preparation (Smith, 1991). These factors guided the construction of a questionnaire for teachers. It assessed two dimensions: a. EFL teachers’ perceptions and thoughts about the test, and b. how teachers respond to the demands/pressures of the test. The questionnaire consisted of 62 items related to the demographic background of teachers (years of experience, age, and role in school, etc.) and their perceptions of the test and the ways they have adjusted their classroom practices.

The questionnaire was administered to 169 EFL teachers who participated in the national test over a period of two years (2004-2006). Preliminary results indicate that teachers believe the results of the test are accurate and reliable and that they change their classroom practice and course content in response to the test. Specific details will be described and elaborated in the paper.

**Employers’ use of CET: From the test takers’ perspective**

Dayong Huang and Mark Garner, *The University of Aberdeen* (hdy124@yahoo.com.cn, m.garner@abdn.ac.uk)

The College English Test (CET) is a large-scale national test intended by Ministry of Education of China to fulfill the two purposes: 1) to measure the English language proficiency of the university students and 2) to promote EFL teaching and learning in universities (Jin, 2007). However, the test results have increasingly been used for other additional purposes. One such unintended use is as a job selection criterion by employers. During the course of interviewing CET stakeholders for my doctoral research into the impact of CET, employers’ use of CET test results emerged as a significant issue. The preliminary analysis of a number of the interviews suggests that employers are actually exerting a greater influence than the other stakeholders of CET. Literature review shows that while there is certain amount of research on the washback of CET (Gu, 2004; Huang, 2002; Jin, 2000; Zhou, 2002), very little has focused on the employers’ use of CET and its impact. Based on the interviews of university students, this paper will report the preliminary findings of a part of a large scale study of the social and educational impact of the CET. This paper will examine the employers’ use of CET from the test takers’ perspective. It will explore: 1) how the CET result is used by the employers in the selection; 2) how do the test takers see this use; and 3) the impact of this use on the test takers’ English learning and other aspects.
Evaluating aviation policy in relation to English language proficiency testing: The case of Korea

Hyejeong Kim, The University of Melbourne (h.kim7@pgrad.unimelb.edu.au)

In recognition of the importance of plain English for aviation safety the International Civil Aviation Organization (ICAO) established the aviation English language proficiency testing policy in 2004. Therefore, as of 5th March 2008, pilots or air traffic controllers who do not show the minimum required level of English proficiency will not be able to practise their profession. There has been strong resistance to this policy both from incumbent Korean airline pilots and also from air traffic controllers within Korean aeronautical society, most notably from those who have years of experience and whose foremost concern should be safety. This strong resistance has raised the question of whether there are factors that the ICAO has ignored or not considered fully in their decision-making process.

The study described in this work in progress session will begin with a needs analysis in order to understand the nature of the communication required for radiotelephony communication in order to evaluate the new English proficiency testing policy enacted by the ICAO. This analysis is expected to draw on insights from professional groups (including test takers). To better understand the resistance to the policy and associated test used for its implementation the study also investigates the policy’s perceived impact on different stakeholders including related government agencies, airline companies, instructors (and teachers) and most importantly the test takers, pilots and air traffic controllers in Korea. Since the policy has the potential to impact on the lives of thousand of Korean aviation personnel, a careful and principled evaluation of the policy and its practical and ethical implications is critical. The study has the potential to raise the awareness of policy makers both in Korea and other aviation contexts and may ultimately bring about refinements of the current policy. The study also models a new multifaceted approach to language policy evaluation thereby making an original contribution to applied linguistics.

Integrating quantitative and qualitative approaches in L2 writing assessment

Ling He, The University of British Columbia (lingheli@interchange.ubc.ca)

The purpose of this study is to explore the interactive continuum of integrating the quantitative and qualitative research methodology in assessing L2 writing performance. To best understand the overall writing quality with the two prompts, one in general and the other required the specific topic and culture knowledge, this study will conduct two-phase analysis on L2 writers’ performance: 1) text analysis based on discourse theory to trace down the specific text features such as topic development, discourse coherence and cohesion, language accuracy and fluency on two prompts and 2) multiple regression analysis to determine the degree to which perceptions of overall writing quality could be predicted from the score on each of the subareas of text features of investigation. By mixing quantitative and qualitative methods, the study is to obtain a more complete picture of the latent variables that might impact the L2 test takers’ writing performance. The implications generated from this investigation will inform both language testing, teaching, and learning.

Interactive assessment in an ESL context: Student learning, teacher constraints and the need for capacity-building

Nicole Tavares, University of Hong Kong (tavaresn@hkucc.hku.hk)

In this work-in-progress session, I wish to discuss with conference participants a longitudinal study that I plan to conduct within the context of school-based assessment (SBA), one of the most recent education reforms in Hong Kong, with ‘Interactive Assessment’ (IA) which is a prominent and integral feature of the local SBA as the prime focus. Grounded in Vygotsky’s Zone of Proximal Development (1978) that knowledge is constructed through dialogue and developed based on the theories of educative assessment (Stiggins 1998), assessment as feedback (Wiggins 2004) and dynamic assessment (Lantolf and Poehner 2004; Poehner and Lantolf 2005), IA calls for re-conceptualisation of the teacher-assessor’s role and strategy use in the process of assessing learners’ oral language.
IA requires that teacher-assessors engage in a collaborative relationship with learners and intervene during their spoken performance to provide timely and appropriate ‘mediated guidance’ to (i) scaffold and support learners in producing an assessable performance that they know learners are capable of producing at their best and (ii) stimulate and challenge them to reach a higher-level performance than that being produced. In line with Valsiner’s future-in-the-making model (1998), IA is believed to have the potential impact of not only promoting improvement-during-the-assessment but also facilitating learners’ development of skills that can be taken beyond the immediate assessment task.

IA is nonetheless counter to prevailing conceptions in most educational cultures of how assessment ‘should’ be done and thus presents ideological, linguistic and pedagogical challenges. Findings of a pilot one-year action-research study of seven Hong Kong junior secondary schools seeking to help ESL teachers improve their IA skills revealed that, while students showed enhanced performance in the organisational aspects of their oral presentations and discussions, teachers failed to focus on the accuracy of learners’ language production, recognise opportunities for positive intervention and use strategies other than questioning.

Inspired by Lantolf and Poehner’s projects on dynamic assessment (2004; 2005), this intended study adopts a Vygotskian sociocultural approach to the analysis of teacher-student while-assessment discourse and teachers’ mediational strategy use. Drawing on pre- and post-assessment interview data, teacher reflections, and the observation and analysis of actual IA episodes in the classroom of an average ESL secondary school, this research study aims to address the following questions: How do teachers interpret IA and what hinders their understanding? What are the factors that limit teachers’ ability to assess interactively? What are the reasons for teachers’ lack of focus on learners’ language accuracy? What mediational strategies are most effective in promoting student learning? What strategies do teachers lack? To what extent does school-based support facilitate IA? What implications does this all have for the local pre-service teacher training and in-service professional development courses? Findings of the study are expected to shed light for the teacher education programmes in Hong Kong at both the school-based and territory-wide levels and contribute towards the improvement of SBA developed by the Hong Kong Examinations and Assessment Authority.

Through discussing the changes in the Hong Kong assessment reform, rationale for this IA study and its theoretical underpinnings, research questions to be addressed, methodology to be used and the value of the study, I hope to invite feedback on how best to proceed with this research and gain (re)new(ed) insights into alternatives or aspects that may have been undermined.

Involving ESP stakeholders in rating scale validation: The case of the ICAO rating scale

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

The International Civil Aviation Organization (ICAO) has stipulated that by March 2008 all pilots and air traffic controllers working internationally need to have an endorsement of their communicative ability in an aviation environment. ICAO has adopted a proficiency rating scale and produced an accompanying manual. The ICAO rating scale has been criticized for, for example, including references to ‘non-verbal clues’ and possibly not being representative of the construct in question.

This work-in-progress session reports on a study which attempts to validate these criteria using a panel of experienced pilots. Ten pilots were asked to listen to ten recordings of candidates (from a range of proficiency levels) taking a large-scale aviation-specific language test and state whether these candidates’ proficiency was high enough to conduct radiotelephony communications in English. The aviation experts were then asked to justify their decisions. After listening to the performances, the aviation experts met for a number of focus group interviews to discuss their findings. In the final focus group interview, the suitability and comprehensiveness of the ICAO rating scale was discussed in relation to the candidate performances.

Some preliminary findings of the study are discussed as well as some of the shortcomings of the research design. In light of this, it is hoped that suggestions from participants will help to shape the design of a larger-scale follow-up study.
Justifying the use of classroom achievement assessment of reading for all stakeholders

Kangzhou Peng, China West Normal University (pengkzhou@yahoo.com.cn)

Using Bachman’s (2007) validation framework: Assessment Use Argument, this on-going work is to articulate a specific assessment use argument to investigate validity issues related to the use of the end-of-course classroom achievement assessment of reading in the fourth semester currently practiced in college reading courses for English majors in China West Normal University to make master/non-master decision about students. The purpose of this study is to determine the extent to which scores on the classroom achievement assessment of reading can be interpreted as indicators of reading progress as defined by the English Teaching Syllabus for English majors (Level Four).

Based on the course objectives of reading (Level Four), three parallel tests are designed to serve as a diagnostic test, a progress test (mid-term test), and a final test to check the learning progress. Latent growth modeling will be also used to measure such progress. Quantitative methods are used to examine the end-of-course classroom achievement assessment in relation to the subtest of reading in TEM-4, another measure of the same ability.

Verbal protocol will be employed to shed some light on the quantitative findings. CFA analysis of multitrait-multimethod data will be used to investigate the convergence and discrimination. Especially, questionnaires and interviews will be used to investigate the consequences of such classroom achievement assessment on the students. Special attention will be paid to the dependability index of cut-score and test-developers’ professional training.

The study intends to prove the usefulness of classroom assessment for accountability and promoting student learning based on Bachman’s framework of Assessment Use Argument.

Justifying the uses of a high-stakes EFL test: An application of the assessment use argument

Huan Wang, University of California, Los Angeles (wanghuan@ucla.edu)
Ru Xing, Tsinghua University (xingru9912@sina.com)

Test scores are often used to make decisions about test takers. In addition, those decisions are intended to have beneficial consequences, such as promoting teaching and learning. To justify the score uses, we may want to provide evidence to support four types of warrants: relevance, utility, intended consequences, and sufficiency (Bachman, 2005). We may also want to collect evidence to address concerns raised by the stakeholders against those warrants. Although the Assessment Use Argument (AUA) proposed by Bachman (2005) is a promising framework to integrate research on construct validity, test fairness, and backwash, its usefulness has not been fully understood by test researchers and users, and the framework has been seldom tested by real-world applications. The present study therefore sets out to apply the AUA to justifying the uses of a high-stakes EFL test in China. To achieve the purpose, the study takes three steps. Step 1: Reviewing the AUA and its applications. Step 2: Applying the AUA. Step 3: Evaluating the AUA. In Step 2, the existing uses of the test scores are first identified, and the AUA is then articulated to see what evidence needs to be collected to support the uses and to address the concerns against the uses. After that, the relevant evidence is collected via both qualitative (interviews and surveys) and quantitative (e.g., structural equation modeling) methods. The evidence is then used to evaluate each type of the existing test uses. The study sets up an example of applying the AUA to justifying existing test uses. This experience with justifying a test from the real world may also provide valuable feedback to the AUA framework. Moreover, the study offers insight into how to integrate research on construct validity, test fairness, and backwash, and how the three areas are related to one another.
Washback, characterized as ‘the influence of testing on teaching and learning’ (Shohamy, 1993), has long been mentioned in the language testing literature, but the phenomenon has only started to attract attention from researchers in recent years. Ever since Alderson and Wall’s (1993) Sri Lankan Impact Study, more and more empirical studies on different tests and educational contexts have been published to enrich the scope of the washback research, to testify the existing washback hypotheses, and furthermore, to reach a more comprehensive washback model.

This present study will investigate the washback effects of the English graduation benchmark policy on college teachers and students in Taiwan. With its exam-oriented tradition, the Ministry of Education (MOE) of Taiwan has been campaigning for the use of the locally developed GEPT (General English Proficiency Test) in an attempt to boost people’s English proficiency across the country. One of such policies was the English graduation benchmark policy for college undergraduates, which encourages each university and college to set its own graduation benchmark for English using the GEPT as an exit test.

Unfortunately, there has been a striking discrepancy between the MOE’s expectation and students’ actual performance on the test over the years, with the latter being much worse. As a result, almost every university and college which has set its English graduation benchmark has had to provide other options for those students who can not meet the requirement before graduation. The most common way was for students to take extra courses instead. Some schools have offered an internal test for students in addition to the external test. A few others have lowered the passing scores originally set for the GEPT. However, there is no evidence to show that a student passing an extra course or an internal test has the same level of required English proficiency for graduation as those who passed the external test.

Since the consequences of failing the external test are not fatal and can be easily made up by other ‘shortcuts’ to fulfilling the graduation requirement, the stakes of the current benchmark policy are questionable. Alderson and Wall (1993) propose in the Washback Hypothesis that “tests that do not have important consequences will have no washback.” Based on this assumption, the GEPT serving as an exit test and its substitute measures probably had no washback effects on college teachers and students. However, empirical evidence is still needed to reach any conclusions.

This present study will also investigate college students’ English language needs and how these needs have been matched to the graduation requirement as well as the EFL higher education in Taiwan. Saif (1999, 2006) concludes that positive washback can be achieved by designing the test in question in accordance with the stakeholders’ language needs. Conversely, it is possible that a test with no linkage to the stakeholders’ language needs will bring about no or even negative washback. In terms of the English graduation benchmark policy in Taiwan, the administrative force was entirely relying on one test to bring about changes in the EFL higher education, but exactly how the GEPT, serving as an exit test, reflects colleges students’ language needs has never been a major issue, not to mention its substitute measures. It is thus worth thinking whether there is a connection between students’ poor performance on the test and the ignorance of students’ needs in the English graduation benchmark policy. The goal of the EFL higher education in Taiwan will also be investigated at the same time in order to capture a broader view.

This present study will employ a mixed methods approach on two technological universities in northern Taiwan, which have similar graduation benchmarks for English. On the ground that washback is a complex and elusive phenomenon, it is expected that the qualitative data will add meaning to the quantitative data and the quantitative data will add precision to the qualitative data (Johnson & Onwuegbuzie, 2004).
Supporting teacher change in assessment practice through professional development

Bordin Chinda, The University of Nottingham (aexbc@nottingham.ac.uk)

In the past decade there has been a great shift in language education toward the use of alternative teacher-based assessment. However, many teachers are not ready to conduct assessments in their classroom mainly because they have not had the support they need. When teacher-assessors receive adequate support, they have better opportunities to participate in their education society, including assessment. One type of support these teachers need is a professional development programme in language testing and assessment.

The main purposes of this study are to implement a professional development programme in language assessment and to examine the changes in teachers’ cognition (their beliefs, attitudes and knowledge) and practices in relation to assessment that result from this programme. This paper will report the results of a 3-month preliminary study of the implementation of the professional development programme and its progress, and of the changes in the teachers’ cognition and practices resulting from the implementation of this programme.

From the pilot study conducted at Chiang Mai University, Thailand, which was based mainly on the qualitative data, I found that teachers who participated in the study reacted positively to the shift toward task-based performance assessment but there were significant differences among the teachers’ views toward the rating criteria and how they were applied. The factors which were responsible for the differences included lack of in-service professional training as well as differences in teachers’ educational background, teaching, assessing, material and test development experiences. These factors, in consequence, affected the teachers’ competence and confidence, and the ways they conducted assessment in their classrooms.

The solution I have adopted, in this context, is to implement an in-service professional development programme in language assessment for teachers. This programme aims to get five teachers to cooperate in a series of seven workshops to validate the existing rating criteria, and to create and validate a new set of rating criteria. The programme, moreover, intends to bring teachers to a level of fundamental understanding of language assessment, judging the students’ actual performances, as well as working in a team or as an assessment community. For the study of the changes in the teachers’ cognitions and practices, a longitudinal qualitative case study method, which includes individual and group interviews and self-report protocols, is employed.

Test interactiveness as assessment of assessments in portfolio-based EFL writing program: The potentiality to incorporate test usability into test usefulness

Hongxia Zhang, Fanglin Zhou and Lihui Chen, Anhui University (zhang007008@gmail.com, zhoufanglin@163.com, zwy630@163.com)

The present study is built on a doctoral dissertation research into the test interactiveness of holistic and analytic approaches in comparison conducted in a college-level EFL learning context in China in 2004. This research not only evidences the effective treatment with analytic scoring but also the practical value of test interactiveness as an important test quality. To further explore the value of test interactiveness as an in-class evaluation tool for monitoring portfolio assessment in our EFL Writing Classroom thus becomes the kernel of the present in-progress study. Peripherally, the current study also attempts to investigate the potentiality to incorporate test usability into test usefulness validation to ensure fair and useful assessment.

Test interactiveness is one of the six qualities (others are reliability, construct validity, authenticity, impact, and practicality) in the test usefulness model proposed Bachman and Palmer (1996). It refers to “the extent and type of involvement of the test taker’s individual characteristics in accomplishing a test task” (Bachman & Palmer, 1996: 25). It comprises four components, test-takers’ language knowledge, topical knowledge, affective schemata, and their strategic competence. The least researched or applied as it is in test validation, test interactiveness is an essential index in assessment quality control, especially in validating performance assessment.
With an increasing interest in and use of alternative portfolio approaches in measuring EFL writing ability in China, comes the inevitable issue as regards how to guarantee accurate and fair measurement in portfolio assessment, which has been seen to be fraught with indecision and indeterminacy, especially in handling a complex collection of works of various types. To reduce the indecision and indeterminacy in portfolio-based writing assessment, test interactiveness may promise a useful method largely because the applying of test interactiveness, while balancing cognitive products and processes, considers simultaneously test-takers’ affective schemata and strategic investment.

For assessments to be fair and useful, test validation may have the necessity of stepping beyond test interactiveness and other traditional test qualities to consider test usability. Usability, a term widely used in information technology and product testing processes, is the measure of a product’s suitability for its users. It is directed at measuring the effectiveness and satisfaction with which specified users can achieve specified goals in particular contexts of use. In language testing, test-takers have long played an obscure role in test development and validation. In actuality, test-takers are one of the real end-users of a test. Do they find the test effective and easy-of-use? Are they satisfied with the test? How do they perceive their experience of being tested? What usability factors matter most? Do they converge with test usefulness qualities or not? What is the bottom-line test usability like? These questions, as supplementary to the kernel issue afore-mentioned, will be examined and answered mainly through a series of questionnaire investigations and verbal reports. This part of the study will help test developers and other test users to hear test-takers’ voice and hence ensure veritably fair and useful assessment.

Test-takers’ perceptions of and performances in a classroom-based oral assessment

Yongfei Wu, Anhui University (alexwuabc@yahoo.com.cn)

Test-takers’ perceptions of the tests can often provide abundant information for the test designers. Shohamy (2001:373) advocated democracy in language assessment and suggested that assessment practices should include different voices in assessment. Yu (2007:539) investigated the students’ voices in an evaluation of these students’ written summaries and discussed the value of involving test-takers in the assessment criteria development. In addition, Huhta et. al. (2006:326) researched the ideas, feelings and experiences of the students who were preparing for and taking a high-stakes national examination.

This research is intended to investigate 112 test-takers’ perceptions of a classroom-based oral assessment and their performances in such oral assessment at a key comprehensive university in China in January, 2008. A questionnaire is being designed for the purpose of collecting the data on the test-takers’ perceptions for quantitative research. In addition, a semi-structured interview will be conducted with these test-takers to deepen the understanding of their perceptions. Their performances in this assessment are going to be observed with scores given according to their oral proficiency.

The result of this research is expected to provide information to the test designers so that the oral assessment is going to be designed and administered in a more democratic way. In addition, it is also expected to provide feedback to the language teachers of these test-takers so that they may make reasonable adjustments to their teaching of oral English in the classroom.

The effect of visual options of MCQ on listening comprehension

Zhi Li, Hunan University of Arts and Science (d9401lizhi@yahoo.com.cn)

From the perspective of task characteristics, this study investigates the effects of visual options of multiple-choice question on listening comprehension in Japanese language proficiency test (JLPT), utilizing retrospective verbal report and follow-up interview methods. The relevant research on listening shows that listening comprehension is a rather complex active inference process; apart from linguistic knowledge and competence, interactions with other world knowledge also play an important role in this process. Japanese Language Proficiency Test is a widely
acknowledged language test for learners of Japanese. The item format of listening comprehension is multiple-choice question, with content visuals or other written forms as its options in the first part, and aural options only in the second part. Previous visual-related studies on listening are very limited, and so far clear conclusions have not been reached, especially for this kind of visual option form.

The subjects in this experiment are 6 seniors of Japanese major from Hunan University of Arts and Science. The testing materials are adopted from the authentic items in Japanese Language Proficiency Test 2000, level 1. The score analysis indicated that the subjects did not perform well on these visual option items, among which the monologue items appeared to be relatively easier; the average score on the visual option items was lower than that of the other items. The analyses of verbal reports indicate both top-down and bottom-up processing were influenced by visual options; some subjects relied on these visuals to verify their listening understanding. On the other hand, different preference toward visual forms was revealed, individual difference in visual reading ability may exist, and some visuals became a burden and distracters for some subjects, thus increased the item difficulty. The implication of this study was discussed and a further quantitative study will be complemented.

Keywords: Listening comprehension test, visual options, retrospective verbal report

The impact of the language of examinations on students’ performance in national examinations: Individual, social and political perspectives

Guoxing Yu and Pauline Rea-Dickins, University of Bristol (Guoxing.Yu@bristol.ac.uk, P.Rea-Dickins@bristol.ac.uk)

A significant number of children in Sub-Sahara Africa demonstrate their subject learning in formal school examinations through the medium of English. The role of language in education as a critical factor for effective learning is crucial, but controversial too particularly in Sub-Sahara Africa, in determining the outcomes of education and language policies and strategies of national governments and international NGOs and donors aimed at addressing poverty reduction and providing quality Education for All (UNESCO 1990) to meet the challenges of globalization. The opportunities to use examinations as a lever for change (Kellaghan & Greaney 1992, 2004) in monitoring and improving education quality in Africa are oftentimes missed, misused or even abused, leading to a “serious waste of scarce educational resources” (Kellaghan & Greaney 2004:13) and discrimination “against minorities, rural populations, girls, and students whose first language different from that of the examination” (ibid: 7). The frequency of speaking English at home is found to be a statistically significant predictor of students’ performance in reading comprehension test across fourteen Sub-Sahara African countries using data of SACMEQ (Yu 2007). The language of examinations has implications not only in relation to the ethics and fairness of the examinations towards the individual students but also on the social and political arenas of the educational systems.

The central aim of the ESRC/DfID funded research project is to investigate the impact of the language of examinations and media of instruction in secondary schools in Zanzibar on the examination performance of learners acquiring subject knowledge and understanding (Science, Mathematics and English) through a language that is not their first. This presentation reports the findings from the systematic literature review on the impact of language on examination performance, addressing two questions: (1) what evidence is there that language constitutes a determining factor in the demonstration of achievement in the formal examination of school subject knowledge? (2) Which country(ies) in Sub-Sahara Africa demonstrates “best practice” (i.e. in relation to validity and reliability in examination development, evidence of positive impact on individuals and society) in the processes of examination development and implementation when/if providing students with linguistic opportunities and accommodation to facilitate students to demonstrate fully their subject learning in examinations. This presentation will also report the social and political impact of language of examination in Sub-Sahara African contexts (e.g. reducing dropout rates and increasing girls’ enrolment and participation).

*SACMEQ: Southern and Eastern Africa Consortium for Monitoring Educational Quality (www.sacmeq.org)
What does the Story Retelling Speaking Test measure? Comparison with other types of speaking tests

Rie Koizumi, Tokiwa University (rie-koizumi@mwa.biglobe.ne.jp)
Akiyo Hirai, University of Tsukuba, Japan (hiraiaki@sakura.cc.tsukuba.ac.jp)

One of the difficulties of speaking tests is ensuring that they have sufficient practicality in administration and scoring. In this context, a tape-mediated Story Retelling Speaking Test (SRST) was developed as a highly practical tool to assess speaking ability of especially producing extended discourse without an interaction. In the test, students read a story silently, and, then, looking only at keywords, they are asked to retell the story and add their opinions about it. Their utterances are recorded. Besides taking the practicality issue into consideration, the SRST also aims to provide positive washback on teaching and learning by having students read a text with the future output opportunities in mind so that chances to intake the language of the story while reading would increase.

As part of the SRST validation process, we have so far examined effects of story lengths on the production amount, students’ reactions toward the test quality, and functions of empirically derived rating scales. Therefore, in this presentation, we will briefly describe the test procedure and report the initial results. Next, we will elaborate future plans to investigate relationships between the SRST and other types of speaking tests, such as the Versant, the SST, (Standard Speaking Test, a modified version of the Oral Proficiency Interview), the T-SST (a telephone format of the SST), and a picture-based speaking test, which all seem to share some aspects of speaking ability with the SRST. This exploration will enable us to grasp the strengths and weaknesses of each speaking test as well as the construct of the SRST and to select appropriate speaking tests according to test purposes.
A construct validation research of the writing task in NMET (GD) — Text-based English writing for Chinese EFL learners

Xinling Zhang, School of Foreign Languages, Shanghai University; Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies (xlzhang_632@163.com)

The present study, within Messick’s unitary validity conception, collects theoretical and empirical evidence for the substantive and generalizability aspects of construct validity of the text-based writing task in National Matriculation English Test (Guangdong Version), a newly-designed large-scale high-stakes test in China. It adopted a constructivist reading-to-write model specifying the metacognitive (planning and monitoring) and cognitive (selecting, organizing, and integrating) operations elicited in text-based writing. Three general research questions are generated to explore: 1) whether the theoretical processes are actually tagged by the assessment task; 2) whether the two sub-tasks manifest the text-based writing construct differently; and 3) whether performance regularities entail suitability of the text-based writing task for NMET (GD).

Data were drawn from different sources via instruments constructed for this study. In response to the first two general research questions, questionnaire data from experts (N=25), the instructors (N=150), and the target candidates (N=532) were collected. In addition, students’ (n=36) interview data complemented the questionnaire data qualitatively. And the aggregation of the foregoing qualitative data, the coding and the rating results of 189 compositions responded to the third research question.

Qualitative and quantitative analyses of the collected data yielded the following major findings. First, the five metacognitive and cognitive operations were engaged in the text-based writing process by the target candidates. Thus, reliable variance of the test scores can be largely accounted for by the construct being test.

Second, the aggregation of questionnaire data and the textual coding results converged to show that the two complementary sub-tasks manifested the text-based writing construct differently to some extent. Therefore, the multiple sampling via summary and response argumentation substantiated wider generalizability of the inferences about students’ text-based writing ability without redundancy.

Third, various stakeholders thought that the text-based writing task was suitable for NMET (GD). In addition, Multi-faceted Rasch Analysis indicated that the rating scores reliably (Cronbach alpha=.95) discriminated candidates with distinct levels of text-based writing ability. MANOVA of the rating and textual coding results authenticated the conviction.

In conclusion, the blend of evidence for the substantive aspect of construct validity entails match between students’ empirical response regularities and the theoretical text-based writing construct. The assemblage of quantitative and qualitative evidence indicates that inferences based on the test scores can be generalized to and across different
population groups covered. Altogether, the test task is testified a fair measure of students’ targeted ability for the assessment purpose of NMET (GD).

This paper presents findings from a larger project that is supported by the Research Grants Council (RGC), Hong Kong (HKU7483/06H). The presenters gratefully acknowledge the RGC’s support.

A different topic, a different test? Investigating stakeholder concerns about the fairness of a writing test

Barbara Dobson and Gad Lim, University of Michigan (bdobson@umich.edu, limgs@umich.edu)

Shortly after the administration of a large-scale, intermediate-level EFL examination, many test takers and other stakeholders complained that one of the writing topics was culturally insensitive, upsetting to test takers, and inappropriately difficult – in a word, unfair. In response, the test developers offered examinees the opportunity to retake the writing section of the exam, on a different topic, approximately one month later. This poster reports on research investigating whether there was a measurable difference in test taker performance on Topic 1 and on Topic 2 and, if so, what factors might account for any difference observed.

First, the retest population was compared to the group that chose not to retest, and differences between the populations were noted. Then, a principled random sample (n=4,193) of the retake population was selected, and possible explanatory factors were analyzed for main effects and interactions using multi-faceted Rasch methodology. These factors include topic, rater, and examinee gender and ability level. Findings indicate that test takers’ performance differed slightly between administrations, and that (1) the difference cannot be attributed to rater or examinee variables, or to interactions among them, and (2) topic was the only significant, though small, factor. Factors not quantifiable in this analysis were also considered and examined. Taken together, it appears that, in practical terms, the effect of topic on examinee outcome was negligible. Implications of this study for test stakeholders and for future research on writing assessment are presented.

A dual-testlet approach to student placement

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

As an integral part of an EFL curriculum revision project at a Japanese university, an instrument was needed to place incoming students into three level divisions of the newly introduced Core English course. Practical and theoretical considerations have led to an in-house development and administration of a placement test, the process of which the presenter wishes to describe.

Among the notable features of the test development was an attempt at a close interface between the test items and the prospective contents of instruction. Further, the test was designed such that it consisted of two “testlets” of distinct difficulty levels corresponding to the higher two of the three level divisions. Student performance on each testlet constituted the primary source of placement decision, i.e., sufficient performance on the Easier Testlet was judged as evidence of readiness for at least the Middle Level, in which case performance on the More Difficult Testlet was also consulted for decision on the student’s readiness for the High Level.

A Rasch-based calibration of item difficulty was carried out following a pilot testing with 1028 students, and the test revision took account of both such quantitative test data and qualitative feedback data from 629 individuals, who proctored and/or took the pilot test. The completed test demonstrated signs of good functionality overall, when subsequently administered to the target population of 1460 for the actual student placement. The presenter will provide further relevant information on the project while addressing known issues and limitations of the approach.
Adding flexibility to automated scoring of low-level ESL essays

Diane Strong-Krause, Brigham Young University (dsk@byu.edu)

Automated essay scoring (AES) has made major advances in the past 50 years. Benefits of AES include reduced cost in terms of time and money, increased consistency, and automatically-generated feedback. However, evaluating essays of lower-proficiency level ESL speakers has proven a challenge because of the high number of language errors and ill-formed sentences in virtually all the essays. Nevertheless, a number of software programs have been available for large-scale assessments and for use in large institutions, including programs for ESL students (i.e., e-rater). While these programs have been effective in a number of contexts, most require students to write on specific topics and don’t allow much flexibility in terms of what aspects of the writing to analyze. Some language programs and researchers may want more control over topic selection and features to be analyzed.

This poster session reports on an AES study that used Linguistic Technology Inc.’s WordMap grammar checking and linguistic analysis package to estimate holistic writing ability of low proficiency (novice through intermediate-mid) ESL students attending an intensive English program in the United States. Based on tenets of junction grammar, the program calculates a number of statistical measures of linguistic maturity including parts of speech, part-of-speech patterns, vocabulary density, and sentence and essay length. Standard single and multivariate regression techniques were used to derive customized prediction algorithms. Average agreement with human raters was 43% for exact agreement, 68% within .5 band of the human ratings, and 89.5% within 1 band of the human ratings.

Results of this study suggest that this tool may be useful to analyze essays of low-level language learners. It may also be of interest to other institutions who may want more flexibility in computerized essay scoring for assessment and for research purposes.

Assessing the degree of knowledge in vocabulary use

Yuko Morimoto, University of Tsukuba (y_morimoto@hotmail.com)

Vocabulary knowledge has been considered multi-faceted since Nation (2001) classified it into three components: form, meaning, and use. In order to assess the various aspects of vocabulary knowledge, constructing appropriate test-formats according to the construct and test purposes is required. However, each component is continuous in nature, so assessing degree of knowledge is not an easy matter. Attempts to construct a test to measure levels of knowledge in vocabulary started with Wesche and Paribakht’s (1993) Vocabulary Knowledge Scale, and subsequently new formats of measuring degree of knowledge were constructed by Laufer and Goldstein (2004) and Laufer, Elder, Hill, and Congton (2004). Their tests focused on mapping of form-meaning, and some other studies have also been conducted focusing on assessing degree of difficulty in vocabulary meaning (e.g., Jenkins, Matlock, & Slocum, 1989; Marshalek, 1981), but no studies have been done with respect to other components, especially vocabulary use. Therefore, this study suggests the possibility of a test which measures the degree of knowledge in vocabulary use.

In order to measure the aspect of use in vocabulary knowledge, a multiple-choice sentential gap-filling test is used as shown below.

I had started to feel very tired and I decided to try and find somewhere to ( ) a sleep.
a. experience b. lose c. order d. have (correct answer)

Each item has one sentential context and four choices including one correct answer and three distractors. One of the words in the context is replaced with brackets, and test-takers have to select one choice that meaningfully fits into the context. Therefore, the test aims to assess which word can be used in the particular context, and the construct falls into the “constraints on use” in the use component of Nation’s (2001) framework. In order to measure the degree of knowledge, three kinds of distractors were prepared: (a) a paradigmatically-related distractor that had a paradigmatic relationship with the word in the correct alternative; (b) a syntagmatically-related distractor which collocated with the word that formed the minimum semantic unit (cf. Kintsch, 1998) in the context (sleep in the
Assessing young EFL learners in China: An introduction of a new test - Tsinghua Children’s English Proficiency Test

Wenxia Zhang, Meihua Liu, and Xiangjun Deng, Tsinghua University (wxzhang@mail.tsinghua.edu.cn, ellen-lmh@yahoo.com, dengxj98@gmail.com)

With the widespread of EFL teaching and learning in primary schools, and evenly in the kindgartens, across the different cities of China nowadays, how to evaluate young learners’ English proficiency has become an important issue.

This paper first introduces the design, development and implementation of the Tsinghua Children’s English Proficiency Test, a large-scale English proficiency testing system for Chinese EFL children of 6 to16 years old who have completed an English training program for their age and level. It will then describe the overall structure and format of the tests for different age groups and proficiency levels. It will also explain the rationale behind the test construction and test methods. It will further discuss the problems and issues, such as developmental factors, social factors, and cultural factors, that are pertinent to assessing young EFL children.

Up to now, the test has been run to around 1,200 EFL children in about 10 big cities in China. It is hoped that the introduction of this new testing system for Chinese EFL children and discussion of the relevant issues will shed some light on assessing young EFL learners in China, and hopefully will have some implications for assessing the same age group in other EFL/ESL contexts.

Automatic assessment of spoken English skills in aviation professionals

Masanori Suzuki, Alistair Van Moere, and Ryan Downey, Ordinate Corporation (msuzuki@ordinate.com, avanmoere@ordinate.com, rdowney@ordinate.com)

International Civil Aviation Organization (ICAO) have implemented a new language requirement which states that by March 2008 all pilots and air traffic controllers must demonstrate at least Level 4 on the ICAO language proficiency scales (Levels 1 – 6) in all six subskills: Structure, Vocabulary, Fluency, Pronunciation, Comprehension, and Interactions. A cooperative research project was carried out between the Federal Aviation Administration (FAA) and Ordinate Corporation to develop an automated test of spoken aviation English using speech processing technologies that can meet these ICAO requirements.

This spoken aviation English test lasts about 25 minutes. It is delivered over the telephone or via computer, which ensures authenticity that aviation professionals have to cope with in their real-life work, i.e. voice-only situations. The test is composed of eight tasks which are designed to measure facility in spoken aviation English in both technical ICAO phraseology and plain English. The eight tasks are Aviation Reading, Common English Reading, Sentence Repetitions, Short Answer Questions, Readbacks, Corrections and Confirmations, Story Retellings, and Open Questions. Responses from all these tasks except for Open Questions are automatically scored by using...
speech processing technologies. All these test items integrate listening and speaking skills requiring the test-taker to process spoken English skills in real-time, which is undoubtedly critical for aviation professionals particularly in case of emergency.

One of the challenges of the development of this test was to develop a task which assesses the Interactions subskill. A task called ‘Corrections and Confirmations’ was developed. In this task, the test-taker hears a radiotelephony message between a pilot and air traffic controller which might contain incorrect information or a request for some information. The test-taker is expected to respond to the message appropriately by correcting the wrong information or by providing information to the request.

Quality and appropriateness of the test items were controlled by the FAA and Oklahoma State University. The items were further reviewed independently by external ICAO experts for conformity of the items to the ICAO phraseology.

Field-testing was conducted with approximately 1200 aviation professionals (natives = 478, non-natives = 628) to collect spoken responses from pilots and air traffic controllers. These spoken responses were then transcribed by trained transcribers and rated by ICAO experts to optimize speech recognition technology and develop automated scoring models.

In this high-stakes test, test reliability is crucial. Validation studies demonstrated a high test reliability of 0.93 on the final ICAO level. The correlation between the scores obtained by ICAO experts and the scores by the automated scoring system was 0.94 (N=140) suggesting that the machine scoring can provide sufficiently similar information to the judgments of the ICAO experts for this high-stakes testing situation.

Cognitive processing and test validity in the IELTS academic reading test

Tony Green, University of Bedfordshire (tony.green@beds.ac.uk)

Reading for academic purposes typically involves extensive reading of books and journals and synthesising diverse sources of information to build an understanding of an academic subject. Tests of academic reading ability, in contrast, seem predominantly to involve questions requiring understanding of propositions contained within a single sentence, without a requirement either to process extensive stretches of text or to integrate information in order to arrive at the correct answers (Weir et al. 2006).

This paper will report the results of a study of the reading types employed by test takers responding to the widely used IELTS test of academic reading ability. Over 250 test takers responded to tasks on the IELTS academic reading test and completed a previously validated retrospective protocol relating to the range of cognitive processes they employed when performing the test tasks and the extent of information from the text(s) that they drew on in arriving at their answers.

The results of this exercise were compared with responses to an earlier survey of over 1000 students reporting on the reading types they employed in their academic studies. This comparison, employing factor analysis, provides evidence for the degree of congruence between the construct measured by the test and academic reading in the target domain.

Construction and validation of assessor-centered rating scale in Chinese

So-young Jang, Kadeessa Abdul-Kadir, Youngshin Chi, Huei-Lien Hsu, Jiyoung Kim, and Chih-kai Lin, University of Illinois at Urbana-Champaign (sojang@uiuc.edu, kabdulk2@uiuc.edu, ychi1@uiuc.edu, hhsu9@uiuc.edu, jykim@uiuc.edu, cjlin4@uiuc.edu)

Alderson (1991) discussed three different systems in designing rating scales to assess language proficiency: user-oriented, assessor-oriented, and constructor-oriented systems. This study is focused on developing assessor-cen-
tered rating scales where its purpose is not only to justify the rating scale descriptors but to examine untrained raters’ perceptions and their rating process which included three major steps: 1) initial rating scale development; 2) feedback from key stakeholders (rater group); and 3) revision of the rating scale.

The rating instrument was developed for the Chinese Fluency Project at the University of Illinois by the Foreign Language Assessment Group on the basis of the current literature on second language fluency, and further revisions were made based on raters’ perceptions. In this study, teaching assistants who were native Mandarin speakers were involved in workshop to provide feedback on the rating instrument. They were asked to rate twelve video-taped speech samples using the prototype.

Both quantitative and qualitative methods of analyses were used for this study. Descriptive statistics, G-theory and FACETS were used to analyze the analytical as well as holistic scorings of the speech samples. FACETS analysis, particularly, is a major method to analyze raters’ different perceptions on rating scales, and assessment criteria as well as matter of raters’ consistency. Initial findings led to the collapsing of the number of bands in the rating instrument. FACETS also reveals examinees’ ability level for each band. In addition to the above analyses, data from feedback questionnaire were analyzed to explore raters’ opinion on the usefulness of the rating instrument in helping them with the scoring. Feedback on discrepancies in descriptions of fluency in the rating instrument was further investigated and reflected in the revisions of the scale. Implications of this study to the assessor-centered rating scale development and rater training will be discussed.

Describing test taker performance in relation the Common European Framework of Reference: How should it be justified to stakeholders?

Spiros Papageorgiou, University of Michigan (s.papageorgiou@lancaster.ac.uk)

Examination providers frequently refer to the Common European Framework of Reference -CEFR- (Council of Europe, 2001) to describe the performance of test takers. Common standards such as this provide stakeholders with a familiar framework against which they can compare test takers’ performances on different examinations. Consequently, the CEFR has rapidly gained popularity in Europe and across the world. However, researchers have reported problems when using the CEFR in a number of contexts (e.g. Alderson et al., 2006) and have raised concerns about its use and hegemonic power (cf. Fulcher, 2004). It is therefore important to better understand the process of using it for the description of test taker performance.

This paper will report on a study to relate an examination to the CEFR. It will focus on the decision-making process of a group of trained judges who acted as panellists in the linking process. The analysis will shed light on two issues: the factors affecting the panellists’ decision-making when they were assigning particular CEFR levels to test taker performances and the problems that they encountered.

The paper will show that even trained judges draw on factors that are not relevant to the judgement task of assigning CEFR levels to performances. Moreover, the paper will present a number of difficulties in interpreting and applying the CEFR scales. The paper will discuss the implication of these findings for stakeholders who have hitherto assumed the CEFR to be transparent and self-explanatory.

Development of a large scale placement test

Yuji Nakamura, Keio University (nkyj@flet.keio.ac.jp)

Placement tests are one of the most widespread uses of tests within institutions. In Japan, the placement test has been a major issue for the last several years because each institution must deal with the wide range of abilities of the entering students. Hughes (2003) claims that placement tests should be developed by the users so that they meet their needs. Though some schools still choose commercially-produced proficiency tests, a school’s placement test should be connected with its curriculum, and therefore, the tests should ideally be developed by the test users. This poster suggests one way this might be done.
The faculty of letters at Keio University has started a project to create a large scale placement test designed to assess students’ English proficiency and to make decisions about their admittance into appropriate reading-centered English programs. The test is composed of four sub-sections (grammar, vocabulary, cloze and reading comprehension). It has in total 50 multiple choice items and is administered to 800 university students. The purpose of the present study is to examine the validity, reliability and practicality of the test using the Rasch statistical program and to suggest the establishment of an item bank.

The validity can be examined whether the results fit the model or not. The construct validity in the Rasch model is investigated through the examination of five elements: 1) Chisquare examination, 2) Fitsresidual examination, 3) Location examination, 4) Item Characteristic Curves, and 5) Targetting information. Also, the content validity and the face validity are discussed by using the interview survey and the questionnaire analysis. The reliability is investigated by the person separation index, which is equivalent to the cronbach alpha. The benchmark for the acceptable boundary is above 0.7. The practicality of the test is examined mainly by the timing factor for administration and scoring.

Some of the results are as follows: 1) The validity issue was examined in terms of the five aspects (Chisquare examination, Fitsresidual examination, Location examination, Item Characteristic Curves, and Targetting information) in addition to the content and face validity ideas and supported to a certain extent. 2) The reliability was investigated by the person separation index. The index 0.802 cleared the benchmark 0.7. 3) The practicality of the test was examined by the timing factor for administration and scoring. There seems to have been no problems in administering and scoring the test itself.

Once all the items are calibrated and the difficulty of each item is determined each item can be put on the continuum of the scale according to their logit scores (difficulty level). These items along with a task can be stored as items in a bank.

What is needed for the future development is the improvement of the validity by adding the predictive validity plus the establishment of an item bank for a wider use of the test by increasing the test items.

Dimensionality and the construct of language proficiency

John H.A.L. de Jong, Pearson Language Assessments (john.dejong@pearson.com)

The introduction of modern test theory in the field of language testing has raised the awareness among practitioners of unidimensionality as a basic requirement for measurement. The concept, however, has led and continues to lead to debates and misunderstandings.

In this paper we will present data form a large scale research project involving over 6000 subjects with more than 90 different language backgrounds. Subjects responded to English a set of language test tasks purporting to tap all four language skills, in isolation as well as integrated. In addition a number of subskills were identified and scored. Language tasks were divided over 21 different item types. Item types are defined according to the skill or skills intended to be involved, the mode of presentation and the format of the required response. Presentation modes and response requirements were varied over intended skills to allow for trait and method comparisons across and within skills. More than 1400 different tasks were presented to subjects in a linked missing data design.

A first objective of the research was to establish under which conditions a single underlying dimension can be assumed to explain overall score variance.

A second objective was to investigate the possibility to parameterize the impact of method on the response data.

Initial results suggest that items can be calibrated using a unidimensional latent trait model. Further investigations will be directed at answering the second research question and investigating to what extent the assumption of unidimensionality holds for different compositions of the sample of subjects.
Feedback for feedforward: Understanding teacher feedback to student stakeholders in assessment for learning

Geraldine Davies, *The University of Hong Kong* (gdavies@hkucc.hku.hk)

A school-based oral assessment component was introduced into the high stakes Hong Kong Certificate of Education Examination in English Language in 2007. Students are the main stakeholders in the examination although the reputation of teachers and schools also rests on the results of these examinations.

School based assessment is embedded firmly within the teaching and learning process. It is recurrent and formative as well as summative in nature. In assessment for learning literature (e.g. Wiggins, 2004, Black and Wiliam 2003, Hattie, 1999 and Tunstall and Gipps, 1996) there is a strong consensus that providing quality feedback to students following their participation in assessment tasks is the crucial link between assessment and learning. Giving constructive, quality verbal feedback to students about specific aspects of their oral language skills in relation to mutually understood success criteria as well as information on how to improve provides students with opportunities for learning and maximizes their potential to do better in subsequent assessment tasks. The provision of quality feedback which optimizes learner development is therefore at the heart of justifying the use of formative language assessment practices to stakeholders.

Despite its pivotal role in the learning process, the nature of teacher feedback in the classroom-based assessment of oral English as a second or additional language is under-theorized compared with studies of feedback in second language writing assessment (eg. Hyland & Hyland, 2007) and teacher-student interaction in classrooms more generally (eg. Wells, 2006, Gibbons, 2006).

This paper reports on the results of an action research study of 12 Hong Kong secondary schools which sought to help teachers improve their oral English feedback practices as part of an initiative to enhance formative assessment. Drawing on pre and post-feedback interview data, teacher reflection, and the observation and analysis of the actual feedback episodes between teachers and students following assessed interactive tasks, this paper aims to discuss how local Hong Kong English teachers interpret ‘feedback’ and the contextual factors which inhibit or enhance the provision of quality feedback. This paper also aims to present a typology of teacher feedback.

The grounded typology of teacher feedback derived from the analysis provides all stakeholders with a common language through which to discuss feedback strategies. It provides teachers with a framework to use in reflecting on, understanding and improving their own feedback practices. In addition, the explicit understandings about the nature of feedback revealed through this typology can also be used by teachers to enhance their students’ practice of providing reflective peer feedback and will thus be of great relevance to all the stakeholders in formative, school based assessment practices.

Informing ability grouping in EFL listening with a placement test

Hongwen Cai, *Guangdong University of Foreign Studies* (hwcai@mail.gdufs.edu.cn)

This study addresses the use of a placement listening test to inform ability grouping in EFL listening in higher education in the Chinese context, where tremendous variation in listening proficiency is found among English learners. High-achieving freshmen in the placement listening test were singled out as a special class for accelerated instruction, who skipped the regular course of listening skills for the first semester and attended the successive course of news listening regularly designed for the second semester. To validate the placement listening test, it is hypothesized that: 1) Ability grouping based on the test accelerates the progress of high-achieving students and maintains their advantages in listening comprehension; 2) Skipping the course of listening skills causes no harm to high-achieving students; and 3) Singling out high-achieving students causes no harm to students in regular classes.

Two cohorts of freshmen EFL learners were included, 509 from the 2005 cohort (81 male, 428 female) and 361 (56 male, 305 female) from the 2006 cohort. After the placement listening test, 30 students with top performance
were singled out from the 2006 cohort for the special treatment, while 40 students who scored highest among the 2005 cohort were listed as their counterparts, who accepted the same regular instruction for the two semesters with all other students. Two quasi-experimental designs were used in the study: the regression discontinuity design and the pretest-posttest nonequivalent-groups design. The former involved students of the 2006 cohort, with the experimental group of 30 students and the contrast group of 331 students. The latter involved both cohorts. Both the 30 high-achieving students and the 331 regular students were contrasted to their counterparts of the 2005 cohort to test the treatment effects.

The 40-item placement listening test in the forms of spot dictation and multiple choices was administered to both cohorts. A composite score was attained for this multidimensional test by summing up the z-scores of its 5 sub-tests. A dictation task and the final test of news listening were used as the posttests. The 149-word dictation was administered to all subjects at the end of the first semester. The final test of news listening, in the form of spot dictation and free responses, with subtest-level internal consistency α = .79, was administered to the high-achieving students of the 2006 cohort at the end of the first semester and to all students of the 2005 cohort at the end of the second semester.

Major findings of the study are: 1) High-achieving students made significantly greater progress than regular students in the 2006 cohort; 2) No significant difference in progress was found between high-achieving and regular students in the 2005 cohort; 3) High-achieving students in the 2006 cohort made significantly greater progress than their counterparts in the 2005 cohort; 4) No significant difference in progress was found between regular students from different cohorts.

The findings validated the hypotheses that ability grouping enhanced by acceleration maintains and extends the advantage of high-achieving students in EFL listening and that regular students suffer no harms from such arrangement. The placement listening test has thus provided valuable information to guarantee the success of the experiment. In a broader sense, this implies the usefulness of placement tests in informing EFL teaching, to the benefit of both teachers and students.

**Investigating the effects of reading and writing on summaries written by L2 writers**

Yasuhiro Imao, University of California, Los Angeles (yimao@humnet.ucla.edu)

Summary writing has played an important role in academic contexts over the years. Traditionally, summary writing has been used to measure students’ reading comprehension formally and informally in academic contexts (Alderson, 2000). However, this use of summary writing to measure reading comprehension can be problematic because of its complex nature. Writers constantly go back and forth between the processes of comprehending the original text and condensing it into an well-organized summary. This recursiveness of the summary writing processes makes the border between reading and writing ambiguous (Hare, 1992).

Judging the quality of summaries by L2 writers adds one more layer of complexity. In addition to cognitive/meta-cognitive capabilities, L2 language proficiency plays a key role in L2 summary writing just as in any L2 language related activities. Even if L2 writers are capable of successfully completing a summary writing task in their L1, it can be expected that their L2 reading ability and/or their L2 writing ability may affect their summary writing performances.

The majority of the previous studies on summarization have focused on processes of summarization and how well they reflect summary products of L1/L2 writers (i.e. Brown et al., 1983; Johns and Mayes, 1990; Winograd, 1984). The writing aspects of summarization have often been ignored or not been considered as an essential part of summarization studies. In order to fill this gap, this study examines the relationships among reading, writing and summary writing of L2 English writers using a multitrait-multimethod (MTMM) analysis (Campbell & Fiske, 1959).

The participants (N=156) were L2 students at a large research university in the U.S. They completed four tests as part of an ESL placement test: 1) multiple-choice reading comprehension, 2) reading gap-fill, 3) essay writing, and 4) summary writing. The summaries and essays were scored with separate analytic scales, each of which had five
categories. Each summary and essay was rated by two trained raters, who had advanced degrees in the area of English language teaching (α: .45 - .76). An MTMM analysis was conducted with four methods (the tests described above) and two traits (reading and writing) using EQS (Bentler, 2004). Effects of reading and writing abilities on aspects of summary writing performance as well as effects of methods will be reported. Theoretical and practical implications will be discussed.

**Issues in developing a model for verbatim translation exams**

Maria Brau and Rachel Brooks, Language Testing and Assessment Unit, Language Services Section, Federal Bureau of Investigation (maria.brau@ic.fbi.gov, Rachel.Brooks@IC.FBI.GOV)

Since 9/11, the number of languages needing translation to be efficiently and reliably assessed has drastically increased, causing a consentient demand for the development of tests in critical languages as they arise. Consequently, a model by which translation performance tests can be quickly created and are likely to have an acceptable level (phi coefficient > 0.8) of validity is needed.

With this in mind, the Language Testing at the Federal Bureau of Investigation (FBI) has produced the Verbatim Translation Exams (VTEs) 15 languages designed to have high construct validity by using authentic tasks. Problematic issues, such as the familiarity of certain examinees with vocabulary genres, are accounted for. The criterion-referenced framework produced guides native readers in the selection passages appropriate to particular Interagency Language Roundtable Skill Level Descriptions for Translation Performance.

This application of generalizability theory, following and extending work as suggested by Molloy and Shimura (2005), Brown (1999), Brown and Ross (1996), and Bachman, Lynch, and Mason (1995) among others, showed results indicating a complex pattern of significant effects, with clear interactions between the various background factors and the language indicators. Language learning profiles, language use characteristics, personal background factors, family characteristics, as well as linguistic characteristics of the tests used, were included in the analyses.

A validation study of each test was conducted using approximately 80 subjects for each language. Measures, in addition to the translation performance, gathered and analyzed for participants included reading proficiency in the source language, writing proficiency in the target language, background variables, and self-assessments. G-studies and D-studies were conducted for each test to measure significant effects of variables for examinee performance, inter-rater reliability, inter-form score correlation, and the increase in difficulty of passages. The discussion will focus on the assessment of translation and writing proficiency, and the design and implementation of translation tests.

**Justifying test design decisions to stakeholders using a consensus log: A case of an argument-based approach to test development**

Jiyoung Kim, So-young Jang, Kadeessa Abdul-Kadir, Youngshin Chi, Huei-Lien Hsu, and Chih-kai Lin, University of Illinois at Urbana-Champaign (jykim@uiuc.edu, sojang@uiuc.edu, kabdulk2@uiuc.edu, ychi1@uiuc.edu, hhsu9@uiuc.edu, cjlin4@uiuc.edu)

Influenced by Messick’s (1989) unified validity, an argument-based approach has been suggested as a useful framework for validation studies (Kane, 1992, 2001; Bachman, 2005). Moreover, argument-based evidentiary reasoning has been emphasized in test design (Davidson & Lynch, 2002; Mislevy, Steinberg, & Almond, 2003) to justify test properties to stakeholders. Responding to a call for research on collecting validity evidence through qualitative or documentary methodologies (Bachman & Palmer, forthcoming), the current study suggests that documenting consensus-making process of test development is a useful way to articulate and collect a priori validity evidence.

The research is based on an on-line English proficiency test development project between a test provider and a test development team. Our priori validity argument rests on advancing several claims: 1) that the construct of the on-line test be based on a strong theoretical framework for assessing language proficiency for each of the skill tested;
2) that the strength of the assessment lies in the specs-driven model of language test development; and 3) that due diligence be observed during the test-construction phase. Of interest to this study is the creation of a “consensus log” resulting from the extensive documentation of the due diligence process. The project comprises three phases: 1) test specifications, 2) item construction, and 3) validation, and the team has recorded consensus making process during each phase by filling out a consensus log. The log was modeled after Toulmin’s (2003) argument structure, containing a request for changes (claim), justification of the request (warrant), rationales for the changes (backing), and counterclaims for the request (rebuttal).

Initial analyses of the log showed that the most common warrants are concerned with construct representation, cognitive processing, authenticity, and practicality. Furthermore, it was found that backing for the warrants raised by the test developers and the provider complementarily constitutes a priori evidence for context and theory validity, discussed by Weir (2005). The findings of the study suggest that a consensus log is useful in guiding test developers to collect a priori validity evidence as well as in facilitating “shared responsibility” between stakeholders (Shohamy, 2001) thereby ensuring “agreed transparency” of a test.

Language, culture, and the California High School Exit Examination (CAHSEE)

Liberty Van Natten, University of California, Davis (llvan@ucdavis.edu)

The prevalence of high stakes standardized tests is increasing in US public schools. The California High School Exit Examination (CAHSEE) is one such test given by the state of California in compliance with the federal requirements of the No Child Left Behind Act. California requires all students in the state to pass the CAHSEE in order to receive a high school diploma. This has spawned hot debates over issues of fairness and validity, particularly for students speaking languages other than English. In the 2005-06 school year, there were over 56 different languages spoken by 1.5 million students in California public schools (www.cde.ca.gov). While these numbers are staggering, they tell us nothing about the numbers of dialect speakers in California. Little is known about issues of fairness and validity of standardized tests for non-standard English dialect speakers. Yet in 2006, of the 75,376 African American high school students taking the CAHSEE, many of whom speak African American Vernacular English, 50% did not pass the English Language Arts portion of the test (CAHSEE results. www.cde.ca.gov).

Research has examined the effects of high school exit exam requirements on high school graduation rates (Warren et al., 2006). Other research has looked at the existence of different dialects of English in schools and their impact on students’ achievement both in the classroom and on standardized tests (Christian, 1979, 1997; Hoover et al., 1995a, 1995b; Love & Kruger, 2005; Shohamy, 2001; Taylor & Lee, 1995; Wolfram et al., 1999). However, at the time of this study there were no studies which questioned the fairness and validity of the language in the CAHSEE for African American students. Because the CAHSEE is administered in Standard English, such research is critical in order to assess the fairness of the test and justify its continued use as a high school graduation requirement.

This research paper examines ways in which the cultural and linguistic practices of high school students are in conflict with the language and cultural knowledge present in the English Language Arts component of the CAHSEE. The project is a case study of three African American bi-dialectal high school students who were preparing to take the CAHSEE. The researcher tutored the students using a test preparation booklet provided by the California Department of Education. The tutoring sessions were audio recorded, transcribed and later analyzed to identify challenges the students faced. Analysis of the transcripts uncovered differences between a) the dialects and life experiences of the students and b) the language used in the CAHSEE. These differences were most evident when the students struggled with multiple genres of text and vocabulary. Their abilities to correctly answer questions were often hindered due to their experiences outside of their classroom learning environments with the genres and vocabulary contained in the test. Students’ perspectives and opinions about the CAHSEE were also particularly revealing. These findings delineate cultural bias on the test and illustrate challenges for students and educators which have previously been overlooked. This study shows that high stakes standardized tests and the systems that use them need closer inspection by researchers for the students who are required to pass such tests and the educators and schools which are held accountable for the pass and fail rates of their students.
The presentation will begin with a brief overview of the major research influencing this study followed by a description of the methods used in the data collection and analysis. Next, descriptions of the findings will be given with specific examples from the tutoring sessions to illustrate the linguistic and cultural differences which posed the greatest challenges for the students. The presentation will then give a discussion of this study’s implications and conclude with suggestions for educators and future research.

**Rater variation in web-based assessment of KFL learners’ pragmatics using FACETS analysis**

Soo Jung Youn, *University of Hawai‘i* (soojung@hawaii.edu)

With growing interest in L2 pragmatic assessment and web-based language testing, the present study attempts to address an issue of web-based scoring method in performance-based pragmatics assessment. The purpose of the study is to investigate variability in two scoring methods, web-based scoring and paper-based scoring, across various aspects of pragmatic assessment of Korean as a foreign language, including test types, speech acts, and levels of examinee. For this purpose, multi-faceted Rasch measurement is employed using the computer program FACETS (Linacre, 1996), to explicate each scoring method’s systematic sub-patterns of harshness and leniency in ratings across various factors in pragmatic assessment. This study uses three different pragmatics tests, written DCT, Language Lab, and Role-play, adapted from Hudson, Detmer, and Brown’s (1992, 1995) pragmatics prototype tests. Within each of these three test types are three speech acts: refusal, apology, and request. The two different scoring methods are used: (a) a scoring website where all examinees’ responses are uploaded and the raters assess the responses on this website, and (b) paper-based scoring materials with rating sheets. Two groups of raters separately assess examinees’ responses either using a scoring website or using regular paper-based scoring materials. The detailed measurement reports of seven facets (examinee, scoring method, rater, test type, speech act, item, and rating category) and bias analysis indicate that the two scoring methods show reliable consistency in their rating results. However, unique and systematic interactions between the each of two scoring methods and various facets are found. Implications of the research findings will be discussed, in terms of the issue of rater variability in performance-based pragmatic assessment, and applications of the Web in language testing.

**Relationships between lexical access and selection of choices in multiple-choice vocabulary tests**

Chikako Nakagawa, *Graduate School, University of Tsukuba* (meluchan@hotmail.com)

Vocabulary knowledge plays an essential role in second language learning; hence many widely-used tests such as TOEFL assess it as a predictor of language proficiency. Knowing a word’s meaning is not the only aspect of vocabulary knowledge; other aspects include knowing its form, function, and position in context (Nation, 1990), and the issue of fluent lexical access (Koda, 2005). Lexical access is the activation of relevant word knowledge by particular input, and it has an important role in reading and vocabulary tests (e.g., Dekeyser, 2001; Grabe & Stoller, 2002). However, large scale tests do not focus on efficiency of lexical access, nor do they consider multiple aspects of vocabulary knowledge. They simply assess learners’ knowledge of word meaning using a multiple-choice format. Some researchers have developed tests which can assess depth of word knowledge (Paribakht & Wesche, 1996; Read, 1993, 1998), but these tests have unfamiliar formats which may lead to poor performance from examinees (Baba, 2007).

Addressing these shortcomings, the present study was conducted (a) to develop a new test that reflects learners’ depth of word knowledge in multiple-choice tests, and (b) to investigate how the efficiency of lexical access affects examinees’ performance in multiple-choice tests. Prior to conducting the main experiment, link strengths between target words and choices in multiple-choice tests were assessed to provide an index of efficacy of lexical access. Forty-five Japanese EFL learners were asked to judge, as quickly as possible, whether two words were related or not. Their reaction times for two types of pairs (the target and its synonym, and the target and a phonologically related word) indicated that link strength with phonologically related words was stronger than with synonyms when the target words were unknown, while opposite results were obtained when the target words were known.
In the main experimental study, three types of multiple-choice items were prepared: (a) target words in context (e.g., The project was denounced* as a scandalous waste of public money), (b) target words only (e.g., denounce*), and (c) known words as targets (e.g., blame*). Each item had four choices: (1) semantically related (e.g., charge), (2) same part of speech with the target (e.g., enjoy), (3) not related (e.g., floor), and (4) phonologically related (e.g., bounce, frame). Furthermore, each type of item was divided into two groups according to the link strength between the target word and the correct choice, strong and weak. It was predicted that the examinees’ performance would be higher for items which had strong link strengths, and that the examinees would choose phonologically related words more frequently when the meanings of the target words were unknown. The results after testing 278 Japanese EFL learners showed that link strengths did not affect the percentage of correct answers in any type of multiple-choice test. Second, a comparison of incorrect choices selected by the examinees demonstrated that participants selected more phonologically related words than others when the target words were unfamiliar to them. In contrast, the percentage of words selected by the participants did not differ significantly in the case of known target words. These results suggest the possibility that we can develop a multiple-choice test which can assess the depth of vocabulary knowledge by controlling the types of choices.

The influence of partner proficiency in a paired oral assessment task

Larry Davis, University of Hawai‘i, Manoa (davisle@hawaii.edu)

The use of paired speaking assessments has become increasingly popular despite the potential for a partner to influence an examinee’s performance. The practical importance of this issue remains unclear; for example, empirical studies into the effect of interlocutor gender and/or familiarity on oral performance have reached varying conclusions (Berry, 1997; Katona, 1998; O’Laughlin, 2002; O’Sullivan, 2002). Differences in interlocutor proficiency may also introduce error into measurements of speaking, especially in EFL classrooms where the oral proficiency of students can vary widely. Investigations of the effect of interlocutor proficiency on oral language assessment have produced contradictory results (Iwashita, 1998; Nakatsuhara, 2004) and though pairing students for classroom tests is desirable in terms of efficiency and washback, its reliability remains in doubt.

In this study the influence of interlocutor proficiency on speaking performance was examined in a group of 24 first-year students at a university in Hangzhou, China. The students were divided into groups of relatively high and low English proficiency and tested once with a partner of similar proficiency and once with a partner of higher or lower proficiency.

Partner proficiency level had no effect on scores, but lower-level individuals produced more language (words) when working with a higher-level partner. Moreover, there was a positive correlation between word count and score, although this correlation is likely a result of differences between, rather than within individuals (i.e., more proficient individuals tend to say more, regardless of partner). An examination of interaction type and conversational dominance (Dimitrova-Galaczi, 2004) suggested that the highest-scoring individuals tended to engage in collaborative, high dominance interactions unless paired with a very low-proficient partner, in which case they took a dominant role in an asymmetric interaction. The lowest-level individuals tended to take a passive role in an asymmetric interaction, regardless of partner. Despite overall trends in the data, considerable individual variation was seen in the relationship between scores, language quantity, and discourse.

The role of alignment in the content validation of a large-scale, standards-based English language proficiency test: Background, methodology, and applications to other contexts

Carsten Wilmes and Gary H. Cook, University of Wisconsin-Madison (wilmes@wisc.edu, hcook@wisc.edu)

The No Child Behind Act of 2001 (NCLB) requires U.S. states, among other things, to demonstrate that L2 students in kindergarten to grade 12 make continuous progress in their ability to function in an English-speaking school environment, i.e., to attain English academic language proficiency. Even though our understanding of the construct of English academic language proficiency is still emerging (Scarcella, 2003), the high stakes nature of
the federal mandate led to the development of a number of tests that assist states in assessing and demonstrating the linguistic growth in their L2 student populations.

This poster describes the alignment issues faced in the development of one such test: The WIDA Consortium’s ACCESS for ELLs test, a large-scale, standards-based English L2 test that is administered to over 500,000 students annually. Alignment is a content validation process that has now matured to the point that it is applicable not only to the narrow field of NCLB-related testing in the United States, but can also serve as an additional tool for L2 test validation in general. This is especially the case since the Webb/Cook methodology employed by us naturally integrates a variety of stakeholders in the validation process (Cook, 2007). We provide the background and demonstrate the methodology and examples of alignment, including how alignment, standards development, and washback produce an iterative validation loop.

Variability and validity of automated essay scores for ESL writers
Yong-Won Lee, Seoul National University (ylee01@snu.ac.kr)

One central issue investigated in this study is the feasibility and defensibility of using generic (or non-prompt-specific) scoring models with e-rater for independent writing tasks of the internet-based TOEFL (TOEFL iBT). More specifically, the present study is designed to: (a) create several substantively different types of automated essay scoring models for independent writing tasks of TOEFL iBT, (b) examine the impact of model type on the variability of automated scores, and (c) investigate the criterion-related validity of automated scores produced by these scoring models using various measures of ESL proficiency as criteria.

Three different types of scoring models, including prompt-specific, generic, and hybrid models, were built and evaluated for nine evaluation prompts (including seven TOEFL CBT prompts and two TOEFL iBT independent writing tasks). A prompt-specific model for each of these prompts was built based on samples of essays (n=500) written for that particular prompt. Generic and hybrid models were built based on a total of 20,000 scored essays written for 40 previously-administered TOEFL CBT prompts and applied for each of the evaluation prompts.

Data analyzed in this study included two human raters’ essay scores, one automated essay score from a prompt-specific scoring model of e-rater, three automated essay scores from generic scoring models of e-rater, and three automated essay scores from hybrid scoring models for a total of 64,215 essays written for nine evaluation prompts from TOEFL CBT and iBT administrations. Included among the criterion measures were scale scores from multiple-choice and constructed-response sections and human rater scores on other independent writing tasks of the related tests.

Results of analyses showed that (a) similar levels of score reliability could be achieved for human and automated scores, (b) human scores turned out to be somewhat better correlates of ESL proficiencies than the automated scores, but (c) the validity difference was greatly reduced when more relevant writing measures were used as criteria. Implications of these findings for ESL writing assessment will be discussed in the presentation along with avenues for further investigation.

Washback or fairness? Insights from the criterion-irrelevant factors in the assessment of Cambridge YLE Speaking Test
Hongtao Ren and Binbin Dai, The Chinese University of Hong Kong (renhongtao11@hotmail.com, tangbo1972@hotmail.com)

Cambridge Young Learners English (YLE) test has been gaining an increasing popularity around the world over the past decade. However, the specific characteristic of children as language learners and test-takers tends to make young learners’ tests a controversial issue. There is always the tension between a positive impact and test fairness, particularly in the speaking component of the YLE test. On the one hand, to remain committed to giving children a positive experience of testing and to encouraging their future language learning, the YLE test has a requirement for
intended positive washback and impact. Taking this into consideration, it seems natural for the YLE oral examiner to resort to some elements other than the candidate’s performance in rating. On the other hand, for the sake of test fairness, score accuracy and reliability must be ensured.

Driven by a desire to justify the usefulness of YLE speaking test to the test-takers and their teachers, this study attempts to address the elements beyond the particular speaking scoring criteria, and whether and how they may influence oral examiners’ marking. The study was carried out with 22 YLE oral examiners from Chongqing, China. A Delphi survey was used to identify the non-criterion factors and these data were cross-referenced with follow-up score comparisons and to explore whether such factors might function in YLE oral ratings.

The results indicated that all the rater participants seemed susceptible to a variety of criterion-irrelevant considerations in their rating, though the degree of influence potential varied individually in terms of both the factors and the oral examiners themselves. The study also appeared to suggest that the significance of the influence of criterion-irrelevant rating factors on test-taker scores was unclear. The survey additionally revealed certain overlap of criterion-relevant and apparently irrelevant variables within the construct of communicative competence and performance.

Session 2

Time: June 26, Thursday, 3:30 - 4:30 Room: Room 118, East 6

A discourse-analytic investigation of topic-related performance variation on a high-stakes writing test

Barbara Dobson, Aaron Ohlrogge, and Sarah Van Bonn, University of Michigan (bdobson@umich.edu, aohlrog@umich.edu, svanbon@umich.edu)

In response to stakeholder complaints about the cultural appropriateness of a writing topic in a large-scale EFL test battery, test takers were offered the opportunity to retest, one month later, with a new topic. This poster reports on research into possible topic-related performance variation using pairs of compositions written by a sample of the examinees who retested.

An earlier multi-faceted Rasch study using data from these test administrations detected a slight difference in difficulty level between Topic 1 and Topic 2. To better understand the significance of that finding, the researchers employed discourse analysis methodologies to study how test takers’ discourse is similar and different in responding to Topic 1 and to Topic 2, and to investigate whether any features of test takers’ writing are associated with greater or lesser success on either topic, as measured by band score.

The research addresses four main questions:
1. For examinees who received the same score on both topics, are there any measurable differences in the features of the discourse generated in response to each topic?
2. For examinees whose scores on Topic 1 and Topic 2 differed, are there any measurable differences in the features of the discourse?
3. If there are differences, are they related to level of writing ability?
4. What implications do the findings have for test developers, raters, test takers, and writing teachers?

To investigate these questions, a sample of 100 pairs of compositions was selected. Fifty pairs were papers from examinees who received the same band score on both topics. Fifty pairs were papers where the scores on Topic 1 and Topic 2 differ. The methodology used for analysis was an adaptation of that of O’Loughlin and Wigglesworth (2007). Measures were selected to represent certain categories of the scoring rubric used for this writing test, including Content and Development, and Linguistic Range and Control. Examples of measures used include number of words, type/token ratio, and error-free T-units.

The poster will present results of the analyses and potential implications for stakeholders.
A comparability study on the old and the new CET-4

Qin Liu and Zhenhua Wang, University of Shanghai for Science and Technology (liuqin_2006@hotmail.com, usstchris@163.com)

The College English Test (divided into CET-4 and CET-6) is a nation-wide high-stakes English test in China, administered by the National College English Testing Committee of China. Since its first implementation in 1987, CET has gone through 20 years of development. In January 2004, the Ministry of Education of China issued the newly enacted College English Curriculum Requirements and began the pilot work of college English reform. Accordingly, reform on CET has been carried out and the new CET-4 made its first appearance in June, 2006. Though the old CET has attracted great attention among researchers and language teachers and has already established its authority (see for example, Yang and Weir, 1998), the new one has stirred little research interest so far. The writers of the present paper, therefore, have conducted a study to investigate the comparability of the old and new CET-4, aimed at answering the following research questions:

1. To what extent is the new CET-4 comparable to the old one in terms of content, format and test taker performance?
2. Which test better measures Chinese students’ communicative language ability?
3. What are the main problems concerning the implementation of the new CET-4?

The comparability study included an analysis and comparison of the content and formats of the old and new CET-4 tests, a comparison of student performance on the two tests, and an analysis of a survey about implementation issues. The study was conducted with 83 students from a college in Henan Province, China. All of them were freshmen and have never received any training in the old or the new CET-4. The students were asked to take the old and the new tests (taken from the 2006 test battery) with an interval of two days. After that, they were asked to complete a questionnaire survey concerning their comments on these two tests. The subjective items (Writing in the old test; and Writing, Skimming and Scanning, Compound Dictation, and Translation in the new test) were double-rated and all data was processed by SPSS.

Analysis of test content was conducted by means of Bachman and Palmer’s framework of test usefulness (Bachman and Palmer, 1996). To be specific, authenticity, interactiveness and reliability were investigated to study characteristics of Target Language Use task types; construct validity was examined to reveal construct definition; practicality was studied to check available resources; and impact was reviewed to determine test purposes. Analysis of student performance was carried out through quantitative study. Firstly, students’ scores on these two tests were equated by using equipercentile measure. Then, statistical instruments such as reliability estimate (including reliability quotient $\alpha$ and inter-rater reliability), descriptive statistics (including mean, SD, skewness, kurtosis, and distribution of the scores), and comparability study on each sub-section (by use of correlation and paired-sample t test) were adopted. The questionnaire designed to cater for implementation issues consisted of three parts. The first part was on students’ self evaluation about English proficiency. The rating scale on students’ proficiency on listening, speaking, reading, writing and translation, issued by the Higher Education Department of the Ministry of Education, was adopted. The second part was their self rating on each sub-section of the two tests. And the third part was their impression on these two tests concerning such issues as test length, test difficulty and anxiety while taking them.

The results of these analyses of test content, student performance, and implementation issues will be discussed in the paper.

A critical inquiry to academic language constructs across U.S. states

Mark Nigolian, St. Michael’s College (snigolian@aol.com)

This report summarizes the academic language constructs used across the United States (U.S.) to instruct and assess English Language Learners (ELLs). Under Title III of the 2001 No Child Left Behind legislation, states are mandated to create aligned English Language Proficiency (ELP) standards and assessments that address academic
language proficiency of English Language Learners. This legislation has led to a wide variety of interpretations of academic language proficiency. These interpretations have in turn led to an array of approaches in creating ELP standards and assessment systems by states and consortia of states (eg. WIDA, Mountain West).

This report compares the different approaches states and consortia have taken to instruct and assess academic language proficiency. It illustrates the variety of ELP standards and assessment systems across the U.S. from those based on slight adaptations of state English Language Arts standards to those that address academic language in each content area. The report also evaluates the theoretical foundations of these different approaches to academic language proficiency. It contends that many states have vague, inaccurate, or poorly developed constructs of academic language proficiency and that these weak constructs could possibly have a significant impact on the instruction and assessment of ELLs in the U.S. This report does not promote a mono-version of academic language proficiency, but urges educators, policy makers, and researchers to question the theoretical foundations upon which states are basing their ELP standards and assessment systems.

A L2 fluency rating scale construction project - Challenges and suggestions

Huei-Lien Hsu, Kadeessa Kadir, So-young Jang, and Youngshin Chi, University of Illinois at Urbana-Champaign (hhsu9@uiuc.edu, kabdulk2@uiuc.edu, sojang@uiuc.edu, ychi1@uiuc.edu)

Rating scales are mostly seen as finished products and are seldom inspected for their valid assessment of performance. This poster documents the construction process of a new rating scale used in speaking assessment as part of a larger study aimed at defining a construct of L2 fluency for the development of an automated speech recognition tool employing criteria based on lay perceptions of L2 fluency.

The rating scales developed in this project were used by untrained raters whose ratings will be compared with the automated assessment scoring results. The challenge during the first phase of the project was to define the nature of L2 fluency construct and further operationalize the construct to be measured in the form of a rating instrument. Three groups of untrained raters (from linguistic, non-linguistic, and non-academic backgrounds) were recruited to participate in the trial of a standard rating instrument developed through an alignment analysis of scales used in large-scaled oral tests and classroom assessment. The findings from the quantitative and qualitative data called for a need for a simplified scale to further facilitate untrained raters’ judgment of L2 fluency. The next phase of the project included a qualitative study to further explore individual untrained rater’s perceptions on L2 fluency by means of think-aloud protocols. The results not only helped to further define the construct of L2 fluency but also were used to construct an empirically derived, binary-choice, boundary definition scale (Upshur & Turner, 1999), which required raters to simply respond to a set of binary questions leading to a final score.

This poster illustrates the new scale for assessing L2 fluency. It also outlines the development processes including presenting the different versions of the rating instrument that were used during the project period. Finally it suggests a methodology for language testers and classroom teachers in rating scale construction for use in their own contexts.

A quantitative study on subject-verb agreement in English

Chih Yuan Cho, California State University, Los Angeles (chochihyuan@hotmail.com)

Interventionist educators affirm that formal grammar instruction of some sort can be of benefit (Purpura, 2004). So when it comes to the target audience which consists of first language (L1) young students in classroom settings like in this study, the assumption is that it should have been beneficial via grammar teaching by the same token.

This study was applying treatment i.e. test taking strategies and grammar instructions to see whether or not the treatment could actually raise students’ awareness in subject-verb agreement and the treatment of both teaching and guidance were effective on boosting test scores. The identical test i.e. previously applied to homogeneous non-treatment group of 12 third-grade-elementary mainstreamed students was then administered to another group of 21
third-grader subjects after engaging in the intensive treatment for duration of a month. A t test for two independent samples i.e. non-treatment and treatment was implemented to determine the significance of test results. Significant differences were found between non-treatment group and treatment group in the use of the grammar instructional practices. In terms of data analysis, the mean test score for the 12 non-treatment participants is 58.08 (SD = 11.33) while the mean test score for the 21 treatment participants is 69.19 (SD = 11.30). After implementing the t test as the stats instrument, accordingly, the difference between the means is statistically significant at the .05 level (t = 2.71, df = 31, two-tailed test). Through the test results, the fact is that in a classroom setting under underlying action-based rationales, it is likely to improve students’ scores on subject-verb agreement by putting into practice a series of intensive teaching and guidance even within a short period of time.

“Assessment is important because it helps me to learn”: How are attitudes and actions related among L1 and FL learners?

Ari Huhta, Mirja Tarnanen, and Reeta Neittaanmäki, University of Jyväskylä (huhta@campus.jyu.fi, tarnanen@campus.jyu.fi, Reeta.Neittaanmaki@campus.jyu.fi)

The presentation reports on research into goal-setting, assessment, and feedback practices in first and foreign languages in the Finnish comprehensive schools. The study is part of a more extensive project that aims at exploring and interpreting mother tongue and foreign language literacy practices.

In the first phase of the study, an extensive questionnaire was designed to chart literacy practices in and out of school contexts, for both students and their L1 and L2/FL teachers. The setting of goals for learning, assessment and feedback were all covered in the questionnaire. In addition, the students assessed their L1 and FL skills along several dimensions and in school-internal and external activities, and both teachers and students replied to a set of statements about attitudes towards different aspects of language education. The student survey was administered in the Finnish-speaking schools in Finland in the spring term of 2006 and it targeted students in their final year of compulsory education (grade 9 of the comprehensives school) when students are 15 years old. The survey was based on statistically representative samples of students and of L1 and FL/L2 teachers, and the results are thus generalisable to all students and language teachers in grade nine. Over 1700 students participated in the survey.

The presentation reports the results of the analyses of how students’ attitudes to the language and its use (mother tongue vs. a foreign language), their teachers, assessment of their skills, and other aspects of language learning were related to their background characteristics and to the actions (practices) they reported to take place in the classroom. In particular, we were interested in finding out if specific links could be established between certain background characteristics, attitudes, and actions when it comes to setting goals, doing assessment and getting feedback.

In the first stage of the analysis, the attitude statements were analysed with exploratory factor analysis and the emerging factor structure was further examined with confirmatory factor analyses until a meaningful set of factors reflecting underlying attitudes was established. In the main stage of the analysis, the relationships among attitudes, background variables and reported actions (e.g. assessment and feedback practices) were studied with a variety of statistical approaches.

The results of the study shed light on, for example, the following questions: What is the relationship, if any, between students’ self-assessments of several in and out-of-school activities in L1/FL and their L1/FL grades? How is that relationship modified by students’ background (e.g. gender, stay abroad), their confidence in their ability to do self-assessment, their attitudes to assessment or by engagement in certain activities (e.g. type and amount of language-related free-time practices)?
An application of multidimensionality model-based DBF/DIF procedures to the Ontario Secondary School Literacy Test

Youn-Hee Kim and Eunice Jang, OISE / University of Toronto (younkim@oise.utoronto.ca, ejang@oise.utoronto.ca)

The ability to read and write effectively is a fundamental aspect of learning, and has long been recognized as a primary indicator of an individual’s success, both in and out of school. Literacy skills usually lead to high academic achievement in school and a satisfying professional career. In today’s increasingly knowledge-based society, attention has focused on the significant challenges for literacy education and assessment among linguistic minorities. According to the Education Quality and Accountability Office (EQAO), Grade 10 students enrolled in ESL/ELD programs in Ontario, Canada performed significantly lower than their native English-speaking peers, exhibiting a 51% success rate on the Ontario Secondary School Literacy Test (OSSLT) compared with a 84% success rate among the overall student population (EQAO, 2006).

Regarding the low performance of ESL/ELD students relative to native English-speaking students, it has been consistently pointed out that the standardized achievement tests, which have been developed on the basis of native English speakers’ performance, might pose an extra challenge for students with limited English proficiency (Abedi, 2004; Abedi & Hejri, 2004; Hofstetter, 2003; Koenig & Bachman, 2004; Snetzler & Qualls, 2000). If a test measures ability or knowledge in a way that favors native English speakers over limited English speakers, test fairness or test equity is threatened. One approach to ensure that a test is not biased against a certain group of test-takers is to examine the possible presence of differential item functioning (DIF). DIF is said to occur when test-takers of equal ability from different groups have a different probability of success on an item (Angoff, 1993; Camilli & Shepard, 1994). Over the past several decades, DIF has become a focus of research in the area of second language education and assessment, and a growing number of DIF studies have examined situations in which test-takers from diverse ethnic, racial, cultural, or linguistic backgrounds take second or foreign language proficiency tests.

This study investigated the extent to which the Ontario Secondary School Literacy Test (OSSLT) functioned differentially for students from different language backgrounds. Two language groups were of interest: Grade 10 students who speak only or mostly English at home (L1 students; n=1,969), and those whose home language is something other than English (ELL students; n=3,675). Based upon Roussos and Stout’s (1996) multidimensionality-based DIF analysis paradigm, a variety of substantive and statistical techniques were used: 1) ESL experts’ content review, 2) dimensionality analyses using HCA/CCPROX and DIMTEST, and 3) DBF/DIF procedures using SIBTEST. The substantive content analysis determined that six literacy subskills represented the construct of the OSSLT: textual comprehension skill, inferencing skill, vocabulary knowledge, grammar knowledge, summary skill, and integrated reading-and-writing skill. In addition, corroborating the results of the HCA/CCPROX analysis, three of six DIMTEST runs presented very strong evidence that vocabulary knowledge, grammar knowledge, and integrated reading-and-writing skill were statistically significant distinct dimensions. The results of the SIBTEST procedure further indicated that items associated with vocabulary knowledge favored L1 students, whereas items requiring grammatical knowledge or integrated reading-and-writing skill favored ELL students. Various instructional implications are discussed for the promotion of an effective literacy education program and the development of a literacy curriculum that can meet linguistically diverse learners’ needs in a multilingual context.

Are teachers teaching to the test? A case study of the College English Test (CET) in China

Hongli Li, Pennsylvania State University (hul151@psu.edu)

The College English Test (CET) in China is a high-stakes standardized test to assess college students’ English ability, and it is believed to have greatly influenced college English teaching in the country. One frequent claim against this test is that teachers may teach to the test, which could narrow the curriculum and turn regular English classes into CET coaching. In order to have more empirical evidence for the influence of the CET on classroom teaching, this study aims to find out whether teachers are teaching to the test and the reasons behind.
The CET traditionally consists of five sections: Listening, Reading, Vocabulary and Structure, Cloze (error-correction or translation), and Writing. In order to gain deeper insights into the influence of the CET on classroom teaching, only the writing section was examined in this study. Also, the huge differences among Chinese universities may make the already complex washback issue more complicated; thus only one university in Beijing was chosen to study. Both quantitative and qualitative methods were used. 20 teachers were surveyed with a questionnaire, and 8 of them participated in the follow-up interview. 122 first-year and 128 second-year students participated in the questionnaire survey, and 20 of the second-year students joined the follow-up group interview. Descriptive statistics, t-test, and Chi-square were used to analyze the quantitative data, and the qualitative data were also presented.

It was found out that the influence of the CET writing on classroom teaching was not as great as what has been claimed. For instance, 80% of the teachers said they would have the same concern over writing with or without the CET. While most teachers did not teach beyond the scope of the CET writing, they hardly changed their teaching methods as a result of the test. The possible reasons were discussed as follows. First, writing was not regarded as an essential language skill and only accounted for 15% of the CET total score; therefore, the general influence of this test on classroom teaching was not large. Second, the teaching of writing also heavily depended on time availability, teachers’ expertise and experience, teaching philosophy, class size, and school curriculum setting, etc, while the CET was only one of the influential factors. Third, some teachers did not have adequate knowledge of the CET, or they doubted the validity and reliability of the test. This might prevent them from teaching to the test. Finally, the impact of the CET writing on teachers was not as much as that on students. Teachers sometimes chose to teach to the test merely due to students’ pressure and request.

The study was significant because it gave detailed description of the teaching of writing as result of the CET, and it added important evidence to the CET washback study. Also, it gave an important signal that the test may not influence classroom teaching as we think, and many other important factors are involved in washback mechanism. Finally, it showed the different intensity of the washback on teachers and students due to their different stakes in the exam system.

Assessing reading comprehension with graphic organizers

Xiangying Jiang, West Virginia University (Xiangying.Jiang@mail.wvu.edu)

Reading comprehension is a complex construct. Current research has begun to move reading assessment beyond basic comprehension, and developments in reading research and theory are pushing the field to assess a wider range of reading abilities in valid and reliable ways (e.g., Britt et al., 1996; Perfetti et al., 1995; 1996; Trites & McGroarty, 2005; Wiley & Voss, 1999).

Graphic organizers (GOs) which reflect the discourse structure of a text have great potential for reading comprehension assessment. GOs are diagrams that reflect text structures and content information visually and hierarchically. Relatively few text structures (e.g., cause-effect, problem-solution, comparison-contrast) are commonly used to organize texts and they can be depicted through GOs. The GO completion task requires readers to fill in the partially completed GOs based on their understanding of the text organization and relationship among ideas.

This study investigates the applicability of the GO completion task as a new item type for reading comprehension assessment. Three GO completion tasks (based on three reading texts) were administered to 340 college EFL students together with the reading comprehension section of three retired TOEFL forms. The test forms were counterbalanced. First, the study shows that the GO completion task is a reliable measure. The internal consistency reliability coefficients (Cronbach’s alpha) for the three forms are .88, .84, and .82. They are even higher than those for the TOEFL reading comprehension tests, which are .83, .78, and .80 based on the current sample of the study. Second, the scoring of the GO completion test is manageable and reliable. With a detailed rubric of possible answers, rater effect in scoring is minimized. In this study, the inter-rater reliability coefficients (Cronbach’s alpha) for the three forms are all above .99. Third, there is a low moderate correlation of .43 (Pearson) between TOEFL reading comprehension and GO completion. This moderate relationship demonstrates that although they are both measures of reading comprehension, they tap different aspects of reading ability.
Results of the GO completion task reveal that students are not fully able to respond to the GOs if they do not have a fairly high “standard of coherence” for the larger organization of the text. The GO completion task represents a departure from traditional reading test tasks and constitutes a more complex task that requires greater cognitive processing. It taps into another level of comprehension assessment that is conceptually important in measuring students’ reading abilities. In addition, GO completion task can cover everything measured in a multiple-choice test and go beyond it. Like any multiple-choice comprehension test, GO items can focus on specific details, key concepts or vocabulary, and main ideas at different levels of the text. One GO diagram can be developed for each text structure to include multiple related items. Because of the fact that multiple GOs can be designed to reflect the main and local text structures and each GO diagram includes multiple items, the instrument creates an opportunity for many items for a text.

Assessment for learning - Providing diagnostic feedback and language enhancement guidance to test-takers

Eva Lui, City University of Hong Kong (eva.lui@cityu.edu.hk)

This poster session presents the development of a semi-automated system that generates feedback on test-takers’ diagnostic test performance and possible pathways for them to work on the weaknesses reported. In the university that this system is being developed, numerous English language enhancement opportunities are available ranging from credit-bearing language enhancement courses to co-curricular practical language activities. Various units are involved in the provision of such opportunities but still students seemed to take up only those that may have implications on their academic results or graduation. One possible reason for test-takers’ lack of motivation to improve their language may be the fact that the various provisions are more trainer-initiated. The approach of the project is to reward the motivated and hardworking students with a longitudinal guidance support. The starting point of the support is an Internet-based diagnostic test that test-takers can take at their convenience. All four language skills are covered and test-takers will receive feedback on their strengths and weaknesses at macro- and micro-skill levels. Related to the weaknesses identified, guidance on relevant enhancement provision that the test-taker could adopt will be provided. Effectiveness of the project relies on the analysis of skills assessed, training of markers and the mapping between skills and guidance feedback. The current poster will report on a study on analyzing skills assessed and the mapping between skills and guidance feedback.


Miyoko Kobayashi, Nobuko Hasegawa, and Nahomi Machida, Kanda University of International Studies (m.kobay@kanda.kuis.ac.jp, hasegawa@kanda.kuis.ac.jp, nahomijp@hotmail.com)

This poster presentation reports on the process of developing a vocabulary test for young learners of English in Japan, and discusses the results of pilot tests with approximately 100 young Japanese learners of English. The main purpose of the study is to work towards the establishment of an appropriate measure of children’s foreign language learning. In addition, the research examines the relationship between children’s test performance and the language learning environment.

The Ministry of Education in Japan has recently proposed that English be introduced to the fifth and sixth graders (final two years) of primary education in 2011. However, the precise objectives of this English language teaching have not been clearly defined and stakeholders - including teachers, teacher educators, and materials developers - are left without a clear sense of direction. This also means that the issue of assessment has been largely neglected. So far, the Ministry has simply recommended that English learning at primary level should be ‘fun’, and that learners should not be graded in terms of their achievement. However, according to recent language-learning theories (e.g. Dornyei 2001; Genesee & Upshur 1996; Tierney et al 1991; Torrance & Pryor 1998), wherever learning takes place, the key stakeholders, i.e. learners, should be given feedback in order to enhance their motivation.

Research suggests that vocabulary plays an important role in children’s early language learning (Dale & Fenson, 1996; MacArthur-Bates, 2003; Ogura 1999), and a significant part of children’s foreign language learning is de-
voted to the acquisition of vocabulary. Therefore it seems natural to assess young learners’ vocabulary knowledge. Nevertheless, the vast majority of research on vocabulary assessment has so far focused on adolescent and adult learners (e.g. Read 2000), and very little research has been conducted on young learner assessment (Cf. McKay 2006).

The research team has examined the words that appear in a number of widely-used international course books and examination materials for young learners of English, and has developed a working word list as a basis for test development. Based on Nation’s categories (1990, 2001), a framework for a vocabulary test has been developed and a series of pilot tests conducted. The poster reports on this test development process and discusses the preliminary findings of the pilot tests. It also shows the results of further examination into the relationship between the children’s test performance and a range of factors such as age, length of English learning, L1 literacy, L2 literacy, and family backgrounds. The presenters further discuss the use of this kind of instrument to measure children’s vocabulary development and suggest future directions for research into young language learner assessment.

Beyond CTT and IRT: An interactional testing model for writing performance assessments

Xin Wang, University of Illinois at Urbana Champaign (xinwang2@uiuc.edu)

The current study as a doctorate dissertation investigates the inconsistency between the nature of writing performance tests and the current measurement theories. In order to fill in the gap in the previous research methodologies, this study proposes an alternative testing model which is able to address the interaction between different stakeholders of writing test.

A writing test as a language performance assessment is a multifaceted entity involving the interaction of various stakeholders, among which essay raters has a great impact on essay scores due to their subjective scoring decision, hence influencing the test validity and reliability (Huot, 1990; Lumley, 2002). This understanding puts forward the demand on the development and facilitation of methodological tools to quantify rater’s decision making process and the interaction between rater and other stakeholders in a language test. Unfortunately, previous studies within the framework of Classic Testing Theory (CTT) and Item Response Theory (IRT) mainly focus on the final outcome of rating or the retrospective survey data and raters’ think-aloud protocols. Very few, if any, have directly examined the very rating process and the interaction per se.

The present study proposes a behavioral model for writing performance test, which addresses raters’ scoring behavior and their reading comprehension as combined with the final essay score. Bearing the “interactional” nature of a performance test, this new testing model is named as the Interactional Testing Model (ITM). Though the focus of this study is writing assessment, the ITM is applicable to the whole field of performance-based testing. The ITM framework considers the process of a language test as the interaction among all stakeholders including test maker, test taker and essay rater, etc. In the current study that focuses on writing performance test, the interaction between test maker and test taker is realized directly through test prompt and indirectly through score; on the other hand, the interaction between test taker and essay rater is directly reflected in the writing responses. This new model defines and explores rater reliability and test validity via the interaction between the text (essays written by test-takers) and essay rater. Instead of indirectly approaching to the success of such an interaction through the final scores, the ITM directly measures and examines the success of raters’ behaviors with regard to essay reading and decision making. In order to examine the online evidence of raters’ decision making, a computer-based interface was designed for this study to automatically collect the time-by-location information of raters’ eye movement, their text comprehension and other scoring events including score assignment, scoring comments and sentence annotation made by essay raters.

As a superset of IRT and CTT, the ITM does not attempt to reject these two measurement methods. Instead, it expands the possible research range into a realm of new phenomenon that is beyond the current consideration of traditional measurement theories. Comparing to CTT and IRT, the ITM bears both practical and theoretical significance. On the practical side, this model may shed light on the study of test reliability and validity. The ITM framework provides a quality control tool to ensure that essay raters comprehend the content of a writing response, follow rating rubrics and assign test scores in a consistent manner, hence providing a comprehensive interpreta-
Bridging the gender gap in computerized speaking test performance

Eunjyu Yu, The Ohio State University (yu.211@osu.edu)

Because of the significant social impact of testing, we should achieve fairness and justice in testing. Currently, computer technology has been increasingly used as a delivery medium of both knowledge and assessment. To date, despite critical theorists’ concerns about gendered technology (Cockburn, 1992; Gill & Grint, 1995; Shohamy, 1998, 2001), few data-driven studies are available on the relationship between gender and technology. In particular, little is known about gender effects on computerized speaking tests. Thus, it is necessary to explore further what relationship exists between gender difference and test medium, especially on technology-mediated speaking tests.

Grounded in critical theory, this study investigated what impacts test takers’ gender, their attitudes, and test medium have on technology-mediated speaking test performance. Using mixed methods, data was collected from 208 non-native English speakers in a major public US university. The ANCOVA statistics found a significant interaction effect of gender and test medium on test performance. Surprisingly, females outperformed males on the computerized speaking test, whereas the taped test results revealed even performance across gender. Further analyses of the qualitative data found that both genders had a positive attitude toward the computerized speaking test but they expressed a negative attitude toward the audio-taped SPEAK test.

This study made suggestions to reduce possible confounding effects and to improve technology-mediated speaking tests. The findings of this study will contribute to narrowing gender gap in technology-mediated speaking tests. This study also helps language test developers use technology, particularly computer technology effectively to measure communicative competence.

Chinese EFL learners’ perceptions of their experiences with a portfolio-based writing assessment project: A case study

Qinghua Li, Shaoxing University (qinghua1965@163.com)

This article describes a case study that attempts to investigate Chinese EFL learners’ perceptions of their experiences with a portfolio-based writing assessment project. The study involved four students who were selected from the experimental group, a sophomore English major class in a local university of Zhejiang province in China. Analysis on the data from journals, interviews and reflective essays showed that taking part in the PBWA project not only empowered them to take charge of their own learning but also promoted their motivation to write in English and they held more interest in English writing, improved their writing strategies and shifted their attention from products of writing to process of writing and welcomed PBWA as a fairer assessment tool than timed one-shot essay tests and thus preferred scores from PBWA as a better indicator of their writing ability.

Key words: Chinese EFL learners; portfolio-based writing assessment; case study

Developing online tests for advanced language assessment

Martyn Clark, Center for Applied Second Language Studies (martyn@uoregon.edu)

Efforts to incorporate proficiency-oriented teaching at the advanced levels are often hampered by the lack of appropriate assessment instruments. This poster will describe the development of advanced level reading and listening items as part of a comprehensive Web-based assessment instrument to measure student proficiency in a second
language. Reliable proficiency data will improve academic placement and articulation of students and support postsecondary language educators in evaluation and program improvement efforts. The Center for Applied Second Language Studies (CASLS) at the University of Oregon is developing online assessments in six languages: Chinese, French, German, Hindi, Japanese, and Spanish.

The presenter will discuss the test design, outline the item development process, show sample items, and present preliminary results from pilot testing. In addition, tensions between practical constraints, end user desires, and validity concerns will be explored. This project is being funded by a Fund for the Improvement of Postsecondary Education (FIPSE) grant.

**Investigation of the Tunghai English Placement Examination using factor analysis**

Jonathan Whittinghill, California State University, Los Angeles (jcwhit1@yahoo.com)

This study examines the test structure of the Tunghai English Placement Examination (TEPE) using item-level data collected between 1998 and 2006. Data is being analyzed to ascertain the structure of the test in each year for which data are available, and determine whether that structure has changed between subsequent administrations. Exploratory factor analysis will be performed at both the section level (grammar, reading, and listening) and at the passage level. Results will be compared with the proposed test structure as outlined in previous research in the interest of construct validity. Furthermore test structures derived for each year will be analyzed for any differences. Such differences, if found, could be indicative of a shift in the population each sample of students purports to represent. Finally, results will also be viewed in light of previous research indicating significant changes in test performance in different sections of the test.

**Language assessment for the business processing outsourcing industry**

Gail Forey and Jane Lockwood, Hong Kong Institute of Education (eggail@polyu.edu.hk, lockwood@ied.edu.hk)

The business processing outsourcing (BPO) industry domiciled in the USA, UK, Australia and New Zealand is rapidly expanding into such non-English speaking destinations as Eastern Europe, South America, India, China and the Philippines. However, many BPO multinational companies are experiencing great difficulty in screening and recruiting for BPO work, particularly in the call centres. These companies are also having difficulty in diagnosing where communication is breaking down on the calls and in maintaining good quality service standards.

There has been great disappointment with commercially available business English tests in providing language assessment solutions for this fast-growing industry. This poster session will:

(i) describe the BPO industry need for language assessment
(ii) describe a solution currently being developed by the industry and a training provider in the Philippines (the Business Processing Language Assessment Scales - BUPLAS)
(iii) outline a series of language assessment research questions emanating from industry and language trainers in BPO worksites
Perceptions on English academic writing rubrics: A comparison between Chinese teacher educators and Chinese pre-service teachers

Danbin Wang, *Shanghai Normal University* (wdanbin@hotmail.com)
Ye He, *University of North Carolina at Greensboro* (y_he@uncg.edu)

The purposes of this study are to explore the similarities and differences between Chinese university teacher educators’ and pre-service teachers’ perceptions of academic English writing, and to identify its implications in English teacher education in China. 10 teacher educators and 10 pre-service teachers in China were asked to rate 3 college student essays based on the rubric they modify and provide an overall rating of the essays. ANOVA were used to identify the differences in participants’ ratings. Content analysis was also conducted based on Bachman’s (Bachman, 1990) language competence model. The findings of this study provided researchers deeper understanding of the design of academic writing rubric for Chinese students, which will certainly benefit both the teacher educators and pre-service teachers, especially pre-service teachers who will have the responsibility of preparing their students for high-stake tests and rating high-stake test compositions themselves.

TBEST: A new on-line English speaking test in China

Jun Liu, *University of Arizona* (junliu@u.arizona.edu)

Task-Based English Speaking Test (TBEST) is designed to measure ESL/EFL learners’ oral English ability through a number of functional tasks (i.e., narrative, interpretive, argumentative, and integrative). Each task requires 90 seconds to complete on-line or through mobile phones. Students’ oral recordings are digitalized and evaluated by two trained raters through a 3-scale/100 point grading rubrics in four areas (i.e. comprehensibility, fluency, accuracy, and complexity). In this presentation, Liu is going to report the design and development of the instruments and also the preliminary results of the testing among 10000 university students in China. Research findings and implications will be discussed apart from a demo.

The assessment use argument: A web-based Spanish placement listening exam

Cristina Pardo, *Iowa State University* (cpardo@iastate.edu)

In order to implement an effective foreign language program it is very important to determine student placement based on their true language knowledge as opposed to how much “seat time” instruction they have had. This poster describes the design of a listening exam and investigates validity issues using Bachman’s (2005) framework of an assessment use argument. The purpose of this study is twofold: First, to interpret the listening scores for the listening exam as a way to know what test-takers can do and, second, to place students in Spanish classes.

First, this study includes investigations of the test usefulness to support the warrants of the validity argument for the Spanish listening exam. To ensure the test usefulness, the test development included investigations of internal consistency, IRT analyses, ANOVA analyses, and planned comparisons for content relevance, criterion-relatedness, and content coverage in order to support the construct validity, and feedback analyses for a post-test study. Moreover, the online listening exam was administrated to 147 Spanish learners and participants’ perceptions about the exam similarity with instructional tasks were analyzed as proof of the usefulness of the test. Second, since the intended use of the test is to place students in Spanish classes, this study also includes investigations of the test appropriateness for this purpose.

This poster concludes that despite a few exceptions that deal with specific linguistic features, the SLE effectively tests Spanish proficiency based on grammatical, discourse and sociolinguistic features. The SLE demonstrated that different proficiency groups of test-takers performed differently. Participants with more listening ability demonstrated higher proficiency levels making the correct placement according to real abilities possible.
The washback effect of a high stakes assessment on EFL teachers’ teaching strategies and assessment practices in Taipei City

Yu-Ching Chan, Graduate School of Children English Education, National Taipie University of Education (yuching@tea.ntue.edu.tw)

Going hand in hand, teaching enriches students’ learning contents, while assessment evaluates students’ learning results. In recent decades, assessment has gained increased attention in EFL education. One of the important issues in the practice of assessment is the washback effect of testing on teaching and learning ((Buck, 1988; Hughes, 2003). Washback effect can be positive (beneficial) or negative (harmful) on teaching and learning. Much research has been focused on the washback effect of assessment on students’ learning (Watanabe, 1996) or large-scale, standardized tests (Imao, 2001). In fact, assessment has powerful influence on teaching, too. Wall (1998) claimed that high stakes tests might induce the impact on teaching methodology and content. In Taipei City, the Basic English Learning Assessment (BELA) is the only official unified English assessment for elementary students. Its purpose is to check students’ English abilities and to decrease the transition problems between elementary and junior high school students. The BELA was first held in 2003 and has become the major work for EFL teachers. However, no studies have been conducted to examine teachers’ attitudes towards the BELA, to explore whether the BELA facilitates or inhibits teachers’ teaching strategies and assessment practices, and to figure out the difficulties EFL teachers encountered. To this end, the quantitative study financially sponsored by the National Science Council was conducted. Three hundred EFL teachers randomly selected from 12 Districts in Taipei City answered a questionnaire containing 41 questions in the Likert five-point scale and 4 open-ended questions. The findings not only indicated most EFL teachers took a slightly negative attitude towards the BELA but also revealed the washback effects of the BELA on EFL teachers’ teaching strategies, teaching methods, time allotment, and assessment practices. The results of this first study about the BELA provided a practical example of how a high stakes assessment affected teaching and assessment in practice.

Using performance descriptors to develop a criterion-referenced test of language proficiency: An exploratory study

Gary Buck, University of Michigan (garybuck@umich.edu)

Recently there has been renewed interest in general performance descriptors of language proficiency: due mainly to the development and expanding use of the Common European Framework of Reference, as well as the recent application of the Interagency Language Roundtable skill level descriptors to a variety of ‘less common’ Asian languages. This is obviously welcome to critics of the norm-referenced test model, who feel that language proficiency should be measured against criterion levels of ability, rather than measured against other test takers. A feeling the authors share.

The question is how do we develop tests that measure test-takers against a set of performance descriptors, or more technically, how do we calibrate our tests against the scale descriptors? There are two basic methods in current use: the first is to administer the test to a group of people of known proficiency level against which the test can be calibrated; regrettably, such groups are rarely available. The second is to estimate the difficulty of items, through multiple expert judgments, somehow aggregate that data, then base cut scores on that information. The problem is that experts often differ considerably on their evaluation of the difficulty of particular items, and the results of standard-setting studies can vary considerably depending on the raters, or the method of aggregating the data.

However, the biggest drawback of these methods is that they are post hoc--essentially, the test is first made, and then the test developer tries to figure out how the results should be interpreted. There is no theoretical basis to go from the scale descriptors to the interpretation of test-taker performance. The purpose of this study is to address this lack by exploring a methodology for building criterion-referenced tests based on a set of scale descriptors. There are two parts to the study: the development of a theoretical model, and the validation of that against test data.

To develop the model: First, items are written to address the main characteristics of each level on the scale -- in this case the scale is the ILR scale, but the method applies just as well to any similar set of performance descriptors.
Second, the performance of people at each level of the scale is modeled in probabilistic terms against items at each level of the scale. Third, these probabilities can then be summed in order to place people on the descriptive scale, based on their test performance.

To validate the model: the test is administered to a group of test takers, the test is scored and scaled using standard IRT procedures, and actual person scores can be compared to predicted theoretical scores. The extent to which the actual scores support the theoretically derived scores indicates the extent to which the test model describes actual performance.

Results will be discussed, implications will be drawn, and recommendations made for the development and further exploration of the methodology.

**Where do teachers go wrong in item writing?**

Nathan Carr, *California State University, Fullerton* (ncarr@fullerton.edu)

As Bachman & Palmer (1996) point out, it is important for language teachers to develop a certain level of competence in language assessment. One area in which a lack of such competence can often be seen is in the construction of multiple-choice test questions. For all their flaws (see, e.g., Hughes, 1989), this task type is sometimes the best choice available, particularly in cases where the rapid scoring of comprehension items is necessary. Little has changed since Oller (1979) noted that creating good multiple choice tests is generally not practical at the classroom level, partly because of the difficulty associated with constructing them appropriately. Nevertheless, when language tests are to be written for use by an entire language program, such as in the case of placement or graduation tests, most or all of the actual writing of the test is often delegated to teachers in that program. These teachers may be given specifications for writing items, perhaps with lists of common item-writing errors which they should avoid (see, e.g., Brown, 2005; Brown & Hudson, 2002); nevertheless, the items they produce tend to require extensive editing and revision.

The present study focuses on the types of errors commonly found in teacher-created multiple-choice items written to assess reading and listening comprehension. The reading and listening tests in question were developed to be administered along with tests of speaking and writing, and are intended to be administered to students attempting to earn credit by examination in intermediate level language courses at a large public university in the United States. These tests were developed for courses in nine languages following the same specifications. The present study focuses on the types of problems identified with items prior to review by the committee overseeing the test development process. Items that were rejected or flagged for revision were typically problematic for one of several broad categories of reasons: low passage dependence, item interdependence, or problems with the technical quality of the items (e.g., grammatical inconsistency across distractors, overlapping options, or items using “not” or “all of the above”). Additionally, some teachers showed a tendency to write too many specific detail items, or items requiring only the recognition of basic factual information from the passages, to the exclusion of other item types, such as inference items.

This presentation will discuss these categories in greater detail, with illustrations from the tests in question. It is hoped that the results described here will be of use in training teachers as item writers in other programs as well.
Presenters Index

A
Abdul-Kadir, Kadeessa 95, 100
Alderson, J. Charles 47
Amdur, Lisa 82
Andrews, Stephen J. 61
Atia, Samira El 47

B
Bachman, Lyle 59
Bachman, Patina 59
Baffi-Bonvino, Melissa 65
Banerjee, Jayanti 47
Brau, Maria 100
Brooks, Rachel 100
Brown, Annie 42
Buck, Gary 116

C
Cai, Hongwen 98
Carless, David 74
Carr, Nathan 55, 117
Chang, Sandy 59
Chan, Yu-Ching 116
Chau, Michael 60
Cheng, Liying 43, 61
Chen, Lihui 87
Chinda, Bordin 87
Chi, Youngshin 95, 100, 107
Cho, Chih Yuan 107
Christison, Mary Ann 34
Chu, Hsiu-yu 86
Chun, Christian 75
Clark, Martyn 113
Consolo, Douglas 65
Cook, Gary H. 103
Cummins, Jim 75

d
Dai, Binbin 104
Davidson, Chris 45
Davies, Alan 38, 76
Davies, Geraldine 98
Davies, Gerry 33
Davis, Larry 103
Davison, Chris 33
de Jong, John H.A.L. 97
Deng, Xiangjun 94
Dobson, Barbara 92, 105
Downey, Ryan 80, 94

E
Elder, Cathie 39, 41

F
Fang, Felicia 46
Forey, Gail 114
Fox, Janna 39

G
Gao, Manman 67
Garner, Mark 82
Green, Tony 68, 95

H
Hamp-Lyons, Liz 33, 42, 45, 76
Han, Baocheng 74
Han, Su-mi 64
Harding, Luke 41, 56
Hasegawa, Nobuko 111
He, Ling 83
Hilden, Raili 66
Hill, Yao 63
Hirai, Akiyo 90
Hsueh, Ching-Ni 69
Hsu, Huei-Lien 95, 100, 107
Huang, Dayong 82
Huhta, Ari 108

I
Imao, Yasuhiro 99
Inbar, Ofra 38, 40
In’nami, Yo 59

J
Jang, Eunice 53, 75, 109
Jang, So-young 95, 100, 107
Jeong, Heejeong 73
Jiang, Xiangying 110

K
Kadir, Kadeessa 107
Kim, Hyeyeong 83
Kim, Jiyoung 49, 95, 100
Kim, Youn-Hee 53, 109
Kinoshita, Toru 51
Knoch, Ute 42, 84
Kobayashi, Miyoko 111
Koch, Elize 78
Koizumi, Rie 59, 90
Künschak, Claudia 70

L
Landon, John 78
Lee, Carolyn 77
Lee, King Siong 71
Lee, Yong-Won 104
Leon, Seth 59
Li, Hongli 69, 109
Lim, Gad 92
Lin, Chih-kai 95, 100
Li, Qinghua 113
Liu, Jinhui 55
Liu, Jun 115
Liu, Meihua 94
Liu, Qin 106
Li, Zhi 88
Llosa, Lorena 35
Lockwood, Jane 114
Lui, Eva 111
Lumley, Tom 46

M
Machida, Nahomi 111
Marcoulides, George A. 32
Mathew, Rama 79
McCollum, Robb 80
Morimoto, Yuko 93

N
Nakagawa, Chikako 102
Nakamura, Yuji 96
Neittaanmäki, Reeta 108
Nigolian, Mark 106
Nissan, Susan 54
<table>
<thead>
<tr>
<th>Presenter</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walther, Ingeborg</td>
<td>77</td>
</tr>
<tr>
<td>Wang, Danbin</td>
<td>115</td>
</tr>
<tr>
<td>Wang, Huan</td>
<td>57, 85</td>
</tr>
<tr>
<td>Wang, Jing</td>
<td>43</td>
</tr>
<tr>
<td>Wang, Xin</td>
<td>112</td>
</tr>
<tr>
<td>Wang, Xu</td>
<td>75, 81</td>
</tr>
<tr>
<td>Wang, Zhenhua</td>
<td>106</td>
</tr>
<tr>
<td>Watt, David</td>
<td>52</td>
</tr>
<tr>
<td>Whittinghill, Jonathan</td>
<td>114</td>
</tr>
<tr>
<td>Wigglesworth, Gillian</td>
<td>51</td>
</tr>
<tr>
<td>Wilmes, Carsten</td>
<td>103</td>
</tr>
<tr>
<td>Wolf, Mikyung Kim</td>
<td>59</td>
</tr>
<tr>
<td>Wu, Jessica</td>
<td>52</td>
</tr>
<tr>
<td>Wu, Weiping</td>
<td>50</td>
</tr>
<tr>
<td>Wu, Yongfei</td>
<td>88</td>
</tr>
<tr>
<td>Xie, Qin</td>
<td>76</td>
</tr>
<tr>
<td>Xing, Ru</td>
<td>85</td>
</tr>
<tr>
<td>Xi, Xiaoming</td>
<td>36, 55</td>
</tr>
<tr>
<td>Yang, Luxin</td>
<td>74</td>
</tr>
<tr>
<td>Youn, Soo Jung</td>
<td>102</td>
</tr>
<tr>
<td>Yu, Eunjuyu</td>
<td>113</td>
</tr>
<tr>
<td>Yu, Guoxing</td>
<td>58, 89</td>
</tr>
<tr>
<td>Yu, Ying</td>
<td>61, 72</td>
</tr>
<tr>
<td>Zhang, Hongxia</td>
<td>87</td>
</tr>
<tr>
<td>Zhang, Jie</td>
<td>73</td>
</tr>
<tr>
<td>Zhang, Lily</td>
<td>70</td>
</tr>
<tr>
<td>Zhang, Wenxia</td>
<td>94</td>
</tr>
<tr>
<td>Zhang, Xinling</td>
<td>91</td>
</tr>
<tr>
<td>Zheng, Ying</td>
<td>79</td>
</tr>
<tr>
<td>Zhong, Qi</td>
<td>69</td>
</tr>
<tr>
<td>Zhou, Fanglin</td>
<td>87</td>
</tr>
<tr>
<td>Ockey, Gary J.</td>
<td>35</td>
</tr>
<tr>
<td>Ohlrogge, Aaron</td>
<td>105</td>
</tr>
<tr>
<td>Oishi, Harumi</td>
<td>51</td>
</tr>
<tr>
<td>O’Loughlin, Kieran</td>
<td>41</td>
</tr>
<tr>
<td>Palmer, Adrian</td>
<td>34</td>
</tr>
<tr>
<td>Papageorgiou, Spiros</td>
<td>96</td>
</tr>
<tr>
<td>Pardo, Cristina</td>
<td>115</td>
</tr>
<tr>
<td>Peng, Kangzhou</td>
<td>85</td>
</tr>
<tr>
<td>Qin, Jingjing</td>
<td>64</td>
</tr>
<tr>
<td>Palmer, Michelle R.</td>
<td>68</td>
</tr>
<tr>
<td>Rea-Dickins, Pauline</td>
<td>89</td>
</tr>
<tr>
<td>Ren, Hongtuo</td>
<td>104</td>
</tr>
<tr>
<td>Rubin, David</td>
<td>80</td>
</tr>
<tr>
<td>Ryan, Kerry</td>
<td>41, 57</td>
</tr>
<tr>
<td>Sankar, Vyjayanthi</td>
<td>66</td>
</tr>
<tr>
<td>Saville, Nick</td>
<td>48</td>
</tr>
<tr>
<td>Sawaki, Yasuyo</td>
<td>36, 54</td>
</tr>
<tr>
<td>Shiotso, Toshihiko</td>
<td>92</td>
</tr>
<tr>
<td>Shohamy, Elana</td>
<td>38, 40</td>
</tr>
<tr>
<td>So, Youngsoon</td>
<td>71</td>
</tr>
<tr>
<td>Spolsky, Bernard</td>
<td>38</td>
</tr>
<tr>
<td>Strong-Krause, Diane</td>
<td>93</td>
</tr>
<tr>
<td>Suen, Hoi K.</td>
<td>69</td>
</tr>
<tr>
<td>Suzuki, Masanori</td>
<td>80, 94</td>
</tr>
<tr>
<td>Tan, May</td>
<td>44</td>
</tr>
<tr>
<td>Tarnanen, Mirja</td>
<td>108</td>
</tr>
<tr>
<td>Tavares, Nicole</td>
<td>33, 45, 83</td>
</tr>
<tr>
<td>Tufts, Clare</td>
<td>77</td>
</tr>
<tr>
<td>Turner, Carolyn E.</td>
<td>43</td>
</tr>
<tr>
<td>Urmston, Alan</td>
<td>46</td>
</tr>
<tr>
<td>Van Bonn, Sarah</td>
<td>105</td>
</tr>
<tr>
<td>Van Moere, Alistair</td>
<td>80, 94</td>
</tr>
<tr>
<td>Van Natten, Liberty</td>
<td>101</td>
</tr>
</tbody>
</table>
Notes
Two levels of certification for English language students outside the USA—

ECPE Examination for the Certificate of **Proficiency** in English

ECCE Examination for the Certificate of **Competency** in English

- Documentation of language proficiency for university admission, employment, or career advancement
- Group exams administered in 25 countries
- Secure test batteries with measurements of listening, speaking, reading, and writing
- Recognized by thousands of institutions worldwide

Other ELI-UM Publications—

MELICET • LCT • MTELPE • EPT

Designed for commercial and educational institutions for placement in intensive programs and for assessment of academic readiness

Founded in 1941 by Charles C. Fries, the University of Michigan English Language Institute originated the first intensive English language research and training program of its kind in the world. The ELI is committed to serving the ESL/EFL community through conducting research, promoting teacher education, and by providing high-quality English language education programs and materials. The Testing and Certification Division develops and administers internationally recognized examinations in accord with the highest professional standards.

www.lsa.umich.edu/eli
Foreign Language Teaching and Research Press (FLTRP) is the largest of its kind in China in terms of the number of foreign languages, categories of publications, and formats of products. Its range of products includes textbooks, supplementary materials, dictionaries, academic works and readers for all age groups from primary to graduate levels. Its publications, in more than 30 languages, come in print and a variety of digital forms, including CDs, CD-ROMs and online materials.

A true leader in the market of foreign-language publications in China, FLTRP puts forth more than 800 new books every year and has a backlist of more than 3,000 titles. In the past 15 years, its annual sales have maintained a steady growth of over 30% on average to reach RMB1.6 billion in 2007.

FLTRP looks forward to deepening cooperation with existing partners and to developing new ties with new friends.

Higher English Education Publishing (HEEP), the largest publishing division of FLTRP, is committed to providing English teaching and learning materials to higher learning institutes around China. It publishes English coursebooks and supplementary multimedia materials for starters, higher vocational school students, and for undergraduates and graduates of both English majors and non-English majors in colleges and universities. For the better use of the coursebooks, HEEP provides supporting materials and teacher training programs to college English teachers. It pays close attention to trends in English teaching and organizes conferences and workshops on language teaching and research at various levels, including the 3rd, 4th and 5th International Conference on ELT in China. It also holds the “CCTV Cup” National Speaking Contest for college students each year.

HEEP also publishes two journals, *Teaching English in China*, with the China English Language Education Association (CELEA), and *Foreign Language Education in China*, with the National Research Centre for Foreign Language Education. With their influence reaching far and wide, the journals have become a window on English language teaching and research in China.
Founded in December 1979, Shanghai Foreign Language Education Press has grown into one of the largest and most prestigious foreign language publishers in China. Its main areas of publication include foreign language textbooks for institutions of higher learning and all other levels of schools, pedagogical reference books, academic works, dictionaries and periodicals, materials for teaching Chinese as a foreign language, and multi-media publications.

Since its foundation, SFLEP has published more than 6,000 titles in nearly 30 different languages, with the total number of volumes exceeding 500 million. Of the titles published so far, more than 500 have won prizes at provincial and ministerial levels. The textbooks for general English are used by almost a thousand colleges and universities in China, while the course books for English majors are used in more than 60% of Chinese universities. It has published over 600 academic works reflecting cutting-edge studies in linguistics, teaching methodology, literary, cultural and translation studies, and so on. Most of them have had a profound influence on China's foreign language research and are the fruits of key research projects in social science in China.

SFLEP actively participates in international copyright cooperation. Currently, there are 34 foreign presses which have cooperative relationships with SFLEP and more than 50 have established friendly cooperative relations with SFLEP.

Shanghai Foreign Language Education Press is working hard to develop a comprehensive platform combining publishing, scientific research, and education. With emphasis on product quality, scale development, distinctive products, and brand names, SFLEP has now gained a leading position in China's publishing industry and will forge ahead in pursuit of professional excellence.
AimHigher!

Pearson Language Tests provide examinations to assess and accredit the English Language usage of students who are speakers of other languages.

We deliver the London Tests of English and the London Tests of English for Children and we are developing a new academic international English test, the Pearson Test of English, in collaboration with Graduate Management Admission Council® (GMAO)®.

Visit us at our table to find out more.
Make the right decision using the TOEIC® tests — an accurate measure of workplace English skills.

The TOEIC tests are given to employees or applicants to determine if they have the ability to comprehend and use English effectively in a workplace environment.

The TOEIC tests can help you:
- Find out who has the English skills to succeed in the global workplace
- In recruiting, developing and retaining the most qualified people
- Save time and resources in finding the right person for the right job
- By giving you a common standard of English-language measurement

Put the TOEIC test to work for you

To learn more about the TOEIC tests and how they can benefit your organization, call +1-609-771-7170, go to www.ets.org/toeic/ltrc or e-mail globalforum@ets.org.
Language testing resources from Oxford

NEW online practice tests
www.oxfordenglishtesting.com

Interactive practice tests supported by:
- Online dictionary look-up
- Instant marking & feedback
- Exam tips
  ... and more

Available now for IELTS, PET, FCE, CAE exams

For more information on Oxford’s resources for language testing, please refer to the brochure in the Conference Pack or visit the online catalogue at www.oup.com/elt
Cambridge University Press advances learning, knowledge and research worldwide.

Cambridge University Press was founded by a royal charter granted to the University of Cambridge by King Henry VIII in 1534. It is the oldest printer and publisher in the world, having been operating continuously since 1584, and is one of the largest academic publishers globally. Its purpose is to further the University's objective of advancing learning, knowledge and research.

Cambridge Learning:
Cambridge is a global leader in English Language Teaching (ELT) books and resources and is well established as a publisher of books for schools in the United Kingdom, Australia and Africa. Cambridge-Hitachi - a joint venture between Cambridge University Press and Hitachi Software Engineering - produces a range of curriculum-based software developed with the changing needs of today's teachers in mind. Cambridge Learning consolidates these strands of educational publishing into a single brand.

Cambridge Language Assessment
This series provides systematic and comprehensive coverage of the field of language assessment. It combines up-to-date research and theory of good practice with a practical, 'how to' element.

<table>
<thead>
<tr>
<th>Assessing Grammar</th>
<th>9780521003445</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessing Listening</td>
<td>9780521666619</td>
</tr>
<tr>
<td>Assessing Reading</td>
<td>9780521599993</td>
</tr>
<tr>
<td>Assessing Speaking</td>
<td>9780521804875</td>
</tr>
<tr>
<td>Assessing Vocabulary</td>
<td>9780521627412</td>
</tr>
<tr>
<td>Assessing Writing</td>
<td>9780521784467</td>
</tr>
</tbody>
</table>

Contact us:
Cambridge Shanghai Office:
Block N, Floor 19, Zhiyuan Building 768 Xiu Road, Shanghai, China
Tel: 021-5301 4700-802
Fax: 021-5301 4710
Email: shanghai@cambridge.org
ppan@cambridge.org

Cambridge Beijing Office:
Room 1208-09, Tower B, Chengjian Plaza
No. 18 Beitaizi Road, Haidian District, Beijing 100088
Tel: 010-8227 4100
Fax: 010-8227 4185
Email: beijing@cambridge.org

Cambridge Guangzhou Office:
Room 1501, East Tower Dongshan Plaza
69 Xian Li Zhong Road
Guangzhou, 510095
Tel: 020-8732 6915
Fax: 020-8732 6993
Email: guangzhou@cambridge.org
The Encyclopedia of Language and Education offers the newest developments in the field, including two new volumes of research and scholarly content essential to the field of language teaching and learning in the age of globalization. Each of the ten volumes presents core information and is international in scope, as well as diverse in the demographics and socioeconomic populations it covers. The Encyclopedia includes a special emphasis on multicultural and bilingual education, reflecting current and emerging trends in the field.

This essential reference:

► Is necessary for every university and college library in the world that serves a faculty or school of education
► Features more than 250 contributors drawn from numerous countries
► Presents information and authoritative insights relevant to every country and every language in each volume

Print Edition
2008. Hardcover
► $1,500.00

eReference
2008. eReference
► $1,500.00

Print + eReference
2008. (Set of print and e-version)
ISBN 978-0-387-35420-0
► $1,875.00

Easy Ways to Order for the Americas ► Write: Springer Order Department, PO Box 2485, Secaucus, NJ 07096-2485, USA ► Call: (toll free) 1-800-SPRINGER
► Fax: 1-201-348-4505 ► Email: orders-ny@springer.com or for outside the Americas ► Write: Springer Distribution Center GmbH, Haberstrasse 7, 69126 Heidelberg, Germany ► Call: +49 (0) 6221-345-4301 ► Fax: +49 (0) 6221-345-4229 ► Email: SDC-bookorder@springer.com ► Prices are subject to change without notice.
All prices are net prices.
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

From Single Words to Passages: Contextual Effects on Predictive Power of Vocabulary Measures for Assessing Reading Performance, David D. Qian

Strategic Competence as a Fourth-Order Factor Model: A Structural Equation Modeling Approach, Aek Phakiti

Using Microsoft Excel® to Calculate Descriptive Statistics and Create Graphs, Nathan T. Carr

Validating TOEFL® iBT Speaking and Setting Score Requirements for ITA Screening, Xiaoming Xi

Recommending a Nursing-Specific Passing Standard for the IELTS Examination, Thomas R. O’Neill, Chad W. Buckendahl, Barbara S. Plake, and Lynda Taylor

To access a free online sample, visit: www.tandf.co.uk/journals/titles/15434303

2008 Subscription Rates

Individual print and online discounted rate:

US$43 / £22 / •27

Institutional online-only rate:

US$299 / £179 / •239

Institutional print and online rate:

US$315 / £189 / •252

325 Chestnut Street, Philadelphia, PA 19106
Email: customerservice@taylorandfrancis.com
Call Toll Free: 1-800-354-1420, Press “4”
Fax: (215) 625-8914
Getting a message across is not always this easy.

Our mission is to provide test development, translation/adaptation, and other related services in the areas of second language proficiency testing and the testing of nonnative English-speakers.

Second Language Testing, Inc.

For further information:
Please be sure to visit our table in the exhibits area or contact Charles Stansfield at cstansfield@2LTI.com or at:

6135 Executive Boulevard
Rockville, MD 20852-3901
United States of America
Ph: 301 231 6046 • F: 301 231 9536
www.2LTI.com
Studies in Language Testing

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 Peter Falvey (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
Lijing 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)

Alan Davies (2008)

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

25 IELTS Washback in Context: Preparation for academic writing in higher education
Anthony Green (2007)

26 Examining Writing: Research and practice in assessing second language writing
Stuart D Shaw and Cyril J Weir (2007)

Forthcoming volumes

27 Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity
Proceedings of the ALTE Berlin Conference, May 2005
Lynda Taylor (Ed) (2008)

28 Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams
Roger Hawkey (2009)
Built on a solid research base
Our in-depth needs analysis, academic corpus, 125 research reports and 30 monographs, and multiple pilot and field tests provide a strong foundation to support the unique design of the TOEFL® iBT.

A highly innovative test
The TOEFL iBT is designed to assess all the critical language skills needed for academic success — reading, listening, speaking and writing — in a groundbreaking way. It features integrated tasks that simulate the kinds of communication expected of students in university classrooms and in campus-based situations.

Meets the highest scoring standards
TOEFL iBT Speaking and Writing responses are encrypted and sent electronically to the Online Scoring Network where they are scored anonymously by multiple, highly trained raters who must pass daily calibration tests and constant quality control checks.

To learn more about TOEFL research and test quality, visit www.ets.org/toefl/research.