

Table of Contents

LTRC 2009 Conference Organizers and Committee Members	2
LTRC 2009 Conference Sponsors	8
Message from the ILTA President	9
About ILTA	12
Hotel Map	14
LTRC 2009 Program Overview	15
LTRC 2009 Full Program	19
Samuel J. Messick Memorial Lecture	30
Abstracts	
Workshops	31
Symposia	33
Papers	41
Posters	65
Works-in-Progress	81
Index of Presenters	98

LTRC 2009 Conference Organizers and Committee Members

Organizing Committee

Margaret Malone (co-chair), *Center for Applied Linguistics*

Sara Cushing Weigle (co-chair), *Georgia State University*

Liz Hamp-Lyons, *University of Bedfordshire*

Craig Deville, *Measurement Inc.*

Micheline Chalhoub-Deville, *University of North Carolina at Greensboro*

Rachel Brooks, *Federal Bureau of Investigation*

Jamie Schissel, *University of Pennsylvania*

Program

Cover Design: Cambridge ESOL

Program Assistance:

Krisanne Post, *Georgia State University*

Heather Conlin, *Georgia State University*

Luke Amoroso, *Center for Applied Linguistics/Georgetown University*

Yen-Tsu Chang, *Center for Applied Linguistics/Georgetown University*

Round-Up Planning Committee

Megan Montee, *Center for Applied Linguistics*

Anne Donovan, *Center for Applied Linguistics*

Jessica Hoover, *Center for Applied Linguistics*

Victoria Nier, *Center for Applied Linguistics*

Proposal Reviewers

Lyle Bachman	UCLA	lfb@humnet.ucla.edu
Alison Bailey	UCLA	abailey@gseis.ucla.edu
Jayanti Banerjee	University of Michigan	banerjej@umich.edu
Vivien Berry	University of Hong Kong	vberry@hkucc.hku.hk
Annie Brown	Ministry of Higher Education, UAE	annieb@napo.ae
Nathan Carr	California State University, Fullerton	ncarr@fullerton.edu
Martyn Clark	Center for Applied Second Language Studies	martyn@uoregon.edu

Christine Coombe	<i>Dubai Men's College</i>	christine.coombe@hct.ac.ae
Allister Cumming	<i>OISE</i>	acumming@oise.utoronto.ca
Alan Davies	<i>University of Edinburgh</i>	a.davies@ed.ac.uk
John H.A.L. De Jong	<i>Pearson</i>	john.dejong@pearson.com
Barbara Dobson	<i>University of Michigan</i>	bdobson@umich.edu
Dan Douglas	<i>Iowa State University</i>	dandoug@iastate.edu
Liz Hamp-Lyons	<i>Hong Kong University</i>	lizhl@hkucc.hku.hk
Dorry Kenyon	<i>Center for Applied Linguistics</i>	dkenyon@cal.org
Ute Knoch	<i>University of Melbourne</i>	u.knoch@auckland.ac.nz
Lorena Llosa	<i>New York University</i>	lorena.llosa@nyu.edu
Sari Luoma	<i>Ballard-Tighe</i>	sluoma@ballard-tighe.com
Tim McNamara	<i>University of Melbourne</i>	tmcna@unimelb.edu.au
Adrian Palmer	<i>University of Utah</i>	apalmer@linguistics.utah.edu
Spiros Papageorgiou	<i>University of Michigan</i>	spapag@umich.edu
India Plough	<i>Teachers College, Columbia</i>	indiac@umich.edu
James Purpura	<i>University of Cambridge</i>	jp248@columbia.edu
Nick Saville	<i>ESOL Examinations</i>	saville.n@cambridgeesol.org
Lynda Taylor	<i>Cambridge ESOL</i>	taylor.l@cambridgeesol.org
Randy Thrasher	<i>Okinawa Christian University</i>	thrasher@ocjc.ac.jp
Carolyn Turner	<i>McGill University</i>	carolyn.turner@mcgill.ca
Elvis Wagner	<i>Temple University</i>	elviswag@temple.edu
Yoshinori Watanabe	<i>Sophia University</i>	yoshin-w@hoffman.cc.sophia.ac.jp
Gillian Wigglesworth	<i>University of Melbourne</i>	gillianw@unimelb.edu.au

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)

Award Committees

2007 ILTA Best Article Award Committee

Constant Leung (chair), *King's College London*

Vivien Berry, *Independent Language Testing Consultant*

Annie Brown, *Ministry of Education, United Arab Emirates*

Johannes Eckerth, *King's College London*

2009 Sage/ILTA Book Award Committee

April Ginther (chair), *Purdue University, USA*

Dan Douglas, *Iowa State University, USA*

Annie Brown, *Ministry of Higher Education, UAE*

Pauline Rea-Dickins, *University of Bristol, UK*

Liz Hamp-Lyons, *University of Bedfordshire, UK*

2009 Selection Committee for the Robert Lado Award for Best Graduate Student Paper at LTRC

John Read (chair), *University of Auckland*

Yasuyo Sawaki, *Educational Testing Service*

May Tan, *McGill University*

2009 IELTS Masters Award Committee

The IELTS Joint Research Committee chaired by Nick Saville

2009 Jacqueline A. Ross TOEFL Dissertation Award Committee

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

The 2009 Selection Committee for the Spaan Fellowship for Studies in Second or Foreign Language Assessment

Jeff S. Johnson (chair), *University of Michigan*

Barbara Dobson, *University of Michigan*

India Plough, *University of Michigan*

Eric Lagergren, *University of Michigan*

Natalie N. Chen, *University of Michigan*

2008 ILTA Grant Funding for Workshops and Meetings

Rob Schoonen (chair), *University of Amsterdam*

Craig Deville, *Measurement Incorporated*

Jim Purpura, *Teachers College, Columbia University*

Award Winners

2007 ILTA Best Article Award

Marilyn Abbott, Alberta Education

A confirmatory approach to differential item functioning on an ESL reading assessment

2009 Sage/ILTA Book Award

Tim McNamara and Carsten Roever, University of Melbourne

Language testing: The social dimension

2009 IELTS Masters Award

Susan Clarke, Macquarie University

Thesis Advisor: *John Knox*

Thesis Title: *Investigating interlocutor input and candidate response on the IELTS Speaking test: A Systemic Functional Linguistics approach*

2009 Jacqueline A. Ross TOEFL Dissertation Award winners and affiliations

Spiros Papageorgiou, Lancaster University

Dissertation Advisor: *Dr. Charles Alderson*

Dissertation Title: *Setting standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*

2009 Spaan Fellowship for Studies in Second or Foreign Language Assessment

Christine Goh and Seyed Vahid Arydoust, Nanyang Technological University, Singapore

Investigating the Construct Validity of MELAB Listening Test through the Rasch Analysis and Correlated Uniqueness Modeling

Kornwipa Poonpon, Northern Arizona University

Expanding a Second Language Speaking Rating Scale for Instructional and Assessment Purposes

Hyun Jung Kim, Teachers College, Columbia University

Investigating the Construct Validity of a Semi-direct Speaking Test: A Structural Equation Modeling Analysis

Gad Lim, Ateneo de Manila University, the Philippines, and University of Michigan

Topic and Rater Effects in Second Language Writing Performance Assessment

2009 ILTA Student Travel Award

Hongwen Cai, University of California, Los Angeles

Clustering to inform standard setting in an oral test for EFL learners

Yao Hill, University of Hawai'i

DIF investigation in TOEFL iBT reading comprehension: Interaction between content knowledge and language proficiency

Talia Isaacs, McGill University

Judgments of L2 comprehensibility, accentedness and fluency: The listeners' perspective

Jiyoon Lee, University of Pennsylvania

Analysis of test-takers' performance under test-interlocutors' influences in paired speaking assessment

Yi-Ching Pan, National Pingtung Institute of Commerce, The University of Melbourne

Washback from English certification exit requirements: A conflict between teaching and learning

Anja Roemhild, University of Nebraska

Assessing domain-general and domain-specific academic English language proficiency

Jie (Jennifer) Zhang, Shanghai University of Finance and Economics/Guangdong University of Foreign Studies

Exploring rating process and rater belief: Transparentizing rater variability

LTRC 2009 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

Center for Applied Linguistics

Educational Testing Service

Elsevier Publishing

Pearson Language Assessments

Second Language Testing International

STEP EIKEN

Taylor & Francis Group

The University of Michigan English Language Institute

Message from the ILTA President

Dear Friends and Colleagues,

How exciting it is that we are once again meeting to share in our common interest of language testing and assessment. On behalf of the ILTA Executive Board, the LTRC Organizing Committee and all the members of the International Language Testing Association, I would like to welcome you to the 31st Annual Language Testing Research Colloquium in Denver, Colorado, USA.

This year's conference theme is "*Reflecting on 30 Years: Learning, Transparency, Responsibility and Collaboration.*" In response to the Call for Papers, 195 proposals were submitted for review. In the end, 23 papers, 21 posters, 20 works-in-progress and 3 symposia were selected. In addition, LTRC has 3 pre-conference workshops: assessing listening comprehension, hierarchical linear modeling and standard setting. A new feature this year is the Newcomer's Panel right before the Welcoming Reception. This idea evolved from first-time conference goers as they tried to navigate their first LTRC, as well as from those experienced in organizing LTRCs.

As is tradition, we find it important to acknowledge those in our field who have demonstrated distinguished scholarship. This year several such people will be recognized and presented with awards at the LTRC Banquet ("Round Up" in Wild West terms). We will honor the winners of the following awards: 2007 ILTA Best Article Award, 2009 IELTS Masters Award, 2009 Jacqueline A. Ross TOEFL Dissertation Award and 2009 Spaan Fellowship for Studies in Second or Foreign Language Assessment. In addition, we have a new award that will be presented for the first time; the 2009 SAGE/ILTA Book Award. The winner of the final award, the 2009 Robert Lado Memorial Award, is selected during the conference. It is given to the best student paper presented at LTRC. So don't miss the Banquet where we can personally honor and congratulate these people.

I would like to put in a special word to students. You are our community's future. Many of us started out as student LTRC attendees, volunteers and gradually presenters of posters, works-in-progress and papers. ILTA has and continues to encourage students to get involved – to be part of LTRC, to present research, and to join ILTA. Several of the awards mentioned above concern students. In addition, this year ILTA is happy to have supported several students through the ILTA Student Travel Award.

This year LTRC is back-to-back with the AAAL Conference (American Association of Applied Linguistics). This is not the first time this has happened. In fact this time, we are sharing the same venue. The evening between the two

conferences (Friday, March 20) there will be a joint activity, the AAAL/LTRC Beer Tasting event. Also during AAAL, there will be various presentations and colloquia focusing on language testing and assessment. We want to nurture our relationship with AAAL and work together for more joint activities and symposia in the future.

We are pleased to announce that LTRC 2010 will be in Cambridge, UK. Come to the ILTA meeting on Thursday at noon to learn more about the details of LTRC 2010, 2011 and 2012.

This is an exciting time for ILTA. The present Executive Board has been in place only since January, but the past year has shown lots of activity. We were able to launch our newly designed interactive website. A big thank you to Dan Douglas and Erik Voss. One of the many things this enables us to do is to be more transparent in posting all sorts of information including archival documentation on both ILTA and LTRC. We still have lots to do, but we're encouraged to have it up and running. This past year, we were also able to re-activate ILTA's Grant Funding for Workshops and Meetings in areas of need throughout the world. Much of the Executive Board's time was spent on continuing to develop a set of Standard Operating Procedures (SOPs) for ILTA. As our association grows, there is the need to formally document how we function. In this same spirit of growth, ILTA is now beginning its third year with Prime Management, our managing company. In a positive light, we have become large enough to need this.

We have come a long way from functioning on the goodwill of a small number of volunteers, but yet it is still those volunteers that maintain the professional values of ILTA. On behalf of ILTA, I would personally like to thank all the volunteers who have been chairs of and served on the selection committees mentioned above, as well as on many other service committees. The Executive Board has and continues to put in hours of work on ILTA issues. I thank them for their endless work.

It is my pleasure to now recognize the tremendous amount of time and effort put in by this year's LTRC Organizing Committee. The co-chairs, Margaret Malone and Sara Weigle, have endless energy and have managed this multifaceted task admirably. Much thanks also goes to the members of the Organizing Committee: Liz Hamp-Lyons, Craig Deville, Micheline Chalhoub-Deville, Rachel Brooks and Jamie Schissel. A sincere word of gratitude to Xiaoming Xi, the ILTA Secretary and special liaison to LTRC 2009 for her contribution to the conference.

Please join me in thanking all of these people and all of those volunteers not mentioned by name. Due to their work, we can look forward to an exciting and stimulating conference on many fronts (e.g., academically, intellectually,

culturally and socially) as we interact and discuss and debate language testing and assessment issues.

All the best for a great LTRC in Denver!

A handwritten signature in cursive script that reads "Carolyn E. Turner". The ink is dark and the handwriting is fluid and legible.

Carolyn E. Turner
President of ILTA, 2009

About ILTA

ILTA Goals

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

ILTA 2009 Executive Board

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

Members-at-Large 2009:

Gary Buck, *University of Michigan, USA*
Barbara Dobson, *University of Michigan, USA*
Yong-Won Lee, *Seoul National University, South Korea*
Rob Schoonen, *University of Amsterdam, NL*

ILTA Newsletter

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

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

ILTA Task Force on Testing Standards Update

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*

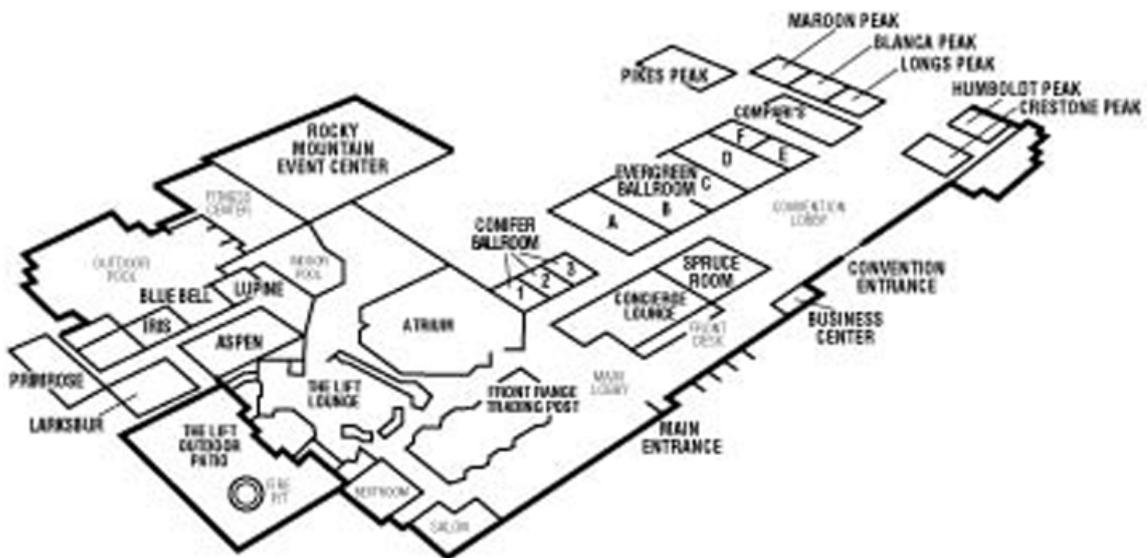
ILTA Institutional Affiliates

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	2001	Caroline Clapham
1993-4	Charles Alderson	2002	Fred Davidson
1995	Bernard Spolsky	2004	Antony Kunnan
1996	Lyle Bachman	2003	Liz Hamp-Lyons
1997	John Clark	2005	Dan Douglas
1998	Tim McNamara	2006	Glenn Fulcher
1999	Elana Shohamy	2007-8	James Purpura
2000	Alan Davies	2009	Carolyn Turner

Marriott Denver Tech Center Hotel Map



PROGRAM OVERVIEW

LTRC 2009 Program Overview

MONDAY, MARCH 16		
8:00 – 5:00	Registration	Convention Foyer
9:00 – 4:00	Workshop 1: Assessing Listening Comprehension	Larkspur
9:00 – 4:00	Workshop 2: Hierarchical Linear Modeling	Aspen Amphitheater
TUESDAY, MARCH 17		
8:00 – 7:00	Registration	Convention Foyer
9:00 – 4:00	Workshop 1: Assessing Listening Comprehension	Larkspur
9:00 – 4:00	Workshop 3: Standard Setting	Aspen Amphitheater
4:00 – 4:30	BREAK	
4:30 – 6:00	Newcomers Panel	Evergreen CD
6:00	Reception	Atrium
WEDNESDAY, MARCH 18		
7:30 – 5:30	Registration	Convention Foyer
8:00 – 6:00	Publishers' Exhibits	Evergreen Foyer
8:30 – 8:45	Welcome to LTRC 2009	Evergreen Ballroom
8:45 – 8:50	In memoriam: Donna Ilyin	Evergreen Ballroom
8:50 – 9:00	Introduction to Messick Memorial Lecture	Evergreen Ballroom
9:00 – 10:00	Messick Memorial Lecture <i>Lorrie Shepard</i>	Evergreen Ballroom
10:00 – 10:05	Presentation of Messick Award	Evergreen Ballroom

PROGRAM OVERVIEW

10:05– 10:15	BREAK	
10:15 – 11:45	<p>Paper Session 1</p> <p><i>The social impact of English certification exit requirements</i></p> <p><i>Addressing transparency and accountability through a strong program of validity inquiry: The Malaysian public service experience</i></p> <p><i>The role of quantitative and qualitative methodologies in the development of rating scales for speaking</i></p>	Evergreen Ballroom
11:45 – 1:15	<p>Lunch</p> <p>LAQ Editorial Board Meeting</p>	Larkspur
1:15 – 2:45	Poster Session	Atrium
2:45 – 4:15	<p>Paper Session 2</p> <p><i>Clustering to inform standard setting in an oral test for EFL learners</i></p> <p><i>Investigating the effectiveness of individualized feedback to rating behavior on a longitudinal study</i></p> <p><i>Exploring rating process and rater belief: Transparentizing rater variability</i></p>	Evergreen Ballroom
4:15 – 6:30	Symposium 1: Investigating the impact of assessment for migration purposes	Evergreen Ballroom
7:00 - 10:00	ILTA Executive Board Meeting	Larkspur
THURSDAY, MARCH 19		
7:30 – 5:30	Registration	Convention Foyer
8:00 – 6:00	Publishers' Exhibits	Evergreen Foyer
8:30 – 8:45	Announcements	Evergreen Ballroom
8:45 – 10:15	<p>Paper Session 3</p> <p><i>From the periphery to the centre in Applied Linguistics: the case for situated language assessment</i></p> <p><i>The extent to which differences in L2 group oral test scores can be attributed to different rater perceptions or different test taker performance</i></p>	Evergreen Ballroom

PROGRAM OVERVIEW

	<i>The analysis of test takers' performances under their test-interlocutor influence in a paired speaking assessment</i>	
10:15 – 10:30	BREAK	
10:30 – 12:00	<p>Paper Session 4</p> <p><i>Towards a transparent construct of reading-to-write assessment tasks: The interface between discourse features and proficiency</i></p> <p><i>DIF investigation of TOEFL iBT reading comprehension: Interaction between content knowledge and language proficiency</i></p> <p><i>Completion as an assessment tool of L2 reading comprehension: Building a validity argument</i></p>	Evergreen Ballroom
12:00 – 2:00	<p>Lunch</p> <p>ILTA Membership Meeting</p>	Larkspur
2:00 – 3:30	Works in Progress	Evergreen A & B
3:30 – 5:30	Symposium 2: The use of integrated reading/writing tasks: international, institutional and instructional perspectives	Evergreen Ballroom
6:30	Round-up	off site
FRIDAY, MARCH 20		
7:30 – 12:30	Registration	Convention Foyer
8:00 – 6:00	Publishers' Exhibits	Evergreen Foyer
8:30 – 8:45	Announcements	Evergreen Ballroom
8:45 – 10:15	<p>Paper Session 5</p> <p><i>Judgments of L2 comprehensibility, accentedness and fluency: The listeners' perspective</i></p> <p><i>In the ear of the beholder dependence of comprehensibility on language background of speaker and listener</i></p> <p><i>Temporal aspects of perceived speaking fluency</i></p>	Evergreen Ballroom
10:15 – 10:30	BREAK	
10:30 –	Paper Session 6	Evergreen

PROGRAM OVERVIEW

12:00	<p><i>Assessing domain-general and domain-specific academic English language proficiency</i></p> <p><i>Defining the construct of academic writing to inform the development of a diagnostic assessment</i></p> <p><i>Profiles of linguistic ability at different levels of the European framework: Can they provide transparency?</i></p>	Ballroom
12:00 – 1:30	<p>Lunch</p> <p>LT Editorial Board</p>	Larkspur
1:30 – 3:00	<p>Paper Session 7</p> <p><i>Relative impact of rater characteristics versus speaker suprasegmental features on oral proficiency scores</i></p> <p><i>A meta-analysis of multitrait-multimethod studies in language testing research: Focus on language ability and Chelle's (1998) construct definition and interpretation</i></p> <p><i>Telling our story: Reflections on the place of learning, transparency, responsibility and collaboration in the language testing narrative</i></p>	Evergreen Ballroom
3:00 – 5:00	<p>Symposium 3: The discourse of assessments: Addressing linguistic complexity in content and English language proficiency tests through linguistic analyses</p>	Evergreen Ballroom
5:00 – 5:15	<p>Wrap-up and Closing Announcements</p>	Evergreen Ballroom
5:30	<p>AAAL/LTRC Beer Tasting</p>	

FULL PROGRAM

LTRC 2009 Full Program

WORKSHOPS

Monday, March 16

9:00 am – 4:00 pm

Workshop 1: Assessing Listening Comprehension (Day 1)

Location: Larkspur

Gary Buck, *University of Michigan*

Jayanti Banerjee, *University of Michigan*

Natalie Nordby Chen, *University of Michigan*

Workshop 2: Hierarchical Linear Modeling

Location: Aspen Amphitheater

Jonathan Templin, *University of Georgia*

WORKSHOPS

Tuesday, March 17

9:00 am – 4:00 pm

Workshop 1: Assessing Listening Comprehension (Day 2)

Location: Larkspur

Gary Buck, *University of Michigan*

Jayanti Banerjee, *University of Michigan*

Natalie Nordby Chen, *University of Michigan*

Workshop 3: Standard Setting

Location: Aspen Amphitheater

Michael B. Bunch, *University of Georgia*

4:30 – 6:00

Newcomers Panel

Evergreen B & C

6:00 – 8:00

Opening Reception

Atrium

FULL PROGRAM

Wednesday, March 18, 2009

ALL SESSIONS IN EVERGREEN BALLROOM UNLESS OTHERWISE NOTED

8:30 – 8:45 **Welcome to LTRC**

8:45 – 8:50 **In memoriam: Donna Ilyin**
Charles Stansfield

8:45 – 9:00 **Introduction to Messick Memorial Lecture**
Craig DeVille

9:00 – 10:00 **Samuel J. Messick Memorial Lecture: Understanding learning (and teaching) progressions as a framework for language testing**
Lorrie Shepard
(Sponsored by the TOEFL Board, Educational Testing Service, Princeton, New Jersey)

10:00 – 10:05 **Presentation of Messick Award**

10:05– 10:15 **BREAK**

10:15 – 11:45 **Paper Session 1**

The social impact of English certification exit requirements

Yi-Ching Pan, *National Pingtung Institute of Commerce, The University of Melbourne*

Addressing transparency and accountability through a strong program of validity inquiry: The Malaysian public service experience

Kedeesa Abdul Kadir, *University of Illinois at Urbana-Champaign*

The role of quantitative and qualitative methodologies in the development of rating scales for speaking

Evelina Galaczi, *University of Cambridge ESOL Examinations*

11:45 – 1:15 **LUNCH**

LAQ Editorial Board Meeting: Larkspur Room

FULL PROGRAM

1:15 – 2:45 **Poster Session, Atrium**

Development and validation of a web-based multimedia Korean oral proficiency test

Seongmee Ahn, *Michigan State University*

Ok-Sook Park, *Michigan State University*

Daniel Reed, *Michigan State University*

The development of an English proficiency test for teacher certification in Quebec

Beverly Baker, *McGill University*

Avril Aitken, *Bishop's University, Quebec, Canada*

Anne Hetherington, *Concordia University, Quebec, Canada*

The development of an on-line training and marking program for the assessment of writing

Annie Brown, *Ministry of Higher Education and Scientific Research*

Issues in developing a proficiency-based assessment in multiple languages

Martyn Clark, *Center for Applied Second Language Studies*

Construction of a proficiency identity within paired oral assessment: Insights from discourse

Larry Davis, *University of Hawai'i at Manoa*; Anne Lazaraton, *University of Minnesota*

Creating valid in-house can-do descriptors for a listening course

Tomoko Fujita, *Tokai University*

Examining collaboration in oral proficiency interviews

Gene Halleck, *Oklahoma State University*

Integrated speaking assessment: Anxiety-reducing or anxiety-provoking?

Heng-Tsung Danny Huang, *University of Texas at Austin*

Developing a placement test for academic Arabic for upper high school and university students

Summer Loomis, *The University of Texas at Austin*

The distance between student writers' intentions of expression and teacher raters' expectations in argumentative essay rating

Lu Lu, *The University of Hong Kong*

FULL PROGRAM

Assessing the language proficiency of Chinese nursing students and professionals:

A brief introduction to METS (for Nurses)

Kaizhou Luo, *Binhai College of Nankai University, METS Office*

Ting Huang, *METS Office*

Mariam Jones, *American Academic Alliance*

Predicting item difficulty: A rubrics-based approach

David MacGregor, *Center for Applied Linguistics*

An analysis of the sentence span task for L2 learners

Toshihiko Shiotsu, *Kurume University*

Effect of guidelines during a consensus discussion in standard-to-standard alignment

Emily Svendsen, *University of Illinois at Urbana-Champaign*

Jiin Yap, *University of Illinois at Urbana-Champaign*

Selling a language assessment to skeptical stakeholders: The GSLPA

Alan Urmston, *Hong Kong Polytechnic University*

Felicia Fang, *Hong Kong Polytechnic University*

Testifying in legal cases: Test your knowledge, flexibility, and creativity!

Margaret van Naerssen, *Immaculata University*

A test to reveal incremental acquisition of vocabulary knowledge

JoDee Walters, *Bilkent University*

Response strategies and performance on an integrated writing test

Hui-chun Yang, *University of Texas Austin*

The effect of different anchor tests on equating quality

Kiyomi Yoshizawa, *Kansai University*

A retrospection study on the construct validity of TOEFL listening comprehension tests with multiple-choice format

Yujing Zheng, *Chongqing Jiaotong University*

Xiangdong Gu, *Chongqing Jiaotong University*

2:45 – 4:15 Paper Session 2

Clustering to inform standard setting in an oral test for EFL learners

Hongwen Cai, *University of California, Los Angeles*

FULL PROGRAM

Investigating the effectiveness of individualized feedback to rating behavior on a longitudinal study

Ute Knoch, *University of Melbourne*

Exploring rating process and rater belief: Transparentizing rater variability

Jie Zhang, *Shanghai University of Finance and Economics/Guangdong University of Foreign Studies*

4:15 – 6:30 **Symposium 1: Investigating the impact of assessment for migration purposes**

Elana Shohamy and Nick Saville, Organizers

Piet Van Avermaet, *Centre for Intercultural Education, University of Ghent* and Max Spotti, *Tilburg University*, "Tests or no tests for integration? Views of the stakeholders"

Nick Saville and Szilvia Papp, *University of Cambridge, ESOL Examinations*, "ESOL skills for life: a micro-level impact study of the tests"

Lorenzo Rocca and Giuliana Grego Bolli, *Università per Stranieri, Perugia*, "The impact of CELI exams for immigrants"

Elana Shohamy, Tzahi Kenza and Nathalie Assias, *Tel Aviv University*, "Language and civil participation: Different requirements for different groups"

Discussant: Tim McNamara

7:00 - 10:00 **ILTA Executive Board Meeting**

FULL PROGRAM

Thursday, March 19, 2009

ALL SESSIONS IN EVERGREEN BALLROOM UNLESS OTHERWISE NOTED

8:30 – 8:45 **Announcements**

8:45 – 10:15 **Paper Session 3**

From the periphery to the centre in Applied Linguistics: the case for situated language assessment

Pauline Rea-Dickins, *University of Bristol*

Matt Poehner, *Penn State University*

Constant Leung, *King's College London*

Lynda Taylor, *University of Cambridge ESOL Examinations*

Elana Shohamy, *Tel Aviv University*

The extent to which differences in L2 group oral test scores can be attributed to different rater perceptions or different test taker performance

Gary Ockey, *Utah State University*

The analysis of test takers' performances under their test-interlocutor influence in a paired speaking assessment

Jiyoon Lee, *University of Pennsylvania*

10:15– 10:30 **BREAK**

10:30 – 12:00 **Paper Session 4**

Towards a transparent construct of reading-to-write assessment tasks: The interface between discourse features and proficiency

Atta Gebriel, *The United Arab Emirates University*

Lia Plakans, *The University of Texas, Austin*

DIF investigation of TOEFL iBT Reading Comprehension: Interaction between content knowledge and language proficiency

Yao Hill, *University of Hawai'i*

Liu, Ou Lydia, *Educational Testing Service*

Completion as an assessment tool of L2 reading comprehension: Building a validity argument

William Grabe, *Northern Arizona University*

Xiangying Jiang, *Northern Arizona University*

12:00 – 2:00 **LUNCH**

LTRC Meeting

ILTA Membership Meeting, Larkspur

FULL PROGRAM

2:00 – 3:30 **Works-in-Progress, Evergreen A & B**

Native versus non-native raters' evaluations of high-level English

Rachel Brooks, *Federal Bureau of Investigation*

Construct validation of listening comprehension test: Effects of task and the relationship between listeners' cognitive awareness and proficiency in listening comprehension

Youngshin Chi, *University of Illinois at Urbana-Champaign*

Bridging the gap: Assessment for learning in a Quebec classroom

Christian Colby-Kelly, *McGill University*

Oral proficiency assessment in a pre-service EFL teacher education programme: Transparency on the validation of vocabulary descriptors

Douglas Altamiro Consolo, *UNESP*

Melissa Baffi-Bonvino, *UNESP*

Factors influencing the pragmatic development Korean study-abroad learners

Sumi Han, *Seoul National University*

Learning outcomes and the focus of the assessment tool

Ching-Ni Hsieh, *Michigan State University*

Weiping Wu, *The Chinese University of Hong Kong*

Developing an oral assessment protocol and rating rubric for applicants to the English for Heritage Language Speakers program

Genesis Ingersoll, *Center for Applied Linguistics*

Anne Donovan, *Center for Applied Linguistics*

Natalia Jacobsen, *Center for Applied Linguistics/Georgetown University*

Computer-based and internet-delivered College English Test in China: IB-CET in progress

Guodong Jia, *Renmin University of China*

Reform from within: A collaborative effort at transparency, reliability and validity in assessment

Claudia Kunschak, *Shantou University*

ESL writing assessment: Does the selection of rating scale matter?

Chih-Kai (Cary) Lin, *Georgetown University*

The Internet-Based TOEFL and test user beliefs

Margaret E. Malone, *Center for Applied Linguistics*

Megan Montee, *Center for Applied Linguistics*

FULL PROGRAM

Developing new lexical measures for diagnostic assessment

John Read, *University of Auckland*

Toshihiko Shiotsu, *Kurume University*

Relationship Among the test, curriculum, and teacher content representations in an EFL setting

Sultan Turkan, *University of Arizona*

Vocabulary knowledge and its use in EFL speaking and writing test performance

Viphavee Vongpumivitch, *National Tsing Hua University*

Video discourse completion tasks for the testing of L2 pragmatic competence

Elvis Wagner, *Temple University*

Tina Hu, *Temple University*

Using integrative task-based assessment to examine the effectiveness of task-based language teaching

Jing Wei, *UMCP*

Cheng-Chiang (Julian) Chen, *UMCP*

Applying protocol analysis in analyzing language test validity: A case study

Huijie Xu, *Zhejiang University of Technology*

Ying Zheng, *Queen's University*

Heying Lou, *Zhejiang University of Technology*

The relationship of TOEFL scores to success in American universities: How high is high enough?

Yeonsuk Cho, *Educational Testing Service*

Brent Bridgeman, *Educational Testing Service*

3:30 – 5:30 Symposium 2 The use of integrated reading/writing tasks: international, institutional and instructional perspectives

Guoxing Yu and Yasuyo Sawaki, Organizers

Yasuyo Sawaki & Thomas Quinlan, *Educational Testing Service, USA* and Yong-Won Lee, *Seoul National University, Korea*, "The value of integrated writing task measures for understanding learner strengths and weaknesses"

Sara Cushing Weigle & WeiWei Yang, *Georgia State University, USA*, "Validation and implementation of an integrated reading and writing test"

Mark Wolfersberger, *Brigham Young University - Hawaii, USA*, "Second language writing from sources: An ethnographic study of an argument essay task"

FULL PROGRAM

Guoxing Yu, *University of Bristol, UK*, "The use of summarization tasks as a measure of reading comprehension: premise, promise, problems and compromises"

Discussant: Alister Cumming, *OISE, University of Toronto, Canada*

6:30 **Round-up; Tivoli Student Union, University of Colorado, Denver**

FULL PROGRAM

Friday, March 20, 2009

ALL SESSIONS IN EVERGREEN BALLROOM UNLESS OTHERWISE NOTED

8:30 – 8:45 **Announcements**

8:45 – 10:15 **Paper Session 5**

Judgments of L2 Comprehensibility, accentedness and fluency: The listeners' perspective

Talia Isaacs, *McGill University*

Ron Thomson, *Brock University*

In the ear of the beholder Dependence of comprehensibility on language background of speaker and listener

Alistair Van Moere, *Pearson*

Ryan Downey, *Pearson*

Temporal aspects of perceived speaking fluency

Nivja DeJong, *University of Amsterdam*

10:15– 10:30 **BREAK**

10:30 – 12:00 **Paper Session 6**

Assessing domain-general and domain-specific academic English language proficiency

Anja Römhild, *University of Nebraska*

Dorry Kenyon, *Center for Applied Linguistics*

David MacGregor, *Center for Applied Linguistics*

Defining the construct of academic writing to inform the development of a diagnostic assessment

Lorena Llosa, *New York University*

Sara W. Beck, *New York University*

Cecilia G. Zhao, *New York University*

Profiles of linguistic ability at different levels of the European Framework: Can they provide transparency?

Rob Schoonen, *University of Amsterdam*

Nivya DeJong, *University of Amsterdam*

Margarita Steinel, *University of Amsterdam*

Arjen Florijn, *University of Amsterdam*

Jan Hulstijn, *University of Amsterdam*

12:00 – 1:30 **LUNCH**

FULL PROGRAM

1:30 – 3:00 Paper Session 7

Relative impact of rater characteristics versus speaker suprasegmental features on oral proficiency scores

Don Rubin, *University of Georgia*

Okim Kang, *Northern Arizona University*

Lucy Pickering, *Georgia State University*

A meta-analysis of multitrait-multimethod studies in language testing research: Focus on language ability and Chelle's (1998) construct definition and interpretation

Yo In'nami, *Toyohashi University of Technology*

Rie Koizumi, *Tokiwa University, Japan*

Telling our story: Reflections on the place of learning, transparency, responsibility and collaboration in the language testing narrative

Lynda Taylor, *University of Cambridge ESOL Examinations*

3:00 – 5:00 Symposium 3

The discourse of assessments: Addressing linguistic complexity in content and English language proficiency tests through linguistic analyses

Jim Bauman, *Center for Applied Linguistics*

Laura Wright, *Center for Applied Linguistics*

David MacGregor, *Center for Applied Linguistics*

Abbe Spokane, *Center for Applied Linguistics*

Meg Montee, *Center for Applied Linguistics*

5:30 AAAL/LTRC Beer Tasting

MESSICK LECTURE

Samuel J. Messick Memorial Lecture

Lorrie A. Shepard

University of Colorado at Boulder

Understanding Learning (and Teaching) Progressions as a Framework for Language Testing

Abstract

Some of the earliest learning progressions were developed by scholars in the area of emergent literacy. These progressions were documented from a developmental perspective and illustrated how children "naturally" gained increasing proficiency in writing, spelling, and beginning reading. While considerable variability was acknowledged, early literacy continua did not explicitly build in consideration of the influence of curricula or learning opportunities. In recent years progress maps and learning trajectories have become powerful tools to conceptually link curriculum, instruction, and assessment and to ensure coherence between classroom level and large-scale assessments. Advances in the development of learning progressions have occurred in a minor way with the assessment of writing, but have primarily occurred in mathematics and science. Research on learning progressions could foster significant improvements in language testing if researchers are attentive to important distinctions among monolingual language development, second-language acquisition, native-language language arts, and foreign-language learning, and if learning continua are understood to be entwined with, and dependent upon, specific curricula and teaching trajectories.

Biography

Lorrie Shepard is professor of education and chair of the Research and Evaluation Methodology program area. Dr. Shepard is currently dean of the School of Education. Her research focuses on psychometrics and the use and misuse of tests in educational settings. Technical topics include validity theory, standard setting, and statistical models for detecting test bias. Her studies evaluating test use include identification of learning disabilities, readiness screening for kindergarten, grade retention, teacher testing, effects of high-stakes testing, and classroom assessment. At the graduate level, Dr. Shepard teaches courses in statistics, research methods, and testing and assessment policy. In the teacher education program, she teaches assessment in collaboration with colleagues in content methods courses.

Dr. Shepard is a past president of the American Educational Research Association and past president of the National Council on Measurement in Education. She was elected to the National Academy of Education in 1992 and served as Vice President of the NAE. She has been editor of the *Journal of Educational Measurement* and the *American Educational Research Journal* and interim editor of *Educational Researcher*. In 1999 she won NCME's Award for Career Contributions to Educational Measurement. Dr. Shepard currently serves on the National Research Council's Board on Testing and Assessment.

WORKSHOPS

Workshop 1: Assessing Listening Comprehension

Gary Buck, *University of Michigan*

Jayanti Banerjee, *University of Michigan*

Natalie Nordby Chen, *University of Michigan*

Gary Buck is an experienced educator and assessment specialist, and is currently the Director of the Testing and Certification Division of the English Language Institute at the University of Michigan. He wrote his dissertation on *The Testing of Second Language Listening Comprehension*, in 2001 he published *Assessing Listening Comprehension*, with Cambridge University Press, and has given numerous workshops and presentations on testing listening comprehension.

Jayanti Banerjee is a language assessment specialist in the Testing and Certification Division of the English Language Institute at the University of Michigan with primary responsibility for the Examination for the Certificate of Proficiency in English (ECPE). She previously worked at Lancaster University where she coordinated their Hong Kong MA TESOL master's degree, taught on various master's degree and PhD courses, and supervised master's degree and PhD projects.

Natalie Nordby Chen is Test Development Manager for the University of Michigan's English Language Institute. An experienced educator and assessment specialist, she specializes in English language assessments and has worked on the design and development of a number of new examinations ranging from Pre-K language to college admissions to professional licensing tests.

Workshop 2: Hierarchical Linear Modeling

Jonathan Templin, *University of Georgia*

Jonathan Templin is an Assistant Professor in Research, Evaluation, Measurement, and Statistics program of the Department of Educational Psychology at The University of Georgia. He obtained his Ph.D. in Quantitative Psychology at the University of Illinois at Urbana-Champaign. Dr. Templin has extensive training in the development and application of multilevel and latent variable models for educational and psychological measurement. His research has been published in *Applied Psychological Measurement*, *Educational and Psychological Measurement*, *Journal of Educational Measurement*, *Psychological Method*, and *Psychometrika*.

WORKSHOPS

Workshop 3: Standard Setting

Michael B. Bunch, *University of Georgia*

Michael B. Bunch is Senior Vice-President of Measurement Incorporated, a test development and scoring company serving the statewide assessment community. An internationally recognized expert in standard setting, Dr. Bunch is co-author, with Dr. Gregory Cizek, of *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests* (Sage, 2007). He received his Ph.D. in psychological measurement from the University of Georgia and currently serves on the Graduate Education Advancement Board of that university.

SYMPOSIA ABSTRACTS

Symposium 1: Investigating the impact of assessment for migration purposes

Time: Wednesday, March 18, 5:00 – 7:00 p.m.

Nick Saville and Elana Shohamy, Organizers

Piet Van Avermaet, *Centre for Intercultural Education, University of Ghent* and Max Spotti, *Tilburg University*, "Tests or no tests for integration? Views of the stakeholders"

Nick Saville and Szilvia Papp, *University of Cambridge, ESOL Examinations*, "ESOL Skills for Life: a micro-level impact study of the tests"

Lorenzo Rocca and Giuliana Grego Bolli, *Universita per Stranieri, Perugia*, "The impact of CELI exams for Immigrants"

Elana Shohamy, Tzahi Kenza and Nathalie Assias, *Tel Aviv University*, "Language and civil participation: Different requirements for different groups"

Discussant: Tim McNamara

A major phenomenon in the past decade, in a growing number of countries in Europe and elsewhere, is the introduction of language assessment in the context of migration. Migrants to those countries are required to pass language tests in order to demonstrate their language proficiency in the national (or other official) language as a condition for entering the country, for long-term residency and/or for obtaining citizenship. Papers that describe this phenomenon and its manifestations, e.g., the type of tests, purposes, uses and rationale have been presented in symposia and colloquia in language testing and applied linguistics conferences in the past four years. In addition, a number of publications are appearing where this phenomena is being documented, explained and critiqued (e.g. upcoming thematic issue of *Language Assessment Quarterly*, 2009; a publication by McNamara and Shohamy, 2008; Stevenson, Mar-Molinero and Hogan-Brun, in press; Extra, Spotti and Van Avermaet, in press). Yet, while detailed descriptions of these tests and the testing policies are available, there is an apparent lack of information and knowledge about the actual uses and consequences of the tests and very little information about the impact on test takers.

All the work so far has been based on macro-level considerations and on the views of language testers about the phenomenon, but there is no information about the attitudes, views and practices from the point of view of the migrants themselves. In an effort to expand the language testing conversation with larger groups of stakeholders (the theme of 2008 LTRC) and for greater transparency, responsibility and enhanced collaboration (the theme of 2009 LTRC), this symposium will focus on the impact of language tests being taken by migrants in several countries at the micro level. Thus, the four papers will report on impact-oriented research which focuses on the views and consequences of the tests as echoed by individuals in their own contexts.

SYMPOSIA ABSTRACTS

Specifically, each of the papers will report on research findings based on data collected from the perspective of those who have been tested in the language of their host country for residency and/or citizenship purposes or are about to be tested for those purposes. They address various consequential dimensions of the tests, including attitudes, local testing practices, costs, benefits and interpretations of the policy itself. Each paper builds on earlier presentations of the testing policy for migration within each of the geographical contexts, i.e. the Netherlands and Belgium, the UK, Israel, and Italy, and will focus on a specific testing system which is now being used and how it affects individual test takers, as described above.

The research methods used include questionnaires, small-scale case studies and comprehensive narrative inquiries, typically with about 10-20 immigrants per country, in which the migrant test takers reflect on the policy and its various effects and consequences in their own lives. Issues to be addressed include: fairness, preparation strategies, perceived intentions, relevance, uses of results, and unpredicted consequences. The data obtained from the test takers in each case will be aggregated and synthesized in order to develop a deeper understanding of the uses and impacts of the tests within each context.

Findings will lead to possible recommendations regarding the appropriate uses of language tests within immigration policies and a wider debate on the role that testing professionals can play in striving for fairness. Most crucially further impact research of this kind is needed to ensure that policy decisions which lead to negative impacts for individuals can be investigated and that these consequences can be mitigated in appropriate ways.

Tests or no tests for integration? Views of the stakeholders

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

Max Spotti, *Tilburg University* (m.spotti@uvt.nl)

Flanders, the Northern region of Belgium, and the Netherlands share Dutch as an official language, yet remarkable differences in assessment traditions and language policies exist between them. These differences are most obvious with regard to language testing for migration and citizenship purposes. Both countries have language requirements for immigrants and for people wanting to obtain citizenship, but the conditions in the Netherlands are stricter. In Flanders immigrants are obliged to follow a language course at A1 level, but don't have to take a language test (see also Van Avermaet & Gysen, 2009). In the Netherlands, on the other hand, immigrants must take a language test at A2-level but are not obliged to follow a language programme. A frequently articulated argument by policy makers in advocating these measures is that they enhance social cohesion and promote social participation. However, little is known about the impact of these tests and integration programmes on the migrants themselves, and whether or not the intended aims are actually achieved in practice. In this paper we report on a small-scale study addressing these and other relevant issues.

SYMPOSIA ABSTRACTS

This research was carried out making use of ethnographic methods, including face-to-face interviews with individual stakeholders: immigrants (test takers as well as non-test takers), employers, language teachers, etc; they were asked to reflect on the integration policies in Flanders and the Netherlands and on the usefulness of the courses and tests for integration purposes. The first findings, including views and perceptions of these stakeholders, will be presented and the implications discussed.

ESOL skills for life: a micro-level impact study of the tests

Nick Saville, *University of Cambridge, ESOL Examinations*

(saville.n@cambridgeesol.org)

Szilvia Papp, *University of Cambridge, ESOL Examinations*

The speakers report on an impact study focusing on test takers who have taken the Cambridge ESOL Skills for Life tests and other stakeholders in the context of applying for indefinite leave to remain (ILR) or citizenship in the UK. Extending the Cambridge ESOL model used for investigating test impact within an educational setting, we report on the results of research to gather feedback from individuals within the main stakeholders groups, especially from the test takers themselves. In addition to analyses based on routinely collected data for test validation purposes, this study reports on the analysis of questionnaire data and interviews with Skills for Life test takers. This data documents their own narratives in relation to their personal language learning and test taking experiences, as well as their views about the government policy and the qualities of the tests themselves. We argue that impact studies of this kind – those that regularly monitor context of learning (including biographical data of test takers) and context of test use (including purposes of test use, interpretation and use of results, role of tests in fulfilling original objectives) - need to be carried out in a cyclical, iterative way to feed into test development and revision, as well as to inform policy making.

The impact of CELI exams for immigrants

Giuliana Grego Bolli, *Università per Stranieri, Perugia*

Lorenzo Rocca, *Università per Stranieri, Perugia*

The CELI I is a suite of exams assessing Italian ability (A1 to B1 level of the Common European Framework of Reference), developed and produced specifically for immigrants with little schooling. The exams function by evaluating cognitive strategies relevant to learning, such as memorisation and generalisation. The purpose of the test is formative and the tests are therefore closely linked to language courses for immigrants. In order to assess the impact of CELI I, CVCL developed and administered questionnaires both to candidates and course teachers. Although data will continue to be collected, analysis of initial response has begun. This presentation will provide results of this preliminary research and discuss the impact of the exam on teaching, learning and immigrant integration.

SYMPOSIA ABSTRACTS

Language and civil participation: Different requirements for different groups

Elana Shohamy, *Tel Aviv University* (elana@post.tau.ac.il)

Tzahi Kanza, *Tel Aviv University*

Nathalie Assias, *Tel Aviv University* (nathali7@post.tau.ac.il)

The notion of 'hollow citizenship' refers to situations whereby citizenship is granted but it does not carry with it meaningful rights. This relates to the language assessment requirements which are imposed differentially to different groups, even those who are considered citizens (i.e., Israeli Arabs). In our earlier work we showed that language assessment are imposed on Jews, Jewish immigrants, non-Jewish immigrants and Arabs via covert language performances that serve as conditions for full participation in Israeli society. In this paper we will report on the applications of these language performances within the three groups: Israeli, Arabs, Ethiopian immigrants who are required to pass a language test as condition for conversion to Judaism, and non-Jews requesting citizenship. For each of the groups we will examine the specific type of language requirement, the impact and consequences of these requirements on their lives and extent to which these are viewed as a valid channel of participation or as gatekeeping devices.

Symposium 2: The use of integrated reading/writing tasks: international, institutional and instructional perspectives

Time: Thursday March 19, 3:30 – 5:30 p.m.

Guoxing Yu and Yasuyo Sawaki, Organizers

Yasuyo Sawaki & Thomas Quinlan, *Educational Testing Service, USA* and Yong-Won Lee, *Seoul National University, Korea*, "The value of integrated writing task measures for understanding learner strengths and weaknesses"

Sara Cushing Weigle, *Georgia State University, USA*, "Recurrent issues in validation and implementation of an integrated reading and writing test"

Mark Wolfersberger, *Brigham Young University - Hawaii, USA*, "Second language writing from sources: An ethnographic study of an argument essay task"

Guoxing Yu, *University of Bristol, UK*, "The use of summarization tasks as a measure of reading comprehension: premise, promise, problems and compromises"

Discussant: Alister Cumming, OISE, University of Toronto, Canada

SYMPOSIA ABSTRACTS

The revival of using integrated writing tasks in large-scale language tests has been a topic of heated debates in the field of language testing, with even a kind of love and hate demarcation. Four papers representing different contexts will be presented in order to describe, from international, institutional and instructional perspectives, the practice of using integrated writing tasks for different assessment purposes. More specifically, we will report the issues, the complexity and the values of using integrated writing tasks (a) in an international large-scale assessment context of the TOEFL® Internet-based Test (iBT), (b) in an institutional language test programme at Georgia State University, (c) in a instruction-focused writing class through an ethnographic angle, and (d) in a study aiming to develop a model of using summarization tasks as a measure of reading comprehension. Titles and presenters are listed below, together with a brief outline of each paper. At the end of the session, a discussant (Alister Cumming) will (a) provide a synthesis of the four papers to highlight the issues and the complexity of using integrated tasks for assessment purposes and (b) discuss the importance of the transparency and responsibility of the institutions (language test providers as well as universities) through the collaboration between language testers, teachers and learners themselves to develop language tests and to co-construct an environment conducive for learning.

The value of integrated writing task measures for understanding learner strengths and weaknesses

Yasuyo Sawaki, *Educational Testing Service* (ysawaki@ets.org)

Thomas Quinlan, *Educational Testing Service* (tquinlan@ets.org)

Yong-Won Lee, *Seoul National University* (ylee01@snu.ac.kr)

Integrated writing tasks require the coordination of writing with other skills such as reading, which supports their use in assessing complex literacy skills. However, integrated writing tasks have also been criticized for blurring the psychometric structure of an assessment, by essentially confounding reading skill with writing skill. However, if it were possible to disentangle reading and writing, it might be possible to extract learner profiles that inform instruction. Building on previous factor analyses of the TOEFL iBT, the present study explores the underlying factor structures across different writing measures obtained from the TOEFL iBT independent and integrated writing tasks and the Reading and Listening scores for different examinee subgroups. In the analyses, both human and automated scoring will be utilized to identify learners' strengths and weakness. A series of confirmatory factor analysis models will be tested to examine 1) whether distinct constructs can be extracted across these measures for different subgroups and 2) whether the strengths of the effects of reading, listening and writing measures on the integrated writing task are the same across subgroups. The results of the factor analyses will be discussed in terms of implications for supporting instruction.

SYMPOSIA ABSTRACTS

Recurrent issues in validation and implementation of an integrated reading and writing test

Sara Cushing Weigle, *Georgia State University* (sweigle@gsu.edu)

WeiWei Yang, *Georgia State University* (eslwyyx@langate.gsu.edu)

Integrated reading/writing tasks have been used for placement and competency testing for non-native speakers of English at Georgia State University since 2002. While there is general consensus among stakeholders that the test is effective for its stated purposes, certain issues regarding the validity and practicality of the test remain unresolved. In this paper we outline an agenda for ongoing test validation from several perspectives. First, we discuss investigations of the perceived usefulness of the test from the point of view of students, raters, teachers, and administrators. Next, we discuss issues in developing parallel forms of the test with an emphasis on institutional constraints and resource requirements for an effective test development and validation program. Next we discuss results of a pilot study investigating the cognitive processes involved in responding to the short-answer questions of the integrated reading-writing test.

Second language writing from sources: An ethnographic study of an argument essay task

Mark Wolfersberger, *Brigham Young University – Hawaii* (maw44@byuh.edu)

Writing from sources is a complex academic task that poses particular challenges for second language students and teachers. This paper will report on an ethnographic study examining the writing of four Chinese (L1) students with varying levels of English (L2) proficiency completing an argumentative writing-from sources essay task for a required writing class. Two somewhat related aspects of the participants' experiences stood out as particularly complex: 1) creating a mental representation of the writing task that was considered acceptable by the teacher and 2) plagiarism. There were a number of personal and contextual factors that influenced the creation of the participants' task representations, factors such as individual background experiences, the writing process, and information from and interactions with the teacher and other people within the writing context. These factors varied in the strength of their influence from the beginning to the end of the essay assignment. Low L2 proficiency constrained the writing performance of some of the participants. This resulted in one participant avoiding elements of the writing task requirements in order to earn passing marks and two other participants producing plagiarized texts and receiving failing marks on the assignment.

SYMPOSIA ABSTRACTS

The use of summarization tasks as a measure of reading comprehension: premise, promise, problems and compromises

Guoxing Yu, *University of Bristol*

This paper will reflect on the debates on the use of integrated reading/writing tasks with specific reference to the premises, promises, problems and certain compromises of using summarization tasks in large-scale language assessments. In an attempt to develop a model of using such tasks as a measure of reading comprehension, several issues were investigated, e.g., the development of assessment criteria, the choice of language(s) for the tasks, the summarizability and the presentation modes of source texts, the relationships between students' summarization performance and their reading abilities as measured by standardized multiple choice tests. Furthermore, this paper will report the differences in discourse features of summaries written by English native speaker experts and test takers, and the values of understanding such differences for students to develop their summarization skills for academic success.

Symposium 3: The discourse of assessments: Addressing linguistic complexity in content and English language proficiency tests through linguistic analyses

Time: Friday, March 20, 3:00 – 5:00 p.m.

Jim Bauman, *Center for Applied Linguistics* (jbauman@cal.org)

Laura Wright, *Center for Applied Linguistics* (lwright@cal.org)

David MacGregor, *Center for Applied Linguistics* (dmacgregor@cal.org)

Abbe Spokane, *Center for Applied Linguistics* (aspokane@cal.org)

Meg Montee, *Center for Applied Linguistics* (mmontee@cal.org)

Efforts over the last decade to better understand and address the requirements for valid and reliable tests of content knowledge and English language proficiency for English language learners has generated much deliberation over how to determine the language load or language complexity of test items. This deliberation has been critical in developing tests of English language proficiency since the language of the test items themselves directly expresses the construct. The typical solution here has been to base items on English language development (ELD) standards that specify levels of proficiency. In respect to content testing, however, the understanding of language complexity has centered on explicating the distinction between language relevant to and language irrelevant to the item's content. Specific methods focus on eliminating irrelevant language, recasting relevant language in simpler or more accessible ways, or in choosing alternative means of representing the construct non-linguistically. The papers in this series will address the issue of language complexity in both content tests and English language proficiency tests.

SYMPOSIA ABSTRACTS

The lead paper will ground the particular treatments in a model of linguistic representation that attempts to unify the various approaches advanced to describe and measure language complexity. This model is broadly based on the theoretical underpinnings of various discourse and cognitively based grammars, in particular the theories of Systemic Functional Linguistics (Halliday), Rhetorical Structure Theory (Mann and Thompson), and Cognitive Semantics (Talmy). The intent of the model is to theoretically unify the various positions and methods advocated by different scholars and practitioners working in the interest of producing fair and effective tests for ELLs.

The second paper will apply a functional linguistic analysis to the discourse of standardized mathematics assessments. Using released items from state and national tests, the grammatical patterns and linguistic features of items will be analyzed against the proposed model. In addition, students' scores will be examined in relationship to items' grammatical features to develop a working hypothesis about the way that grammatical complexity influences the coherence of the items and, by extension, the accessibility test takers have to them.

The third paper will use a discourse analytic approach to identify construct-relevant factors that may contribute to the empirical difficulty of writing tasks on a test of English language proficiency (ELP). Several years of data from an operational K-12 ELP test will be analyzed to determine task difficulty. Based on that analysis, construct-relevant linguistic features in the tasks that may affect difficulty will be identified. The information gleaned from the analysis is used to develop a rubric to help item developers judge how well items fit the ELD standards on which they are based.

The final paper discusses a project to gather data on the discourse of English language proficiency testing from the examinee's perspective. This paper will describe field testing procedures developed to gather qualitative information on test items, and will focus on the methods used to elicit and analyze examinee feedback on speaking and writing items. As a part of the field testing procedures, students are asked to respond to speaking and writing items. They are then interviewed individually and asked to give their perspective of the intent of the item developer and to describe any challenges presented by the item. Information gathered from this process is used to revise test items.

These papers provide a multi-faceted view of how insights from discourse-based and cognitive-based approaches to linguistics can be used to more fully understand the functionality of test items in both content tests and English language proficiency tests and, in doing so, can contribute to understandings of how to develop valid and reliable assessments.

PAPER ABSTRACTS

Session 1: Wednesday, March 18, 10:15 a.m. – 11:45 a.m.

The Social Impact of English Certification Exit Requirements

Yi-Ching Pan, *National Pingtung Institute of Commerce; The University of Melbourne* (huangpan63@yahoo.com.tw)

To motivate students to learn English in order to enhance both their English proficiency and workplace competitiveness, one third of technical universities/colleges in Taiwan have established English certification exit requirements, which require them to select from and pass English proficiency tests such as the GEPT, TOEIC, TOEFL, and IELTS to graduate.

This study investigated the social impact the exit requirement policy has had on the workplace. Many previous studies have focused on the impact of tests on teaching and learning, but very little research has explored the influence of tests on the societal dimension, which has gained more attention in the last decade from a number of researchers. For example, McNamara and Roever (2006) state that since tests can have widespread and unforeseen consequences, a language test that is psychometrically validated does not necessarily denote a test that is favorable for society.

Data was collected in two parts. First, media reports with regard to employers' opinions of tertiary English education in the last decade, when exit requirements were about to be established, were collected. Second, interviews were conducted with 19 business people who are in charge of recruiting potential employees in 17 industries across Taiwan. The number of employees at these businesses ranged from 8 to 30,000. These interviews were undertaken to discover the importance of English certification as an element of job hunting, the opinions of businesses regarding various certification tests, and their attitudes toward the exit requirements.

Findings indicate that although these employers all have a favorable assessment of this policy, only a small percentage (13%) of them require English certificates as a hiring criterion. The impact of the exit requirements on the workplace has been minimal because neither the supply of English certificates in schools nor the demand for English certificates in the workplace is significant. Another interesting finding was that a large number of employers (53%) seem to regard the certificates as evidence that applicants who possess them are diligent students and assume that they'll likely be hard-working employees, when that is not at all the point of the certificates from the testers' point of view. These employers interpreted tests very differently than testers do, where a lot of the interpretation is centered on cultural norms of what personal qualities tests highlight rather than on language ability.

Possible contributory factors are discussed to explain the weak level of test effects perceived in the workplace. Implications are also made with the goal of eliciting a stronger level of washback beneficial for the field of education, test developers, and the workplace.

PAPER ABSTRACTS

Addressing Transparency and Accountability through a Strong Program of Validity Inquiry: The Malaysian Public Service Experience

Kadessa Abdul Kadir, *University of Illinois at Urbana-Champaign*
(kadeessa@gmail.com)

Extending a language assessment developed for particular target population to a wider group having possibly dissimilar discourse needs has the potential for giving rise to issues of validity of test-score interpretations and accountability to test users. In an effort to address these concerns and allowing for transparency and accountability in the dissemination of information to all the relevant stakeholders, this study reports the findings aimed at evaluating the usefulness and impact of an occupational language proficiency assessment, the English Language Proficiency Assessment (ELPA), developed initially for the Malaysian Administrative and Diplomatic service but extended to a wider test population (other services) in the public service following a policy shift in 2003. This large scale validity inquiry opted to use Kane's (2006) interpretive argument approach as a test-bed for a strong validity inquiry program which provided a complete framework for building validity for test use and impact. Using an extended version of the three-bridge argument model, the researcher examined all aspects of the test cycle which included the fidelity of the scoring procedure, generalization of observed scores to universe of scores, extrapolation of observed scores to non-test behavior, and impact of the test on the public service. For each level of inference, the plurality of findings suggests that the ELPA provides an effective and fair indicator of English language competency regardless of the service being assessed. In general, the fidelity of the scoring procedures was supported by positive feedback from test takers despite administrators concerns about the overall management of the assessment.

Using classical test theory, generalizability theory, and multi-facet analysis findings for the generalization inference suggest moderate to high inter-rater reliabilities as well as high G and Phi coefficients indicating that test takers' ability contributed most to the magnitude of variance although other facets associated with the assessment were also examined. An interesting finding about the use of analytical and holistic scoring suggests that across different raters and groups, the analytical scores proved to be less stable than holistic scores if each of the rating scale criteria was incorporated in the analyses. The largest source of variance was contributed by rater-by-item (criteria) interaction which implicates the way the raters used the rating scale for each test taker. Further analyses from FACETS revealed that rater severity, although present, was not an issue. In addition, although no raters were found to be misfitting some were inconsistent in their own ratings.

Findings for the extrapolation inference confirmed, among others, the underlying structure of ELPA to be a three-factor correlated model where the performance indicators which consisted of the sub-tests and tasks test were found to measure the three latent traits quite well. In an attempt to validate the construct of ELPA by extrapolating test scores to actual use of language at the workplace, the data of two competing structural models suggest a weak to modest fit. Non-performance indicators

PAPER ABSTRACTS

such as frequency and type of activities which required the use of English at the workplace suggest a weak to moderate relationship to the performance indicators. The fourth level of inference included findings for test use and impact, an area often neglected in validity inquiry. Feedback received from three stakeholders suggests that on the whole, the ELPA has had a positive impact on the public service. Test takers and policymakers alike provided insights as to the reasons for the utilization of the ELPA across different services such as meeting the higher objectives of enhancing language competency in the public service.

The role of quantitative and qualitative methodologies in the development of rating scales for speaking

Evelina Galaczi, *University of Cambridge ESOL examinations*
(galaczi.e@cambridgeesol.org)

The CEFR (Appendix A) advocates the use of both qualitative and quantitative methodologies in the design of performance assessment scales. This paper will present the revision of Cambridge ESOL's rating scales for Speaking and will argue for the need for triangulation of analytic approaches. The presentation will illustrate this premise with the description of several studies which supported the revision of the rating scales in question and utilised both quantitative and qualitative methodologies.

The presentation will fall into three parts. First, a brief overview of the Cambridge ESOL rating scales for Speaking will be given and a case will be made for the necessity for their validation prior to use in live conditions. Next, the presentation will focus on three general methodologies which supported the revision of the scales, as outlined in the CEFR: an intuitive, a qualitative and a quantitative phase. The intuitive phase, which consisted of the setting out of the design principles for the assessment scales and was based on experts' reviews of current practice and their experience, will be overviewed. The qualitative and quantitative phases of the assessment scales' construction will then be discussed, with a focus on (a) a scaling exercise, which involved a rank ordering of the rating scale descriptors through multifaceted Rasch analysis; (b) a verbal protocol trial, which focused on raters' perception of the descriptors while rating performances and (c) an extended trial which confirmed the soundness of the descriptors prior to their live roll out. The presentation will end with a discussion of the value of gathering evidence from multiple methods and drawing on both qualitative and quantitative approaches to research in language assessment.

PAPER ABSTRACTS

Session 2: Wednesday, March 18, 2:45-4:45 p.m.

Clustering to inform standard setting in an oral test for EFL learners

Hongwen Cai, *University of California, Los Angeles* (hwcai@mail.gdufs.edu.cn)

This study compares two different approaches to setting cut scores in a high-stake test of oral proficiency for college-level EFL majors in China. The test under study, Oral Test of Graded Test for English Majors, Band 4 (TEM4), is a nation-wide examination of students' progress in spoken English toward the end of the second year of college education. Those who pass the test will receive a certificate, which will be an important qualification in their job-hunting efforts. Consequently, setting cut scores between pass and failure is a crucial task for test developers and administrators.

As a starting point, Angoff methods are used to set the cutting scores. After the scores of the five-component test are yielded through double-marking, a panel of expert raters is convened to make the final decision on the borderline cases taken from a sample of 640 cases. Cut scores are then derived from these decisions for the whole population.

Traditionally, multiple cut scores are derived from the sample statistics through panel discussion, with various cut scores for different components of the test. To qualify for a certificate, a test taker must pass all five cutting scores. A more recent approach to assigning test takers to either the pass or failure group applies the k-means clustering procedures using the means of the sample cases in both pass and failure groups as initial cluster centers.

Whether the multiple cut scores approach or clustering is adopted, separate results are obtained from the scores given by each rater. When results across raters conform, the assignment is final. When disagreement arises across raters, however, the relevant test taker's test performance will be assessed by a third rater, who finally decides the status of the test taker.

The focus of this study falls on the different performance and efficiency of the two approaches to assigning test takers to the pass or failure group. Three criteria are used:

1. The degree of conformity across raters, measured through percentages of test takers assigned unanimously across raters;
2. The explanatory power of the five test components in assigning test takers to the appropriate groups through each approach, as a test of whether the approach under concern fully reflects the complexity of component scores; and
3. The stability of the approaches, measured through discriminant analysis as the degree of agreement between group membership predicted by the model and group membership actually assigned through each approach.

Findings on 3,233 test takers, the whole population of one rating venue, favor clustering over the multiple cut scores approach in all three criteria. The use of the relatively new approach of clustering to inform standard setting in language assessment is recommended with certain cautions.

PAPER ABSTRACTS

Investigating the effectiveness of individualized feedback to rating behavior on a longitudinal study

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

The effectiveness of individualized feedback on rater behaviour has been investigated in several previous studies (e.g. Lunt, Morton, & Wigglesworth, 1994; Wigglesworth et al., 1993; Elder et al., 2005). Although not all studies were able to show its effectiveness, overall findings were favourable. However, all of these studies investigated the value of such feedback on a one-off basis, and it is thus not clear how raters use this type of feedback over several administrations of a test. Furthermore, previous research has focussed only on one language skill at a time, and we do not know if raters can incorporate the feedback in a similar way when rating speaking and writing. A better understanding of this could prove useful in further refining rater training and feedback processes.

This study tracks the rating behaviour of 20 raters assessing a large-scale ESP assessment for the health professions over a period of eight months. After each administration, raters received detailed performance profiles of their rating behaviour when rating both speaking and writing. The feedback, which was generated using the multi-faceted Rasch program FACETS (Linacre, 2006), provided details on raters' relative severity when compared to the other raters in the group, their internal consistency, as well as about any individual biases towards categories on the analytic rating scale they might have displayed. Raters' subsequent rating behaviour was monitored and tracked over each subsequent administration by conducting detailed FACETS analyses. Raters also completed a questionnaire and a subset of raters was interviewed to ascertain their views on the effectiveness of this feedback.

The findings showed that only some raters were able to incorporate the feedback into their ratings. Adjusting their relative severity was found to be easier than making any changes to internal consistency or reducing any individual biases towards categories on the rating scale. When writing and speaking raters were compared, it was interesting to note that the rating behaviour was much more consistent across administrations for writing than it was for speaking. The questionnaire and interview data suggested that those raters who were more positively disposed to the feedback were also more likely to be able to incorporate the feedback into their ratings. In the questionnaires and interviews, raters identified and evaluated the usefulness of differing sources of influence on their rating behaviour.

Further refinements to this approach to training are suggested and implications for rater training are discussed.

PAPER ABSTRACTS

Exploring rating process and rater belief: Transparentizing rater variability

Jie Zhang, *Shanghai University of Finance and Economics/Guangdong University of Foreign Studies*, (zhangjie617@gmail.com)

In performance assessment, rater variability has long been held as challenge to ensure test reliability and validity as well as the transparency of score interpretation. The present study takes a focusing-on-rater and process-oriented approach to investigating rater variability by comparing raters with different levels of rating accuracy in terms of their internal rating processes and scoring-related beliefs in the context of CET4 (College English Test Band 4 in China) essay scoring. Three major sources of data were collected through separate procedures in the empirical study, including an independent rating session, a concurrent think-aloud session and a subsequent semi-structured interview session. Raters' rating performance as compared with the expert norm were first assessed using MFRM. Raters were then classified into two groups with higher and lower levels of rating accuracy based on the individual calibrations of their rating performance indicated by MFRM statistics. Based upon verbal protocol analyses of raters' concurrent think-aloud and interview transcripts, comparison was made between the identified HIGH and LOW groups in terms of their internal processes (text focus and mental processes) during rating and the content and structure of their scoring-related belief systems.

The comparison has detected considerable differences in many important ways. It was mainly found that good raters would direct their attention to a wide range of textual features and their constructed text image would be more comprehensive and balanced than the poor raters. When processing the acquired information, good raters would more often adopt effective error-diagnosing, summarizing and inferring behaviors to abstract, integrate and categorize the details and particularities in their text images into evaluation and inference about the essay quality and conduct more self-monitoring strategies such as weighing and assessing own ratings to make self-reflection on their own rating accuracy. Different rater groups also differed in their scoring-related beliefs. The major finding is that good raters often have a comprehensive, balanced, and well-organized beliefs about the assessment target, in which a wide range of requirements on examinees' writing ability and performance are involved and clear differentiation is usually made among different levels of language use and different aspects of text quality, with a well-ordered hierarchy of the importance for those levels and aspects in defining essay quality. As a result, they would also be more clear and systematic in their interpretation and operationalization of the rating rubrics and their perceived effective solutions to the uncertainty during their decision-making.

Through linking the detected variability at each level, it was therefore contended that raters' scoring-related beliefs would serve as the internal context in which the dynamic rating processes take place and that variability in their belief systems is at the root of other sources of rater variability. Therefore, the main objectives in the whole scoring system of large-scale performance assessment should be to make transparent the expected values and understanding about the assessment target and instrument, effectively communicate them from the higher level of administration to the individual

PAPER ABSTRACTS

raters and gradually bring the raters into a common “judgment community” which would share similar core beliefs about scoring in the given test context.

PAPER ABSTRACTS

Session 3: Thursday March 19, 8:45 – 10:15 a.m

From the periphery to the centre in Applied Linguistics: The case for situated language assessment

Pauline Rea-Dickins, *University of Bristol* (P.Rea-Dickins@bris.ac.uk)

Matt Poehner, *Penn State University* (mep158@psu.edu)

Constant Leung, *King's College London* (Constant.Leung@kcl.ac.uk)

Lynda Taylor, *University of Cambridge ESOL Examinations* (Taylor.L@ucles.org)

Elana Shohamy, *Tel Aviv University* (elana@post.tau.ac.il)

This paper calls for an increase in research addressing the specificity and contextual features of situated language assessment practices so as to extend our understandings of fair and equitable educational and social processes. We argue that much language testing and assessment (LTA) research has been narrowly focused and therefore our construct of language proficiency risks remaining under-developed, inadequately capturing the vast and complex range of language needs and assessment requirements in our global world.

Through a reflection on past practice, we first explore ways in which LTA has moved from periphery to centre stage in Applied Linguistics (AL), taking into account effects of globalisation and migration, increased demands of accountability, and the imperative for a socially responsible and ethical positioning in assessment. Second, we report on recent empirical case studies to show how new insights can be developed about language(s) use when mediated through diverse socially situated assessment practices and we discuss the implications for broadening the constructs that inform valid assessment. These include the assessment of multilingual and interactional performance in both subject and language learning classrooms, the role of language in migration/citizenship and in professional practice, and definitions of language proficiency that go beyond NS norms. We raise several questions: e.g. What might spoken language proficiency mean in different contexts? Can/should content knowledge be assessed through more than one language? Does the traditional focus on reliability deny the value of variability? Is LTA a privilege of the developed world?

In summary, we argue that LTA has become a central means of addressing real world problems, one that takes account of different contexts and needs. LTA research should, thus, be oriented around the purposes and effects of assessment on individuals/groups from a real world, holistic perspective. In a world seeking to balance the twin drivers of 'globalisation' and 'localisation', the 'ecology' of language assessment - involving greater transparency, responsibility and collaboration - is ripe for investigation.

PAPER ABSTRACTS

The extent to which differences in L2 group oral test scores can be attributed to different rater perceptions or different test taker performance

Gary Ockey, *Utah State University* (gary.ockey@usu.edu)

Over the past decade, the group oral discussion task has received increased attention as a test of second language speaking ability (e.g., Bonk & Ockey, 2003; He & Dai, 2006; Turner, 2008). In the group oral discussion task, three or more test takers discuss a topic without any prompting or interaction with interlocutors; after a group has been assigned a topic, only the test takers themselves control the discourse that is produced.

An established threat to the validity of the score interpretations of the group oral is that test takers' scores have been shown to be influenced by group membership. That is, a rater might assign a test taker a higher or lower score depending on the group members with whom the test taker is assessed (Berry, 2004, Bonk & Van Moere, 2004; Ockey, 2009). What is not clear is the extent to which it is the test taker's performance or the rater's perception of the test taker's performance which is affected by the group to which the test taker is assigned. The aim of the study was to answer this question.

The sample included videotapes of 60 non-native speakers of English taking an oral English proficiency examination in the group oral format. The videotaped tests were based on a medium-stakes assessment used for placement in which test takers in groups of three were asked to discuss a topic related to previous class discussions for a 12-minute period of time. Each of the 60 test takers was tested in a group of three members on two separate occasions. Test takers represented various first language groups and all were graduate students.

The videotaped group oral test performances were digitized, and four copies were produced for each of the two rating occasions. One copy was not manipulated, while each of the other three was manipulated. In the manipulated copies, two of the three test takers' voices were muted and their appearances blurred so that they could not be recognized. Thus, one version had no muted voices and no blurring, and each of the other three versions had a different person whose voice was not muted and appearance was not blurred. This manipulation resulted in two treatment conditions: regular group oral situation in which all three test takers could be seen and heard discussing a topic and individualized situation in which only one individual in the group could be seen and heard.

Four trained raters with advanced degrees in teaching English as a second language rated all test takers in both conditions on the two occasions. Counterbalancing was employed to control for a rating order effect, and the ratings were spread over a three-month period to diminish the effects of raters remembering how they had assessed a test taker previously. A six-point analytic rating scale was used to evaluate the test takers' performance in three areas: comprehensibility, fluency, and accuracy. Estimates of reliability based on Cronbach's alpha for the four raters were high: 0.96, 0.94, 0.87, and 0.80.

ANOVA and generalizability analyses (Shavelson & Webb, 1991; Brennan, 2001) were conducted on the data. The facets in the completely crossed analyses were raters,

PAPER ABSTRACTS

occasions, and rating situation. Results of the analysis along with implications for rater training and administration of the group oral will be discussed.

The analysis of test takers' performances under their test-interlocutor influence in a paired speaking assessment

Jiyoon Lee, *University of Pennsylvania* (jiyoon@dolphin.upenn.edu)

Since Paired Speaking Assessment (PSA), in which two non-native speaking test-takers perform their target language use ability, has been introduced in second language classroom as well as high-stakes testing, it attracts great attention due to its potential advantages and caveats. It is argued that PSA can measure a test-taker's ability to interact with an interlocutor who has the equal status as test-takers themselves. It is hardly the case in other speaking assessments such as in an interview where a high-authority figure usually initiates and manages the interaction. However, as two test-takers interact in PSA, it is extremely significant to examine the dynamics and variables that these test-takers bring in the situation since these factors can influence the validity of testing results. In this study, the potential influence of interlocutors' proficiency on a test-taker's performance is thoroughly researched in terms of the test-taker's linguistic (i.e., syntactic accuracy and complexity) and interactional performance (i.e., interactional contingency, goal orientation, and dominance). Each test-taker interacts with higher, same, and lower proficiency interlocutors on three statistically equivalent decision-making tasks (i.e., Cambridge FCE section 4 tasks). These three interlocutor groups are distinctive in terms of their language ability; however, their social status, age, and gender are compatible to each other. Raters' reaction to the test-taker's performance in these conditions is also correlated with the test-taker's actual linguistic and interactional performance results. Repeated measure ANOVA is employed to analyze the data. In addition, thorough transcription analysis will provide more detailed information regarding test-takers' performance. The results of this study are helpful to understand 1) whether a test-taker's performance may change depending on his/her interlocutors' proficiency and 2) whether raters' perception may influence on their evaluation practice. The findings of this study are conducive to understand the value of PSA, and provide guidance on cautious introduction and use of this format in second/ foreign language classroom and high-stakes testing.

PAPER ABSTRACTS

Session 4: Thursday March 19, 10:30 a.m. – 12:00 p.m.

Towards a transparent construct of reading-to-write assessment tasks: The interface between discourse features and proficiency

Atta Gebril, *The United Arab Emirates University* (attag@uaeu.ac.ae)

Lia Plakans, *The University of Texas, Austin* (lplakans@mail.utexas.edu)

Reading-to-writing tasks have been recently introduced to a number of L2 exams. However, little empirical evidence has been offered to describe the discourse features in reading-to-writing assessment tasks across different proficiency levels. This information is of significant importance to the validation of these writing tasks, and particularly to an accurate, transparent definition of the reading-to-write construct.

For this purpose, the researchers asked 139 students to write an argumentative essay based on a reading-to-write task - test takers read two passages before writing their essays. The essays were holistically scored by two raters and a third rater was employed in case of disagreement. After scoring, the essays were classified into three proficiency levels. For the analysis of the discourse features, the researchers selected a number of measures to identify the discourse features in the essays based on the literature (Cumming et al, 2005; Wolfe-Quintero et al, 1998): vocabulary richness, accuracy, fluency, and syntactic complexity. In addition, to identify the role of reading in the written results, source use was analyzed in terms of verbatim use across proficiency levels. One-way ANOVA was used to compare the writing features across the three proficiency levels.

Results of the study showed significant differences among the three proficiency levels in most of the measures used in this analysis. More specifically, test takers with a higher proficiency level appear to produce longer essays, more sophisticated vocabulary, and more syntactically complex essays than those with lower proficiency. Verbatim source use revealed somewhat complex patterns across the proficiency levels showing the challenges in interpreting scores from reading-to-write tasks. The study has a number of implications with regard to construct definition and scale development. The analysis of the discourse features could help in better understanding the development of writing across different proficiency levels, and consequently in understanding the reading-to-write construct. In addition, the scoring rubric used in this study, which was a modified TOEFL iBT integrated task scoring rubric, appears to be successful in distinguishing among the different proficiency levels. Suggestions for further research are provided.

PAPER ABSTRACTS

DIF investigation of TOEFL iBT Reading Comprehension: Interaction between content knowledge and language proficiency

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

Ou Lydia Liu, *Educational Testing Service* (lliu@ets.org)

This proposed study aims to investigate potential interaction between background knowledge and language proficiency in affecting performance on the TOEFL iBT Reading Comprehension Test. Five passages from four recent TOEFL iBT test administrations were selected and item level data from 7310 students were analyzed. The five selected passages include three of non-American culture topics and two of physical science topics. Each passage had 14 comprehension items. Focal and reference groups were identified from these 7310 test takers for each passage through an online survey about examinees' academic and cultural background. The students in the focal group and reference group were further classified into high proficiency group (with scores between 91-120) and low proficiency group (with scores between 0-90) on the basis of their TOEFL iBT total score. For the low proficiency group, the number of test takers in the focal group ranged from 107 to 279 for the five passages, and the number ranged from 525 to 647 for the reference group. For the high proficiency group, there were 99 to 330 test takers in the focal group, and 602 to 1013 test takers in the reference group across five passages.

Three types of differential functioning had been investigated: (1) differential item functioning (DIF); (2) differential bundle functioning (DBF); and (3) differential passage functioning (DPF). For DBF, items were classified into item bundle by content experts if they contain academic or cultural specific terminologies. The standardization DIF method was used to perform the DIF, DBF, and DPF analyses. The DIF items were classified into three magnitude levels at small, intermediate or large according to ETS DIF guidelines. The existence of differential functioning indicates a potential background effect on reading comprehension at the item level or bundle level. Interaction effect between the background knowledge and language proficiency can be observed from the difference in the DIF presence and magnitude between the low and high proficiency group.

The results suggest that background knowledge had a mixed effect on reading performance, with some items favoring the focal group and others favoring the reference group. Only one of the five passages investigated had intermediate DBF and DPF, in favor of the focal group. The conclusion is that background knowledge in general has little effect on TOEFL iBT reading performance.

Regarding the interaction effect, the number of DIF items showing interaction between the two proficiency groups varies across passages. The differences in DIF presence or magnitude may be contributed to certain item and passage characteristics. No interaction was observed on the bundle and passage level between the high and low performing groups.

This research sheds new light on the understanding of background effect and its interaction with language proficiency in second language reading comprehension

PAPER ABSTRACTS

literature, it also has significant practical implications for test development to advancing fair assessments.

Completion as an Assessment Tool of L2 Reading Comprehension: Building a Validity Argument

William Grabe, *Northern Arizona University* (William.Grabe@nau.edu)

Xiangying Jiang, *Northern Arizona 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, and 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 addresses the issues related to reliability and validity of the GO completion test in assessing L2 reading comprehension. This study involves data collected during two phases of research. During the first phase, a GO completion test of comprehension of three reading passages was administered to 340 college EFL students together with the reading comprehension section of the TOEFL. During the second phase, a GO completion test of comprehension of two reading passages was administered to 640 ESL students (at various proficiency levels and with multiple L1 backgrounds) and 40 college-level English L1 speakers along with a multiple-choice comprehension test. The reliability, content, concurrent, and construct validity of the GO completion test were substantiated.

The results demonstrated that GO completion task can be a reliable and valid item type for L2 reading assessment. 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. 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.

PAPER ABSTRACTS

Session 5: Friday, March 20, 8:45 – 10:15 a.m.

Judgments of L2 comprehensibility, accentedness and fluency: The listeners' perspective

Talia Isaacs, *McGill University* (talisaacs@elf.mcgill.ca)

Ron Thomson, *Brock University* (rthomson@brocku.ca)

Accents are subject to social evaluation, and listeners' attitudes toward L2 accented speech have the potential to impact their interactions with non-native speakers and to bias their assessments. Enhancing cross-cultural communication, where listeners and speakers assume an equal share of the "communicative burden," entails a greater understanding of which components of non-native speech are the most salient to different groups of listeners and the extent to which features that they reportedly attend to are present in the actual discourse (Lindemann, 2006).

Building on research that has investigated the effects of rater expertise and rating scale length on rater decision making in pronunciation assessment (Isaacs & Thomson, 2008, May), this mixed-methods study examines the extent to which expert and non-expert raters' perceptions of L2 speakers' pronunciation performance align with measures derived from an analysis of the speakers' actual discourse (cf. Brown, Iwashita & McNamara, 2005). The study replicates, expands upon, and critiques conventions for defining and operationalizing major constructs in L2 pronunciation research, with the overall goal of shedding light on which features raters consciously attend to in the absence of scoring rubrics and how these dimensions relate back to the construct.

Twenty expert and 20 non-expert raters assessed speech samples of 38 newly-arrived immigrants to Canada (19 Mandarin, 19 Slavic) on numerical rating scales for each of comprehensibility, accentedness, and fluency. Stimulated recalls and post-task interviews were employed to gauge raters' impressions of the speech and perceptions of the rating process. Following the generation of a preliminary coding scheme, the stimulated recall data were mapped onto the interview data and triangulated with the quantitative rating data. Emergent categories reflecting aspects of the speech that raters reportedly attend to were then used to generate quantitative measures that were employed in an independent analysis of the speech samples (e.g., speaking rate, substitution errors). A major area of focus was on L1 transfer effects that seemed to incite positive or negative responses from the raters.

Results showed some correspondence between the raters' comments, the scores that they assigned, and features of the L2 performance, although the interplay between segmental, prosodic, and temporal aspects of the speech was more complex than raters were able to articulate in the research setting. Comments linking voice quality with personality attributes were persistent in the data and, when viewed in relation to scoring decisions, lent insight into possible sources of construct irrelevant variance and rater bias. A strong L1 effect was observed, though the ratings did not seem to be affected by accent familiarity. Based on commentary from the raters and our data collection experience, we suggest that construct definitions that are provided to raters in operational pronunciation research be more precise and task specific (Fulcher, 1996;

PAPER ABSTRACTS

Chalhoub-Deville, 1996) and that researchers be mindful of the standard that is imposed when speech samples of native speakers are used to establish the upper bounds of the scale during rater training. Broad implications for pronunciation assessment and for enhancing native-speaker interactions with non-native speakers in countries with a large influx of migrant workers are discussed.

In the ear of the beholder: Dependence of comprehensibility on language background of speaker and listener

Alistair Van Moere, *Pearson* (avanmoere@ordinate.com)

Ryan Downey, *Pearson* (rdowney@ordinate.com)

Previous research on the comprehensibility of language learners' speech has shown that there may be differences in how L2 speech is perceived depending on the language background of the listener (e.g., Munro, Derwing, and Morton, 2006). Some research finds a benefit to non-native speakers' comprehensibility when the listeners are themselves non-native speakers (e.g., Bent & Bradlow, 2003). With recent global trends in employing non-native English speakers (NNS) as customer service representatives for English-speaking customers, increasing focus is being brought to studying attributes of NNS speech which contribute to impediments in comprehensibility.

The current research concerns whether naïve, monolingual English speakers with limited exposure to NNS tend to judge the speech of individuals who speak English dialects referred to as "Indian English" (IE) as being less comprehensible than English spoken by individuals from other L1 backgrounds (Other English, or OE, speakers). A study was conducted investigating how NNS speech is rated by listeners from different language backgrounds, designed to address several specific questions, including:

- 1) Do listeners of different language backgrounds systematically assign more or less harsh scores to IE speakers when judging comprehensibility, pronunciation, and fluency?
- 2) Are IE speakers judged as more or less comprehensible when compared to OE speakers with similar pronunciation and fluency scores? To what extent does this depend on judges' language backgrounds?
- 3) Is there a "comprehensibility threshold" at which a speaker's English ability is rated as "not good" (i.e., poor pronunciation and fluency) but nevertheless does not interfere with understanding?

An automated test of spoken English (the Versant English Test) was used to provide a consistent measure of test takers' ability in Pronunciation and Fluency. A total of 640 individual speech samples were extracted from tests taken by 160 test takers (40 IE and 120 OE speakers). Samples were randomized and assigned to human raters in sets, so that all raters first rated all 640 samples according to Comprehensibility, then Pronunciation, and finally Fluency.

PAPER ABSTRACTS

Rater groups from different language backgrounds assigned ratings to all samples. Rater groups included: Indian English experts (those who speak an IE dialect or have extensive experience teaching or assessing English in India); Standard American English experts (native English speakers with experience teaching/assessing SAE); SAE non-experts (native speakers of SAE with no formal experience in linguistics or language assessment); and speakers of Other English varieties (non-native English speakers with varying degrees of experience with English).

The results have implications for making high-stakes employment decisions in domains such as the customer service and call center industries.

Temporal aspects of perceived speaking fluency

Nivya DeJong, *University of Amsterdam* (n.h.dejong@uva.nl)

In second language testing practice, speaking fluency is usually assessed by human raters. In a large scale study (n = 1007), I investigated to what extent temporal aspects of test takers' speech are related to human raters judgements on different levels of speaking proficiency. Previous studies investigating the relation between temporal measures of speech and ratings on fluency have usually found that speech rate incorporating pauses (number of syllables per second on total time) best predicts ratings (Cucchiaroni, Strik, & Boves, 2002; Kormos & Denés, 2004). In this study, I have disentangled speech rate and pausing behaviour by measuring speech rate without pauses, i.e. the number of syllables per second on speaking time, as well as pausing behaviour.

Trained raters assessed oral fluency of 1007 short monologic responses (up to 40 seconds). They were instructed to pay attention to speech rate, unnatural pausing behaviour, and hesitations. At the same time, temporal aspects of speech were calculated using a script written in PRAAT, a computer software program for the analysis of speech. With this script I computed speech rate and pausing behaviour separately: speech rate was measured as the number of syllables per second when speaking, and pausing behaviour was measured in terms of mean length of pauses and the number of pauses per minute.

All speech samples were also rated on overall speaking proficiency by other judges using the CEF-scale. Regression analyses showed that for low-proficient speakers (under B2), human judges were sensitive to both speech rate as well as pausing behaviour when rating on fluency. However, for high-proficient speakers (B2-C2), only pausing behaviour significantly predicted human scores. At the same time, I found that for the low-proficient speakers human fluency ratings could be predicted much more accurately than for the high-proficient speakers.

In conclusion, human raters who are instructed to rate fluency are sensitive to actual pausing behaviour and to a lesser extent to speech rate when they make their judgements. Furthermore, perceived fluency at higher proficiency levels is much less related to global temporal measures of speech than at lower levels.

PAPER ABSTRACTS

Session 6: Friday, March 20, 10:30 – 11:30 a.m.

Assessing domain-general and domain-specific academic English language proficiency

Anja Rumhild, *University of Nebraska* (roemhild@bigred.unl.edu)

Dorry Kenyon, *Center for Applied Linguistics* (dorry@cal.org)

David MacGregor, *Center for Applied Linguistics* (dmacgregor@cal.org)

Within the last decade, academic English language proficiency has become a major focus in the assessment of English language learners (ELLs) in the United States. However, the construct of academic English language (AEL) is still not very well understood. A general definition of AEL (Chamot & O'Malley, 1994) is as the language used in the classroom for the purpose of acquiring content-specific skills and knowledge. Recently, Bailey and Butler (2003) developed a conceptual framework that describes in more detail how the construct of AEL can be operationalized in language tests and language curricula. They hypothesized that an important dimension of variation is between linguistic features that are common to various academic domains (i.e., domain-general) and linguistic features that are unique to individual content areas (i.e., domain-specific). While this distinction is conceptually intuitive, it has not been investigated empirically.

The purpose of this study was to examine domain-general and domain-specific AEL from the angle of a construct validity study using a latent variable modeling approach. Specifically, the goal was to model domain-general and domain-specific variance in a latent factor model to evaluate and compare the salience of these variance sources. The analyses were carried out on data from multiple test forms targeting academic English language proficiency at different grade and proficiency levels, which affords comparisons of the latent factor models across different ELL student populations.

The study is based on test data from a large-scale assessment of English language proficiency for K-12 learners currently used by 19 states. The test is an operationalization of model performance indicators defining five English Language Proficiency Standards. The test purports to measure academic English language proficiency in four academic content areas as well as "Social and Instructional language." The test is organized in five grade clusters within which are three overlapping test forms targeting five different proficiency levels. For this study, data from nine test forms from the upper elementary, middle, and high school clusters were analyzed.

The results of this study revealed that at low levels of English language proficiency, domain-specific variance did not play a significant role in explaining examinee performance on test items across the five standards. At the mid and high levels of proficiency, however, the presence of domain-specific variance was increasingly observable through a general increase in model fit and increasing salience of individual item factor loadings. These results were replicated across grade clusters. The empirical findings suggest that AEL differentiates between domain-general and domain-specific

PAPER ABSTRACTS

dimensions with increasing English language proficiency. Thus, when considering the construct of AEL, level of English language proficiency must not be ignored.

Defining the construct of academic writing to inform the development of a diagnostic assessment

Lorena Llosa, *New York University* (lorena.llosa@nyu.edu)

Sara W. Beck, *New York University* (sarah.beck@nyu.edu)

Cecilia G. Zhao, *New York University* (gz312@nyu.edu)

Given the high stakes currently attached to students' performance on writing assessments in the US, it is problematic that we know little about the nature of the writing tasks students face in classrooms and on tests and about the challenges students experience with these tasks. The purpose of this study was to address this gap by: 1) developing an evidentiary framework for a definition of academic writing at the high school level; and 2) developing an inventory of difficulties that both English Language Learners (ELLs) and native English speakers (EOs), experience with this type of writing. This framework and inventory will serve as the basis for the construct definition of a diagnostic assessment of high school students' difficulties with academic writing.

Following the steps for developing an evidentiary framework for a construct outlined by Bailey and Butler (2003), we analyzed a wide array of documents and data to inform our framework including: the academic writing demands in state English language arts (ELA) and English as a Second Language (ESL) standards and high-stakes tests; interviews with 12 9th and 10th grade ELA and ESL classrooms teachers in New York City (the largest school system in the United States); and observations of their classroom instruction on academic writing. To inform the inventory of difficulties we drew from the classroom observations and teacher interviews as well as interviews with 25 students about their expectations for academic writing and perceptions of sources of difficulty; and verbal protocol data to capture the students' composing processes. Standards, tests, and classroom observation data were compiled and analyzed using a task characteristics framework based on Bachman and Palmer (1996). Interview data were analyzed inductively for themes related to the characteristics of the writing tasks and student difficulties in completing them, while verbal protocol data were analyzed using categories from prior protocol-based studies of composing processes (e.g., Cumming, 1989; Sasaki, 2000).

The analysis of standards, tests, and classroom instruction indicate that the genres of writing represented most prevalently in our sample are analytic in nature, defined by Schleppegrell (2004) and Martin (1989) as genres that require the writer to engage with the question of why: e.g. to explain why something is so, or to argue for why an action should be taken, or why an interpretation or argument is valid. Our analyses of the student interviews and verbal reports revealed that a majority of students lacked awareness of writing as crafted for a particular purpose and audience, and did not use the writing process. Students also experienced difficulties understanding the task

PAPER ABSTRACTS

prompt, producing a narrative-type plot summary even though the prompt was designed to elicit an argument. Our comparison of data from ELL and EO students revealed that even though some of the ELL students struggled with finding the right vocabulary to express their ideas, in general they experienced the same types of difficulties as EO students, especially in the areas of task interpretation and ability to decipher genre expectations. ELL students, however, experienced these difficulties in greater numbers and with greater frequency than EOs.

Our findings suggest that a diagnostic assessment of academic writing for high school students should focus on analytic writing and should include tasks that address the areas of difficulty prevalent among this population: task interpretation, argument construction, planning and revision.

Profiles of linguistic ability at different levels of the European Framework: Can they provide transparency?

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

Nivya DeJong, *University of Amsterdam* (n.h.dejong@uva.nl)

Margarita Steinel, *University of Amsterdam* (M.P.Steinell@uva.nl)

Arjen Florijn, *University of Amsterdam* (A.F.Florijn@uva.nl)

Jan Hulstijn, *University of Amsterdam* (J.H.Hulstijn@uva.nl)

The Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) proposes six levels of language proficiency, defined as a combination of (1) what an L2 user can do in terms of domains, situations, roles, topics, constraints etc. (i.e., the mainly functional scales of chapter 4), and (2) how well the L2 user can perform (i.e., the scales of linguistic quality of performance in chapter 5). The CEFR resulted from the scaling of performance descriptors, derived from a number of examination systems, using the judgments of experts. However, little is known about the actual relationship between language task performances and required linguistic abilities of L2 learners. The study that we report on in this presentation, aims to address this issue.

Almost 200 intermediate and advanced adult learners of L2 Dutch performed in eight speaking tasks, covering formal and personal situations requiring production of descriptive or argumentative speech acts. A panel of six experts rated one task to be at the A2 level, three tasks at the B1, and four at the B2 level. Each of the 1600 productions (200 x 8) was rated by four raters on a rating scale consisting of 6 main categories, exclusively defined in terms of communicative adequacy, not referring to linguistic quality. In this way, we established participants' functional L2 proficiency to be at (at least) B2, B1, or lower than B1. Participants also performed on seven linguistic and psycholinguistic ability tests, in the domains of vocabulary, grammar, and pronunciation.

The main question we address in this presentation is: What is the compositional nature of speaking ability at the B1 and B2 levels of functional adequacy, in terms of linguistic

PAPER ABSTRACTS

and psycholinguistic subskills? To answer this question, we report on discriminant analyses of the speaking performances at the distinguished B1 and B2 level, in terms of the (psycho)linguistic subskills of these participants. We will discuss the extent to which our results can connect the two parts of the CEFR division between functional and linguistic proficiency scales, and thus add to the transparency of the former type of scales.

Session 7: Friday, March 20, 1:30 – 3:00 p.m.

Relative impact of rater characteristics versus speaker suprasegmental features on oral proficiency scores

Don Rubin, *University of Georgia* (drubin@uga.edu)

Okim Kang, *Northern Arizona University* (okim.kang@nau.edu)

Lucy Pickering, *Georgia State University* (esllup@langate.gsu.edu)

True score variance of a test of English speaking proficiency is a function of the comprehensibility of examinees' language (together with the degree to which examinees' discourse is responsive to the demands of the examination tasks).

Variation among raters due to such individual characteristics as familiarity with NNS speech or attitudes toward international sojourners is trait-irrelevant and undermines score accuracy. Yet previous studies in educational settings indicate that evaluations of speech performance are exquisitely susceptible to listener expectations and biases.

Rater background characteristics that may mediate bias in judgments of oral proficiency include native speaker status, level of educational attainment, linguistic sophistication, and ESL/FL teaching experience.

Our contention, then, is that even expert judgments of comprehensibility or accent are potentially subject to rater bias and therefore of limited utility as criterion or covariate measures in studies whose very purpose is to estimate the impact of rater background and bias on oral proficiency ratings. Instead, a proxy for true-score components of oral proficiency scores must be derived from objective measures of pronunciation, such as computer-assisted acoustical measures of suprasegmentals.

Accordingly, this study examined the relative contributions to iBT TOEFL® speaking scores of (a) a set of four suprasegmental parameters of NNS test-takers' speech and (b) a set of six rater background and attitudinal characteristics. Speech samples rated in this project were produced by 28 male speakers, representing four different L1s, each responding to four tasks. Each of the 112 samples was analyzed for 29 different suprasegmental pronunciation variables using computer-assisted methods. In the first phase of this study each sample was also judged by 82 untrained raters of varying backgrounds.

Separate regressions were run to ascertain (a) effects of rater background and attitudinal factors on ratings and (b) effects of acoustically measured suprasegmental pronunciation (i.e., trait-relevant) factors on ratings. Criterion variables for regression analyses included (1) holistic scores, (2) deviation of raters' holistic scores from those of "official" ETS raters, and (3) raters' impressions of examinees' comprehensibility. Other regressions examined predictors of (4) rater severity scores derived from Rasch modeling and (5) official ETS holistic scores (i.e., scores obtained from ETS-trained and employed raters and duly reported to test-takers).

Results indicated that, in general, about 20% of the variance in scores of naïve (minimally trained) raters could be attributable to rater background and attitudinal factors, that is, to trait irrelevant factors. The most potent of those factors was rater

PAPER ABSTRACTS

native speaker status (native vs. non-native speakers of English). The amount of weekly contact with non-native speakers affected one rating outcome, and reverse linguistic stereotyping affected the degree of fidelity or deviation from official ETS scores.

In contrast, over 60% of the variance in official ETS scores and over 70% of the variance in scores obtained from raters in this study could be accounted for by suprasegmental pronunciation factors, that is, by trait-relevant elements. The most potent of the suprasegmental factors was intra-run fluency, which was a positive predictor of oral proficiency ratings. Another factor, indistinct boundary markers was a negative predictor for one rating outcome. The incidence of indistinct boundary markers and of high fundamental pitch reduced fidelity and increased deviation of ratings in this study from official ETS ratings.

Implications for the practice of oral proficiency testing include the recommendation that raters be drawn from either native English speakers or non-native English speakers, but not both. Level of raters' routine contact with non-native speakers should be kept relatively uniform. Raters should be trained to attend to the suprasegmental components of intra-run fluency, but not to over-react to fundamental pitch fluency or to indistinct boundary marking. Test-takers can optimize their scores by maximizing intra-run fluency and avoiding indistinct boundary marking.

A meta-analysis of multitrait-multimethod studies in language testing research: Focus on language ability and Chelle's (1998) construct definition and interpretation

Yo In'nami, *Toyohashi University of Technology* (innami@hse.tut.ac.jp)

Rie Koizumi, *Tokiwa University, Japan* (koizumir@tokiwa.ac.jp)

One of the central challenges in the field of language testing is how to conceptualize the construct, which can be further categorized into two main issues: examining the approaches to be taken in defining the construct (Bachman, 2007; Chapelle, 1998; Messick, 1993) and investigating the internal structure of language ability being measured within the construct (e.g., Bachman, 1990; Carroll, 1993). The most comprehensive review on the approaches to construct definition is perhaps the one by Chapelle (1998), which summarizes three perspectives toward construct definition on the basis of whether test performance is attributable to trait only, to context only, or to both trait and context. The structure of language ability has been continuously investigated since Oller's (1979) unitary trait hypothesis was proposed (and later refuted). These two topics have been simultaneously examined for the past 30 years, for example, by using a multitrait-multimethod (MTMM) design, where two or more traits are measured with two or more methods (Campbell & Fiske, 1959).

In order to gain a better understanding of these two topics and present the best picture currently available, the current meta-analytic study expands previous MTMM studies by reanalyzing and quantitatively synthesizing them. We specifically searched for

PAPER ABSTRACTS

models that were most likely to fit the data well across studies by investigating theoretically competing models in addition to the models originally tested. The two research questions examined were as follows: (1) Which approach to construct definition is most supported in L1 and L2? (2) Which structure of language ability is most supported in L1 and L2?

Forty-four data sources from 36 studies obtained through an extensive literature search were analyzed using confirmatory factor analysis. Eighteen rival models were tested for each MTMM matrix, and the best-fitting model was selected. These best-fitting models were classified into three types (trait, method, or trait-and-method) for Research Question 1, according to Chapelle's (1998) three approaches to construct definition, and into five types (unitary trait, correlated trait, uncorrelated trait, higher-order trait, and method models) for Research Question 2. The results reveal that in both L1 and L2 studies, the most empirically-supported construct definition is the interactionist approach, suggesting (a) the necessity to include both trait and method into construct definition and (b) some generalizability of test performance across contexts. Moreover, the most frequently occurring ability structure is higher-order in L1 and unitary in L2. Further analysis of moderator variables such as trait and method types is conducted to provide directions for future research.

Telling our story: Reflections on the place of learning, transparency, responsibility and collaboration in the language testing narrative

Lynda Taylor, *University of Cambridge ESOL Examinations*
(lynda_and_nigel.taylor@ntlworld.com)

A sure sign that an academic discipline has matured is its decision to look back and reflect thoughtfully, even self-critically, upon the journey so far. The gift of hindsight, a sense of perspective, and the wisdom that comes with time and experience all interact with the desire to give meaning to and learn from our narrative.

For language testing and assessment this tendency is visible in the publication of a series of narrative 'histories' in recent years, beginning with Spolsky's *Measured Words* in 1995. Since then other volumes or chapters have charted the rise (and fall) of specific language tests as well as the evolution of language testing as both an academic, research-oriented discipline and a community of practice, e.g. Weir and Milanovic 2003, Hawkey 2004, O'Sullivan 2006, Davies 2008, Taylor and Angelis 2008, and most recently Hawkey 2009. In addition, language testing journals now celebrate 'significant' anniversaries or publish interviews with eminent scholars as they reflect on a lifetime career in language assessment, and a Lifetime Achievement Award has been instituted to recognise long and distinguished service in our field. The LTRC 2009 theme betrays a similar desire to look back in a considered way and to benefit from the insights generated by this exercise.

Drawing on approaches from our sister field of Applied Linguistics, this paper reviews and evaluates attempts from within our research community to 'tell the story' of our field over several decades. It explores how the discourse of these narratives contributes to the construction of our identity as an academic discipline and community of practice

PAPER ABSTRACTS

in the present age. The threads of learning, transparency, responsibility and collaboration will be examined to uncover how they weave together into a complex and colourful tapestry that illustrates 'our story' – a story which may help us understand what has shaped us to this point and how our field might evolve in the future.

POSTER ABSTRACTS

Posters

Time: March 18, Wednesday 1:15 – 2:45 p.m. Location: Atrium

Development and validation of a web-based multimedia Korean oral proficiency test

Seongmee Ahn, *Michigan State University* (ahnseon2@msu.edu)

Ok-Sook Park, *Michigan State University* (ospark@msu.edu)

Daniel Reed, *Michigan State University* (reeddan@msu.edu)

In recent years, there has been increasing use of computer technologies in language assessment. Web-based tests are considered appropriate for less commonly taught languages, such as Korean, as they promote faster delivery and more convenient test administration among remote locations (Winke & Fei, 2008). However, web-based proficiency tests in teaching Korean are rare compared to web-based placement tests, despite a high demand among teachers of Korean. Thus, there is still a need to establish a useful tool for the measurement of global oral proficiency in Korean.

The purpose of the project is to develop and validate a web-based Korean speaking proficiency test. This poster describes the development of the test in terms of the test construct, operationalization of the construct, innovative use of new web-based technology and, importantly, a rater-training component. In addition, the research questions and a design for a validation study are presented. The main questions addressed in the study were 1) Are the test tasks valid? 2) Is this test scored reliably? 3) How do assessment reviewers and examinees perceive the validity of test tasks and the test as a whole?

Approximately five US educational institutions with Korean language programs were selected for this study. Several statistical procedures for analyzing the study's data will be described including inter-rater correlations, correlations between each task and total test score, and correlations between total scores and external criterion measures (e.g., grades; self-assessment scores). The study also involves assembling both internal and external review boards for the purpose of providing objective, professional feedback on the test items. The discussion of the study will include pedagogical implications and the impact of making the test widely to Korean teachers available online.

The development of an English proficiency test for teacher certification in Quebec

Beverly Baker, *McGill University* (beverly.a.baker@mail.mcgill.ca)

Avril Aitken, *Bishop's University, Quebec, Canada* (aitken@ubishops.ca)

Anne Hetherington, *Concordia University, Quebec, Canada*

(ahetherington@education.concordia.ca)

POSTER ABSTRACTS

All pre-service teachers in faculties of education in universities in Quebec are now required to demonstrate proficiency in their language of instruction (either English or French) as part of their certification by the Quebec Government's Ministère de l'éducation, du loisir et du sport (MELS).

This poster outlines the development activities of a common proficiency test for the English sector from 2008 to the present, undertaken by the three English universities of Quebec at the request of the MELS.

This poster presentation, in addition to outlining the steps in test development (from conception through piloting to validation and drafting of test specifications), addresses all four areas of this year's conference theme: learning, transparency, responsibility, and collaboration.

Learning: The integration of learning and testing has been paramount in test development activities. Test development has proceeded in parallel with discussions on course requirements for students, with the explicit goal of encouraging positive washback in the form of an increased focus on academic and professional English writing for educators. With demand for teachers very high in Quebec at the moment, the goal has not been to exclude students from university programs but to provide what is necessary for students' success.

Transparency: Test development was informed in the initial stages by information provided by many individuals involved in language testing for education students (including current test graders and administrators, student test takers, and those making decisions based on test results). In addition, a great deal of information has been provided to all these stakeholders during the development process. This presentation discusses the benefits received from this consultation, as well as the challenges involved in balancing these benefits with the additional time and effort required.

Responsibility and Collaboration: Many lessons are being learned in coordinating the input and (sometimes contradictory) opinions of test developers, administrators from three universities, representatives from a Quebec Government Ministry and certification body, and representatives from local English school boards. For example, different stakeholders--as might be predicted--have different goals for testing as well as differing conceptions of the construct of quality language for the teaching profession. Opinions also differ on power-sharing among these stakeholders: For example, which responsibilities for test administration should be at the university, school board, or ministry level?

During the poster session, comments and advice from test developers on these and other issues would be welcome and greatly appreciated, as development continues.

POSTER ABSTRACTS

The development of an on-line training and marking program for the assesment of writing

Annie Brown, *Ministry of Higher Education and Scientific Research*
(annieb7@gmail.com)

Language assessment is becoming increasingly dependent on technology, for both test delivery and for scoring. For performance assessments, on-line training and marking is an attractive proposition in contexts where large numbers of test candidates require the services of a large and dispersed cohort of markers, and where security and fast turn-around of results are imperative.

This poster will describe the development of an on-line rater training, accreditation and marking program used for the writing component of a school-leaving / university entrance exam, administered to around 32,000 students each year. The program has been in use for three years in this context. The presenter will describe and demonstrate the components and capabilities of the program, including marker registration, verification, training and accreditation, script delivery, score recording, and marker tracking. Scripts are batched by test version and are allocated to markers iteratively and on-demand in order to minimize marking time and ensure maximum completion of marking assignments. The presenter will discuss unexpected problems that arose in the implementation of the program and how the program developers dealt with them, thus ensuring that the program is fully responsive to the needs of the users. Finally, the presenter will discuss the results of a study into user satisfaction and the effectiveness of on-line rater training.

It is anticipated that this poster session will be of interest to other institutions who require the flexibility and speed in marking that such a program can offer, and are interested to learn about program capabilities and problems.

Issues in developing a proficiency-based assessment in multiple languages

Martyn Clark, *Center for Applied Second Language Studies* (martyn@uoregon.edu)

Proficiency has been the stated goal of many second language instruction programs since the introduction of the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines in the early 1980s. Yet, instruction in many classrooms continues to focus on formal knowledge of grammar and vocabulary, due to the difficulty in measuring proficiency. The Center for Applied Second Language Studies (CASLS) has cultivated partnerships with various organizations to create online foreign language proficiency tests in Arabic, Chinese, French, German, Hindi, Japanese, Persian, Spanish, Swahili, Urdu, and Yoruba. These tests are designed to be consistent with ACTFL proficiency guidelines and are intended to provide useful proficiency information to language instructors.

Over the past two years, CASLS has been refining the test design and exploring a new testlet-based multi-stage adaptive algorithm. This poster will highlight the

POSTER ABSTRACTS

development of the current test delivery system and the challenges involved in implementing a common test design in diverse languages with different populations. Results from pilot testing and simulation studies will be presented. Issues in aligning test scores with proficiency levels are also discussed. This project is being partially funded by a grant from the Fund for the Improvement of Postsecondary Education (FIPSE).

Construction of a proficiency identity within paired oral assessment: Insights from discourse

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

Anne Lazaraton, *University of Minnesota* (lazaratn@umn.edu)

The increasing popularity of the paired format in oral language testing has engendered legitimate scrutiny of its reliability and validity as compared with the more traditional interviewer-interviewee arrangement. Contradictory research findings have emerged; although characteristics such as the gender, cultural/L1 background, and language proficiency of one's interlocutor likely affects the discourse produced with a partner, the question remains whether this interlocutor effect influences scores on the test in any meaningful way.

In this poster presentation, the construct interlocutor effect is further examined. Initially, transcript data from one classroom and two large-scale speaking tests were analyzed by working backwards from testtaker scores to locate discourse features to support those scores. In the course of this analysis, identity formulations, such as "proficient" and "competent", as constructed in and through the discourse testtakers produce, emerged as a salient feature of the talk. Specifically, we posit that a testtaker brings a language proficiency identity (LPID) to a test task, and this identity is constructed, mediated, and displayed in the talk. We argue that "proficiency" is fluid, in that it will shift – on a turn-by-turn basis - based on who we are talking to in an L2 and what sort of identity(ies) we bring to and are mediated in that interaction. We present discourse data (with corresponding sound files) that suggest various ways in which testtakers can position themselves as 'proficient' by being, or, more precisely, "doing" "supportive" and "interactive" on the one hand, and positioning their partners as lacking in proficiency on the other. Implications for rating scale construction and rater training are suggested.

Creating valid in-house can-do descriptors for a listening course

Tomoko Fujita, *Tokai University* (tfujita@xa2.so-net.ne.jp)

This study attempts to create valid and appropriate can-do descriptors as class objectives for each level of a listening course in an English language programme. By relating students' work to the can-do descriptors, they will have a more concrete

POSTER ABSTRACTS

picture of what their listening skills are like, and this will lead them to be better self-evaluators. Students will also receive some ideas regarding what kind of concrete skills they will need to acquire if they want to go to the next level.

Approximately three thousand Japanese college students in the course are divided into three different levels; advanced (A), intermediate (I), and basic (B) according to the results of the in-house placement test. Initially, a need analysis was conducted, and the first draft of the can-do descriptors were written based on the reference level descriptions for CEFR, the European Language Portfolio, the International English Qualifications, STEP, and TOEIC. Ten teachers from the listening course formed a panel to examine the descriptors to see if they were appropriate for each level. This resulted in the completion of the can-do checklists along with 14 descriptors for each level.

Four hundred and twenty randomly chosen students from the three different levels responded to the statements on the checklists at the beginning (can-do 1) and the end (can-do 2) of the semester. First, the results were analyzed by applying 1 parameter IRT model and students' ability level (θ) for each level and were compared. Among three different levels, the students' average θ in the basic level improved the most, and that of advanced level improved the least. Secondly, the relationships between θ of can-do 1 & 2 were compared with their θ of an English general proficiency test (a grammar, a listening, and a reading subtest and a total score), and a listening final test. The correlations between the can-do 2 θ showed stronger relationships with a listening subtest (θ) and a listening final test (θ), whereas can-do 1 θ showed stronger relationships with a reading subtest (θ) and a total score (θ). These results may imply that students become better evaluators after taking listening classes because they have a clearer picture for each can-do descriptor.

The presenter will also discuss how revisions were made to the can-do checklists in order to make them level appropriate by utilizing the item difficulty parameter for each descriptor.

Examining collaboration in oral proficiency interviews

Gene Halleck, *Oklahoma State University* (gene.halleck@okstate.edu)

Much research has shown that the role of an interlocutor in an oral interview can be critical to the outcome of such a test (Brown, 2003; Brown & Lumley, 1997; McNamara, 1996; McNamara, Hill, & May, 2002; McNamara & Lumley, 1997; Reed & Halleck, 1996; Stansfield & Kenyon, 1992). Other researchers, noting the interactive nature of most oral interview discourse, have pointed out that accommodation plays a significant role in the co-constructed language of the interview and suggested that such accommodation could affect the validity of ratings (Malvern & Richards, 2002; McNamara and Lumley, 1997; Ross, 1992; Ross & Berwick, 1992). If the extent to which the interviewer accommodates to the language of the interviewee could have a significant effect on the candidate's proficiency rating (Malvern & Richards, 2002; Ross,

POSTER ABSTRACTS

1992; Ross & Berwick, 1992), then it might be useful to consider a way of eliciting language without the contribution of an interviewer. Although a number of researchers have advocated the use of role-plays for the collection of oral data (Cohen & Olshtain, 1994; Demeter, 2007; Halleck, 2003, 2007; Sasaki, 1998), some have argued that since monologic role-plays do not provide “opportunity for negotiation” (Rosendale, 1989) this may be a limitation of such a method for obtaining a valid sample of proficiency.

This poster reports results of a study that examined the discourse elicited by two types of role-play situations within oral interviews and raises the following questions: 1) Does the language behavior of the interviewer influence the ratable language elicited by the role-play and if so how? and 2) Does variation among interviewers (constructing dialogic role-plays) pose a threat to the validity and reliability of the role-play as an elicitation device?

This poster looks at the role that interlocutors play in two distinct types of role-plays: one in which there are two roles and the ensuing discourse is dialogic; and one in which the interviewer sets up the task but is not required to actually participate in the role-play when the situation is established, thus creating a more monologic situation. The analysis raises questions about the validity and reliability of dialogic role-plays as elicitation devices for the evaluation of oral proficiency because of the co-constructed or interactive nature of the discourse elicited.

The findings of the study indicate that inferences about an interviewee’s oral proficiency are more easily made when the role-play does not require the interaction of two interlocutors. This analysis suggests that to obtain a valid, ratable sample of discourse, a role-play requiring monologic rather than dialogic discourse may be more efficient and could provide a more reliable and valid sample for an accurate proficiency rating.

Using a reading-to-speak task to assess EFL oral performance

Heng-Tsung Danny Huang, *University of Texas at Austin*
(danny123@mail.utexas.edu)

Integrated tests consist of thematically-linked tasks where the supplied input constitutes the basis of the responses to be produced. Vis-à-vis their independent counterparts, such tests are claimed to feature a higher degree of authenticity, to equalize test-takers in terms of the information available for them to formulate argument, and to offer ideational and lexical input to facilitate idea generation. Drawing on these advantages, the current study further examined the issue as to how integrated speaking tests would impact the anxiety test-takers experience and the strategies they call in action. In this study, 47 EFL college students completed a reading-to-speak task and a speaking-only task. For the reading-to-speak task, they recorded their ideas pertaining to 2 oral prompts after reading a thematically-related article provided ahead of time. For the speaking-only task, they recorded their opinions on another 2 prompts without any prior input support. Immediately afterwards, they responded to a state anxiety questionnaire and a list of open-ended questions gleaning

POSTER ABSTRACTS

their anxiety levels, strategy use, and affective reactions. The paired t-test performed on the questionnaire revealed no significant difference between students' anxiety levels on the two tasks, demonstrating that the reading input did not impose additional anxiety on the students in an unreasonable manner. Qualitative analyses suggested that students relied principally on the reading support in tackling the reading-to-speak task while capitalizing largely on their prior world knowledge and note-taking skills in approaching the speaking-only task; moreover, a majority of them favorably perceived the reading-to-speak task for it aptly activated/supplied pertinent background knowledge for engaging the oral task. In light of these results, several pedagogical implications are proposed.

Developing a placement test for academic Arabic for upper high school and university students

Summer Loomis, *The University of Texas at Austin* (s.loomis1@gmail.com)

While the interest in Arabic language has grown exponentially in recent years, the area of Arabic language testing is still limited. With increasing enrollment and interest in expanding programs, there is a growing need for testing instruments designed to assess Arabic language learners' abilities. In particular, Arabic language instructors often must construct their own placement exams to evaluate students who wish to place out of the first two or four semesters of instruction. Programs like the Concordia Villages Arabic School and STARTTALK initiatives across the US will increasingly produce younger students who may wish to continue their Arabic learning at the university level. This poster focuses on the development of a placement test of academic Arabic for use in higher education Arabic language programs in the US context.

This placement exam is being designed for use in evaluating students' abilities to cope with formal Arabic for academic purposes, i.e. eventual advanced study of the language, learning subject material through the language, and engaging in and presenting research in an Arabic speaking context. As the Al-Kitaab textbook series from Georgetown is one of the most widely used in the US for Arabic instruction, this placement test is being designed around the learning objectives and progression of this series. The poster will include the theoretical basis for the design of the testing items as well as details on how the test will be piloted on a large scale. The test is a traditional paper based test and will include reading, listening, grammar and writing components. The poster will include pilot items from these areas and feedback from student testers.

POSTER ABSTRACTS

The distance between student writers' intentions of expression and teacher raters' expectations in argumentative essay rating

Lu Lu, *The University of Hong Kong* (cnu100037@yahoo.com)

So far the studies in the writing assessment area have mostly been focusing on the raters' decision-making process in essay rating. However, inadequate attention has been devoted to the discussions of the gap between student writers' intentions of expression and teacher raters' expectations in reading and grading the essays although such knowledge will surely exert positive effect on the students' learning about what they can do to strike a balance between their purposes and their readers' needs in the context of writing assessment.

This small-scale exploratory study will focus on two types of argumentative essays written by Chinese English-major undergraduate students: in-class timed exam essay and after-class homework essay. The researcher will investigate two research questions: (1) What are the convergences and the divergences between the students' and the teacher raters' inner criteria of persuasiveness of argumentative essays? (2) To what extent does the gap between student writers' purposes of expression and teacher raters' expectations influence the raters' grading of these essays?

Data collected for this study include 24 discourse-based interviews with the students whose essays were graded as the highest-rated (12 essays) and the lowest-rated (12 essays) in the two writing tasks respectively, as well as three Chinese teacher raters' verbal reports on the process of their assessing the 24 essays. The researcher also examined the students' and the raters' perceptions of the two argumentative writing tasks. The research findings reveal that the students and the teacher raters reached consensus on the well-formedness of texts, such as coherence and elaborations of discourse topic. In addition, the teacher raters catered more for the credibility of the writers and the evidence they provided in the argumentative essays, whereas the student writers seemed to favor persuading the raters in an exhortative tone. The researcher argues that some student writers' lack of the understanding of the raters' expectations is partly due to the opaqueness of the writing criteria, especially the teachers' clear interpretations of the writing criteria to the students. It is suggested that transparency of the student writers' expectations for expression and the teacher raters' expectations to each other will be useful for both students' learning argumentative writing and raters' argumentative essay evaluation.

POSTER ABSTRACTS

Assessing the language proficiency of Chinese nursing students and professionals: A brief introduction to METS (for nurses)

Kaizhou Luo, *Binhai College of Nankai University, METS Office*
(kevinlkz@hotmail.com)

Ting Huang, *METS Office*

Mariam Jones, *American Academic Alliance*

China's recent entry into the WTO coupled with high economic growth, and the subsequent and ongoing investment and improvement in health care infrastructure both in physical and human capital, has created an urgent demand for China's current and future healthcare professionals to develop a high level of English proficiency requisite to operating skillfully and safely within an English speaking context.

In order to address the above stated concerns, the Examination Authority of Chinese Education Ministry, in cooperation with the Chinese Medical Association and Association for International Exchange of Personnel, constructed the Medical English Test System (METS for Nurses) in 2006. METS' goals are to assess the English proficiency needed to meet the workplace communication of domestic nursing students as well as other healthcare professionals, produce the positive washback effect on teaching, learning, and curriculum reform of nursing English and English in other healthcare fields in China, and provide a worthy prototype for Chinese LSP testing. The four-level METS is a criterion-referenced test to show how test takers achieve in comparison to the proposed standards which are formulated in METS syllabus. Twice a year the test is administered by the National METS Office, a non-profit testing service located in Beijing. The written test has been conducted nationwide 4 times. The number of test candidates has increased significantly with each iteration of the existing test. To date, more than 84,560 nursing students and 13,260 working nurses have taken METS at the existing network of 194 accredited testing sites throughout China in the period from June 2007 to December 2008. It is commonly believed that communicative language activity can be divided into the receptive type, productive type, interactive type and medium type. METS will assess these communicative abilities in 2 subtests, written and spoken. The written test, which is currently delivered, comprises 5 components: listening, vocabulary and structure, reading, translation and writing. The computer-based spoken test is still being developed and tested. Currently, the METS language proficiency scales are aligned to the corresponding scales of Public English Test System (PETS) which is a recognized large-scale test administered directly by the Chinese Examination Authority. Upon passing both tests, a test taker will be awarded the National Comprehensive Certificate of Medical English.

METS stood poised to be a high-stakes test with its score being increasingly used as the measure of proficiency in English for nursing purposes both in the secondary/tertiary schools, and in the workplace such as granting diplomas before graduation, and seeking employment and advancement in a career. What is of critical importance to the successful completion and improvement of METS is close collaboration at each stage of test development. The METS Office is composed of testing specialists, medical experts, curriculum designers and marketing developers, in order to provide the platform

POSTER ABSTRACTS

necessary for evaluating inputs from various stakeholders. Thus far, the one of the key insights reached through such active collaboration among our experts is the fundamental nature of METS as a language test. Such a test focuses on language performance, and thus the rating criteria should represent language proficiency and not include non-linguistic factors such as content or background knowledge, although there exists a strong debate about the construct of LSP testing in this field.

METS is considered a highly reliable and valid test according to the statistics reported in METS Test and Score Manual (Aug 2007 Version). We regret that we cannot provide more detailed information as the study is yet to be published. Further investigations for reliability, validity and washback effect are still being conducted, and the presenter will provide major findings during the poster session. We welcome any professional suggestions on the research and development of METS.

Predicting item difficulty: A rubrics-based approach

David MacGregor, *Center for Applied Linguistics* (dmacgregor@cal.org)

Can item developers learn to accurately predict empirical item difficulty in a language test? In this study, we describe an attempt to use information about item difficulty to identify factors that may influence the difficulty, and to use this knowledge to help item developers write items that more closely reach the desired proficiency level. Items from a test of English language proficiency for grades K-12 based on standards developed by a multi-state consortium were analyzed using the Rasch model, and actual difficulty was compared to target difficulty in terms of six levels of proficiency defined by the standards. Items that met the target difficulty were analyzed for linguistic features that may influence their difficulty. Likewise, items that were shown empirically to be easier or more difficult than their target level were analyzed for linguistic features that may have contributed to their missing the target. Based on this analysis, a rubric was designed to help item reviewers judge how closely selected-response items fit the standards at their intended proficiency levels.

To examine the rubric's efficacy and potential as an item review tool, two raters were trained on the rubric and then individually rated the features of 59 operational test items. Raters' judgments were analyzed for consistency and investigated for the degree to which they successfully predicted the empirical difficulty of the items. The results suggest that researchers were successful in identifying features that contribute to item difficulty, and indicate that they were moderately successful in their attempt to train item developers to use the rubrics to evaluate items. These findings show how empirical data can be combined with qualitative analysis to evaluate and strengthen the development process of a standards-based test anchored in second language acquisition theory.

POSTER ABSTRACTS

An analysis of the sentence span task for L2 learners

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

It has been argued that individual differences in working memory capacity can account for language performance of various types, both among the native speakers (e.g., MacDonald et al., 1992) and the L2 learners (e.g., Harrington & Sawyer, 1992; Walter, 2004).

Assessing one's verbal working memory is a topic of active debate (e.g., Engle et al., 1999), and one of the methods commonly employed to capture this construct is known as the Sentence Span Task (Daneman & Carpenter, 1980). Reflecting a view that working memory resources must be simultaneously shared by temporal information storage and processing in such cognitively complex activities as language comprehension, this task imposes concurrent on-line requirements of reading semantically isolated sentences in sets of increasing size while maintaining in short-term memory all the sentence-final words in each set for a set-final recall. Research on this complex task has advanced mostly in the L1 context (e.g., Daneman & Tardiff, 1987; Waters & Caplan, 1996; Friedman & Miyake, 2005), and aside from establishing L1-L2 links in working memory and their links with other skills among high proficiency learners (Osaka & Osaka, 1992; Miyake & Friedman, 1998; Fortkemp, 1999), published work involving L2 measures of Sentence Span have revealed little on the inner workings of such measures. The continued interests in the role of working memory in L2 processes (e.g., Koda, 2005) and its obvious significance in L2 performance at any level of proficiency in both test and non-test contexts must justify more in-depth analyses of the L2 versions of the Sentence Span Task.

This presentation reports on a data from an ESL version of the Sentence Span Task as it was employed with a group of 236 Japanese learners of English representing a wide range of proficiency. The analyses focus on the complex relationships between such task variables as the sentence length, lexical density, sentence presentation order, total set length, and recall word difficulty on the one hand and such performance outcomes as word recall and sentence reading time on the other. Implications for future designs and uses of working memory span measures with L2 learners, particularly of lower proficiency, are discussed.

Selling a language assessment to skeptical stakeholders: The GSLPA

Alan Urmston, *Hong Kong Polytechnic University* (ecalanu@polyu.edu.hk)

Felicia Fang, *Hong Kong Polytechnic University*

Since its implementation in 2005 as a compulsory English exit test for all students at the Hong Kong Polytechnic University, the Graduating Students' Language Proficiency Assessment (GSLPA) has attracted increasingly greater attention from local students, employers, EFL teachers, language testing experts and researchers. The Assessment is currently administered twice a year to a population of approximately 3000 students.

POSTER ABSTRACTS

The idea of having a workplace-related English assessment as a university exit test has been regarded as appropriate (Lumley & Qian, 2003). However, in Hong Kong, students in their final year at university are encouraged by the government to take IELTS under the Common English Proficiency Assessment Scheme (CEPAS). Hence, there has been pressure on the Polytechnic University to justify the use of the GSLPA to some stakeholders and to promote it to others, who perhaps cannot see beyond IELTS as a measure of English language proficiency. The GSLPA Team has gone about this by initially ensuring that the Assessment itself (separate tests of written and spoken English) is of a high quality and conforms to recognised standards of language assessment. The Team has also invested a great deal of time and effort in promoting the Assessment to help to establish it within the educational and employment communities in Hong Kong. It has done this through establishing international recognition and by working with employers and employers' organizations to effectively 'sell' the GSLPA to them. Clearly, without stakeholder acceptance, and of course this includes the students, the Assessment will not survive, especially when faced with such high-profile opposition.

This poster presentation will outline the measures taken to ensure the suitability of the GSLPA to provide an accurate and reliable statement of the English proficiency of graduating students in a workplace context. Where appropriate, empirical data is provided as illustration. The presentation will also describe how the Assessment has been promoted to students, employers and others in a continued campaign to enhance its face validity and profile in a competitive market. It is anticipated that fellow assessment providers will find the presentation an interesting case study of how to market an assessment that fulfils a specific role when faced with the challenge of more well-known, mass-marketed, and more generic assessments.

Testifying in legal cases: Test your knowledge, flexibility, and creativity!

Margaret van Naerssen, *Immaculata University*
(margaret.vannaerssen@gmail.com)

How would you handle language proficiency assessment in a legal case? Of course, "it all depends." However, the factors to consider are somewhat different than when doing large-scale testing or institutional placement assessment.

It is not enough for language assessment experts to report levels or scores on a language assessment protocol. Language assessment experts, linguists, attorneys, and judges need to see beyond the numbers and labels.

This poster gives viewers a chance to learn more about the issues, test out their ideas, and add their expertise to this area of forensic linguistics. It focuses on the use of language proficiency assessment data in individual, non-native speaker cases. This might involve assessment data used in conjunction with data in existing language evidence. This might involve data added to assist fact-finders in the evaluation language proficiency issues.

POSTER ABSTRACTS

In an interactive poster format, viewers test themselves at various levels of involvement. First, they can read an overview statement of the need for principled language proficiency assessment in legal cases. Second, they can read brief definitions of language samples, language evidence and linguistic evidence. Third, they can read a list of somewhat unique challenges in the language proficiency assessment of suspects/witnesses/defendants. Each challenge has a “pull-out” tab with a few elaborating phrases. Fourth, they can read cards of three sample cases. Fifth, on the other side of each case card, they can choose to read a proposed (or actual) strategy and a few evaluative comments. Sixth, for cases that interest them, they can provide possible strategies additional factors to consider. This can be done in discussion with the poster presenter and/ or in notes left in the “pocket” next to the case (adding contact information if they wish). Seventh, in another pocket they can leave other comments on concerns or experience (again adding contact information). Writing supplies are provided.

In recent years there has been an increase in courts of non-native speakers of the primary language of a legal system and recognition of special issues. Language assessment findings are increasingly being introduced. Thus, they also are increasingly becoming targets of challenges. Those concerned with the use of language assessment need to be better prepared with variations in responsibilities: in making informed decisions in the choice and use of particular protocols, in evaluating choices of other testers, and in evaluating own findings and those of others.

The poster does not address court challenges involving theoretical or technical issues in large-scale language testing. It also does not address proficiency testing of court or other government interpreters or language specialists.

A test to reveal incremental acquisition of vocabulary knowledge

JoDee Walters, *Bilkent University* (walters@bilkent.edu.tr)

Many tests of vocabulary look at vocabulary knowledge from the narrow perspective of being able to match form with meaning. A test of receptive vocabulary may ask test takers to match one of several definitions to a given word, or vice versa. A test of productive vocabulary may ask test takers to produce the appropriate word, given a definition, or to insert an appropriate word in a given sentence. Such tests ignore the fact that word knowledge is comprised of several aspects, including the recognition and production of the form of the word (written or spoken), and knowledge of its grammatical properties, collocations, associations, and register, all in addition to knowledge of the meaning of the word. Traditional vocabulary tests also do not acknowledge the fact that vocabulary knowledge is acquired incrementally. Learners are unlikely to acquire all of the above aspects of a word at one time; it is more likely that their knowledge of vocabulary words will grow gradually, with each exposure adding to what is known.

In the context of another research project, a test of vocabulary knowledge was needed, to provide information about vocabulary learned while reading (incidental vocabulary

POSTER ABSTRACTS

acquisition). Given that traditional vocabulary tests do not allow test takers to show all of the aspects of word knowledge they possess, and do not allow them to demonstrate partial knowledge, a new kind of vocabulary test was required. This poster describes the development of this test.

The test is designed to reflect all aspects of word knowledge that might be acquired while reading, without the aid of a dictionary. The aspects tested include recognition of the word, word form (i.e. spelling), grammatical information (i.e. part of speech), collocations, associations, and meaning, from partial to precise. The test is paper-based, and consists of a combination of selected-response and short-answer questions. The test was piloted, and the items were analyzed, with analysis focusing specifically on the behavior of different test items for the same target words. In addition, the results of the pilot test were checked for validity by interviewing selected test takers regarding their knowledge of the target words. This interview followed the format of the Vocabulary Knowledge Scale, a test usually used to measure depth of vocabulary.

The resulting test format is believed to be useful in providing a more complete and accurate picture of vocabulary knowledge. Such a test format may be useful for classroom teachers, textbook and materials writers, and researchers – in short, anyone who is interested in more than a meaning-focused, all-or-nothing glimpse of vocabulary knowledge.

Response strategies and performance on an integrated writing test

Hui-chun Yang, *University of Texas Austin* (hcyang@mail.utexas.edu)

Writing strategies are the conscious tools writers use to explore ideas, conceptualize thoughts and organize formats of presentation. A number of studies have examined the effects of strategy use on writing, however, little research has been conducted in the assessment context, particularly the second language writing assessment context. Moreover, very few studies thus far have explored test-takers' strategy use in a writing test in which reading, listening, and writing are combined. This study aims to investigate the relationship of test-takers' use of writing and test-taking strategies to their writing performance in the integrated writing test that requires them to incorporate into their essays relevant information from a reading passage and a lecture. Both quantitative and qualitative data analyses were employed for this study. The 150 international students enrolled in University of Texas at Austin were recruited to take a reading-listening-to-write test, followed by the Integrated Writing Strategy Inventory (IWSI) on how they thought while completing the test. Ten voluntary students, five from the high achievement group and five from the low achievement group, were selected for follow-up interviews. Exploratory factor analysis (EFA) was used to identify the link between strategy use (latent variables) and student essays (measured variables). Drawing upon previous writing research, confirmatory factor analysis (CFA) was utilized to test the hypothetical relations between observed and latent variables. A structural equation modeling (SEM) was used to model relationships between students' self-reported writing strategy use and their writing performance.

POSTER ABSTRACTS

The data collected from follow-up interviews were analyzed to provide supplementary information in interpreting the quantitative data. The results of the study have implications for second language writing assessment and instruction.

The effect of different anchor tests on equating quality

Kiyomi Yoshizawa, *Kansai University* (yoshizaw@ipcku.kansai-u.ac.jp)

The present study is conducted in a testing situation where common internal anchor items cannot be used for linking different forms of a language proficiency test used for university admission purposes. These test forms are mainly based on reading passages and written to the same specifications to assess whether test takers are proficient in English to take English language classes offered by the university.

The present study examines how different anchor items would affect equating quality. Test authenticity has been emphasized in language tests. Concerning reading tests, it is usually the case that tests consist of reading passages and comprehension questions. Test takers are asked to read several passages and respond to the items based on specific passages respectively. However, when internal anchor items cannot be used, different test forms can be linked, using external anchor items (Wright, 1977). The present study examines how different anchor items would affect equating quality. To this end, two forms (Forms B and Y) were selected; each form consists of three sections with a total of 55 items. 1,000 test takers were randomly selected from all the test takers who took each form respectively. A 50-item anchor test was constructed: 28 items from Form Y Sections I and III and 22 items from Form B Section

II. The anchor test was administered to 176 English language learners with equivalent level of proficiency. Rasch measurement model was used to calibrate the items on a common scale. Concurrent equating design was applied. The items on the anchor test were selected in five different ways in terms of passage length and item types:

- (1) Test 1: all the items are included;
- (2) Test 2: items in I (Y) and II (B);
- (3) Test 3: items in III (Y) and II (B);
- (4) Test 4: items in three sections asking for understanding specific parts of a text;
- (5) Test 5: items in three sections reading for over-all understanding of a text

Test 2 and Test 3 differ in the items from Form Y. Sections I and III differ in passage length, but the type of the question items are equivalent. Test 4 contains the items from all the three sections, which ask test-takers to understand specific parts of a text. Test 5 contains the items from all the three sections, which ask test-takers to read for main idea, topic, the author's point of view, attitudes, tone of a passage, or title.

Results of equating quality and practical implications will be presented.

POSTER ABSTRACTS

A retrospection study on the construct validity of TOEFL listening comprehension tests with multiple-choice format

Yujing Zheng, *Chongqing Jiaotong University* (nelliezyj@yahoo.com.cn)

Xiangdong Gu, *Chongqing University* (xiangdonggu@263.net)

The construct validity of a certain test is determined by two factors: the right constructs it measures and the test-method effects it poses on test-takers' performance. As for listening comprehension tests, the multiple-choice (MC) format is the most widely-used test format, however, a review of literature suggests that findings of the construct validity of TOEFL listening comprehension tests with MC format are rather controversial, mainly due to the unclearness of (1) what constructs the tests really measure, and (2) what effects the MC format really has on test-takers' performance.

This study intends to clarify these two questions, thus to shed light on the construct validity of the tests. The retrospective verbal report methodology is employed to examine the test-taking processes. Qualitative data are collected by asking eight Chinese EFL students with seven years of English learning experience at Chongqing University in China to verbalize their test-taking processes immediately after they answer the MC items adapted from the test papers of TOEFL in January 1998 and October 2000, and then to answer the researcher's questions in the later retrospective interview. The transcribed verbal report data are analyzed according to the framework of listening constructs proposed by Buck (2001). The constructs tested by the tests and the specific effects of the MC format on test-takers' performance are identified.

The research findings show that compared to the framework of listening constructs proposed by Buck (2001), all the listening constructs, including language competence and strategic competence as well as their sub-categories of grammatical knowledge, discourse knowledge, pragmatic knowledge, sociolinguistic knowledge, cognitive strategies and metacognitive strategies are all tested. It indicates that the constructs engaged by the test-takers are those required by the constructs definitions, that is to say, the tests are valid in that they are testing the constructs they intended to measure. Meanwhile, the background knowledge, which the tests didn't intend to measure is actually tested. This contaminates the construct validity of the tests. The MC method affects test-takers' performance in the ways of superficial matching, reading and uninformed guessing, which also pose threats to the construct validity of the tests.

The study offers some implications for the design of MC test items. One tentative suggestion is to avoid overlapping words between listening text and the item options. The original words or phrases should be disguised through the use of paraphrase when constructing the MC items. Besides, the effects of reading skills on listening comprehension test performance may be lessened by using lower-level words or simple syntactic structures in the MC item options. From the methodological perspective, the study develops a data collection procedure combining the immediate retrospection with retrospective interview in order to examine more thoroughly the mental thoughts of test-takers. The method provides useful data for the study, which may offer some insights for the application of it to construct validation studies in the future.

Works-in-Progress

Time: Thursday, March 19, 2:00 – 3:30 p.m. Location: Atrium

Native versus non-native raters' evaluations of high-level English

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

Unlike many language testing organizations' speaking testers, the Federal Bureau of Investigation's oral test raters are required to have speaking ability equivalent to a highly articulate, well-educated native speakers (WENS), defined in the Interagency Language Roundtable's (ILR) Skill Level Descriptions (SLD) Level 5. Whether these requirements are necessary for reliable rating, and whether ILR Level 5 status is achievable by non-natives (Coppieters, 1987; Sorace & Robertson, 2001) is debated in the literature. Previous research is inconclusive, with some studies showing discrepancies between native and non-native raters (Hoyt & Kerns, 1999), and others not finding significant differences (Ludwig, 1982). Research to date has investigated this issue largely within the academic context (Davies, 2003; Paikeday, 1985; Hamp-Lyons & Davies, 2008) where achievement rarely exceeds ILR Level 3; whereas, Federal Government contexts often deal with high-level speech, ILR Levels 3-5 (Clark & Clifford, 1988).

This study examines judgments and observations made by native speaker versus near-native speaker trained raters (n=40) of high-level English speech. Structural equation modeling, including profile analyses and MANCOVAs, of quantitative and qualitative data from test scores, tester reports, notes, and think-aloud protocols triangulate data of rater groups' assessments and justifications, controlling for various rater variables (Reed & Cohen, 2001). Research questions investigate whether rater groups a) assign significantly different scores overall and across rating factors, b) focus on different language features c) are affected by other rater variables, such as proficiency and experience.

Preliminary results indicate non-significant differences between rater groups for individual factors and overall scores; however, with English ability as a covariate, profile analyses for the think aloud protocols reveal the native and non-native groups' evaluations differ significantly across rating factors. The research is intended to inform policy on oral proficiency rater selection and address theoretical issues of ultimate attainment.

WORKS-IN-PROGRESS ABSTRACTS

Construct validation of listening comprehension test: effects of task and the relationship between listeners' cognitive awareness and proficiency in listening comprehension

Youngshin Chi, *University of Illinois at Urbana-Champaign* (ychi1@illinois.edu)

The working paper investigates construct validation of listening comprehension test and measures relationship between test takers' cognitive awareness of listening process especially listening breakdown when they are listening academic lectures and their proficiency. Second language learners use various strategies to process both aural and visual information however they also experience comprehension breakdown due to fast speech, redundant words or unfamiliar context. We implement mixed methods to collect evidence of test takers' listening proficiency and their cognitive process. First, listening tasks were developed based on the outcomes of need analysis. Several public lectures were examined to find the possible comprehension breakdown while listeners are listening to the lectures. Second, a listening test which includes breakdown tasks will be administered to students in a university. Task difficulty and test takers' listening comprehension will be measured in quantitative way. The study will explore the potential relationship between task effect and cognitive awareness in listening comprehension.

Bridging the gap: Assessment for learning in a Quebec classroom

Christian Colby-Kelly, *McGill University* (d.colby@elf.mcgill.ca)

Recent educational reform in Quebec, Canada, has incorporated some elements of the assessment for learning (AFL) approach to learning in second language ESL classrooms there, yet a gap remains between the philosophical considerations of the method and actual classroom practice. Many researchers (Black, 2005; Blumfeld, 1992; Cowie, 2005; James & Gipps, 1998; Leung, 2005; McNamara, 1995) in the field of educational measurement have noted a trend towards using assessment to foster learning. Leung and Mohan (2004) declared "There is now widely recognized support for classroom-based formative teacher assessment of student performance as a pedagogically desirable approach to assessment which is capable of promoting learning" (p. 335). The AFL approach has become a part of state educational policy in England, Northern Ireland, and Wales; in Scotland; in Australia; New Zealand; and recently in Hong Kong. Locally in Quebec, while formative assessment practices are in wide use in L2 classrooms, they do not extend to the dynamic, learner-oriented methodology of AFL. Colby-Kelly and Turner's (2007) baseline study of formative assessment practices in a Quebec, pre-university English for Academic Purposes (EAP) classroom setting, found evidence of teacher-engagement in the assessment bridge (AB), which they defined, drawing on the work of the Assessment Reform Group (2002), Bachman (2004), and Lynch (2001), as the area of classroom-based assessment encompassing the fusion of assessment (where learners are in their learning), teaching (where they need to go and how best to get there) and learning (action on the part of the learner). The present

WORKS-IN-PROGRESS ABSTRACTS

research investigates evidence of the AB and the effects of introducing AFL procedures in two Quebec pre-university EAP classes. Thus, the study focuses on the learners' attempts to master the contingent use modal form 'would' in obligatory contexts, adapting an AFL methodology using (1) computer assisted learning (CAL) including self-assessment, concept mapping, and reflective self-questioning components, (2) a group reflective, peer-assessment, knowledge-construction activity, and (3) a teacher-class guided questioning and scaffolding formative assessment activity. This research takes an interactive and dynamic assessment approach, adopting a Vygotskian, sociocultural perspective to the co-construction of learning. It takes a mixed-methods approach, incorporating quantitative and qualitative measures of case studies of two teachers' use of AFL in classroom settings. The data will be analyzed and triangulated. The study participants are two randomly assigned Advanced-level EAP classes. Quantitative measures include pre- and post-treatment fill-in-the-blanks tests, frequency counts of pre- and post-treatment written essays, a CAL instrument, and group-constructed learner written production. Qualitative measures include pre- and post-treatment teacher and student questionnaires to measure student and teacher perceptions on assessment and learning. Baseline, pre-, during-, post- and delayed post-treatment data will also be collected. The learning materials for the following self-, peer- and teacher-class assessment and knowledge construction activities will be presented using a mock mystery theme with the learners challenged to use metacognition to solve a language mystery. The results of the present research will shed light on our understanding of the effects of AFL procedures on learning in an L2 classroom assessment setting, and on evidence of the AB. In addition, it will inform learner metacognition as students grapple to decode second language usage conventions, all the while minding the gap.

Oral proficiency assessment in a pre-service EFL teacher education programme: transparency on the validation of vocabulary descriptors

Douglas Altamiro Consolo, *UNESP* (dconsolo@terra.com.br)

Melissa Baffi-Bonvino, *UNESP* (melissabb7@hotmail.com)

In this presentation we report on a research project initiated in 2005 which aims at investigating criteria for 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 whole study comprises two phases. During phase 1, the study was carried out with the purpose of investigating class work on vocabulary and the assessment of the students' oral production by focusing on oral language data produced in class (in seminar presentations) and in two oral tests, the FCE speaking paper and the TEPOLI (Test of Oral Proficiency in English; Consolo, 2004; Consolo & Teixeira da Silva, 2007). The role of students' expectations and views about the processes of teaching and learning vocabulary in class, aiming at oral language production, was also considered in the first phase. For the assessment of lexical proficiency in oral test situations, four students, who participated in all the occasions for data collecting in phase 1, took two

WORKS-IN-PROGRESS ABSTRACTS

mock FCE speaking tests during the academic year and the TEPOLI, which has been designed and implemented as a test for EFL teachers and teachers-to-be, at the end of the course. A quantitative analysis of the lexis produced in the oral tests was conducted with the help of the RANGE programme, and the overall results from the first phase indicate that students displayed better lexical competence and that vocabulary teaching in speaking tasks reflected positively in the quality of students' oral performance. Similarities in the levels of proficiency were obtained when results from FCE and from TEPOLI were compared. In the second phase of the study, eight students from another class of students graduating from the same course have contributed with data, and were assessed on five different occasions: oral language produced in class (seminar presentations), two mock speaking FCE tests, the speaking paper from IELTS and the TEPOLI. The multiple methods used to collect data include, apart from the three different oral tests, individual interviews, questionnaires, and recordings from the EFL lessons with the class. In phase 2, previous to the occasions when tests were administered, students were also better informed about the tests they would take, as well as about the rating scales to be used in those tests. Such procedure was adopted (a) based on opinions given by the students who participated in phase 1 and (b) with the purpose of establishing a policy of more transparency about criteria for language assessment in the course of the study. The larger amount of data obtained in the second phase of the study, including expectations and perceptions of the class teacher, oral testers and test takers involved in the investigation reinforces the importance of developing linguistic-communicative competence during pre-service teacher education, and the relevance of pedagogical work with focus on vocabulary and oral proficiency in this context. It is also a central purpose of the study to analyze the connections between vocabulary and oral proficiency during the oral assessment process, based on graduating students' language in oral test situations per se, and also on the language produced by these students according to the assessment criteria of the tests taken, mainly with regards to the criteria to assess lexical proficiency by means of the given tests. We will report on the results of the data analyzed to date and discuss the importance of the study as part of the process to validate the vocabulary descriptors of TEPOLI, which encompasses the contributions from both phases and supports the definition of more objective criteria to assess lexical oral proficiency in English with the necessary transparency needed in testing speaking proficiency in EFL at the end of Letters courses.

Factors influencing the pragmatic development Korean study-abroad learners Sumi Han, *Seoul National University* (sumihan8@snu.ac.kr)

In recent years, more and more students have taken part in study-abroad (SA) programs sponsored by colleges and universities. Through SA programs, EFL participants in particular are eager to experience another culture and improve their second language. In this respect, SA programs are popular and considered beneficial to learning a second language and have attracted the interest of language researchers (DuFon, 1999; Hoffman-Hicks, 2000; Kasper & Rose, 2002). Not only are examined the

WORKS-IN-PROGRESS ABSTRACTS

overall language improvement of the participants in student exchange SA programs, but also the effects of studying abroad on the improvement of learners' pragmatic competence as cultural awareness (Barron 2003; Cohen & Shively 2007; Kondo 1997). However, relatively few researchers have attempted to investigate factors, which may impact the change of L2 pragmatic ability of SA learners, in a single study. Thus, there are few efforts to evaluate SA programs comprehensively or systematically, which would then undermine the results of studying abroad in pragmatic learning.

This study aims to contribute to the current body of research by investigating the correlation between five factors - (a) studying abroad itself, (b) an explicit instruction of pragmatic knowledge before SA, (c) language proficiency, (d) motivation, and (e) attitude towards a target community - with advanced EFL learners' change in pragmatic competence after a one-year SA. The subjects in the current study are: 120 Korean undergraduates enrolled at a four-year university in Seoul, Korea, almost at the same age (from 20 to 23 years old) and with various majors. 60 selected students as the experimental group participate in one-year SA programs at major universities in New Zealand; the other 60 of the at-home control group continue to study English in Korea. Concerning data collection, all the subjects fill out a 50-item questionnaire for the two factors (d) and (e), and take a TEPS as a language proficiency test for the factor (c), twice before and after the program. Furthermore, half of the experimental group as the SA experimental group will get an explicit instruction of pragmatic knowledge during 5 2-hour sessions before the program, in order to enhance their awareness about the target-language pragmatics. Finally, a modified discourse completion test (Hudson, Detmer & Brown, 1992, 1995) will be constructed and piloted. Its final version will be divided into two tests (pretest-posttest design) to measure the subjects' pragmatic competence before and after SA, including the control group in Korea.

Descriptive statistics will be computed for the questionnaires, the modified discourse completion tests, and the TEPS scores for the groups. Concerning the 5 factors influencing the pragmatic competence, each item in the questionnaires and test scores will be compared and examined via a series of t-tests for the variables. So a matrix of correlations will be obtained among different groups: between the whole experimental group and the control group in Korea, and between the SA experimental group and the SA control group. The correlation between each factor and the pragmatic development of the SA learners will also be inspected to determine which factor is more crucial to the pragmatic development in SA by capitalizing on multiple regression analysis.

It is expected that results of this longitudinal study will reveal the degree to which the five factors are related to the development of the pragmatic competence of the EFL Korean SA learners. The data would be used to identify what difficulties advanced EFL learners have with acquiring the target language pragmatics. The study results are also expected to provide very valuable information as to design ideal before- and after-courses in SA programs, thereby improving the effectiveness of such programs overall.

WORKS-IN-PROGRESS ABSTRACTS

Learning outcomes and the focus of the assessment tool

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

Weiping Wu, *The Chinese University of Hong Kong* (wwpplc@cuhk.edu.hk)

This paper reports on the findings and related analyses based on data collected in a 5-year (2004-2008) study, focusing on the learning outcome as indicated by the Computerized Oral Proficiency Assessment (COPA), a tool to assess the oral proficiency of learners with the sample collecting approach as used by the Simulated Oral Proficiency Interview (SOPI) developed by the Center for Applied Linguistics (CAL). Scores for 392 participants, who were heritage students enrolled in the program of Chinese as a Second Language and took the COPA on a voluntary basis before they exited the program, were used in the study.

Instead of focusing on the correctness of language structure, the COPA was designed to encourage learners to communicate in simulated real life situations. Each participant was required to produce 12 speech samples randomly selected from a pool of 600 some tasks, which were organized at 3 proficiency levels (Intermediate, Advanced and Superior) according to the ACTFL Proficiency Guidelines. All individual samples were then rated by a team of 3 certified

COPA raters. The finalized rating for each participant was assigned a numerical value on a 1-8 scale, where 1 stands for the lowest proficiency level while 8 the highest, so that differences in leaning outcome over the years could be compared and analyzed. Attempts have been made to find answers for two key research questions:

1. What are the differences, if any, in the speaking samples collected over the years with special reference to the level of proficiency in oral communication?
2. What are the characteristics of the students each year and whether language gain based on such characteristics are statistically significant?

Findings from the study, which show a significant gain in proficiency level as indicated by the results of the test over the years, suggest that the principle of testing “for teaching” rather than “of teaching” is indeed applicable if both teaching and testing serve the same purpose. Specific examples related to the design of the test, including the authenticity of the tasks, organization of the task bank and precautions in randomly generating the test by the computer are also discussed as part of the paper.

This “work in progress” presentation describes the development of a telephone interview protocol used in the English for Heritage Language Speakers (EHLS) program. The EHLS program gives native speakers of languages other than English the opportunity to achieve professional proficiency in English and thus increase their marketability with the U.S. government. To complete the program successfully, participants must have advanced proficiency in English at entry.

Previously, each program applicant submitted a detailed written application; selected applicants then participated in English language testing, and then final selections were made for the program. For the 2009 application process, a 15-minute telephone interview of each applicant was added to the procedure in order to increase the information available to inform the selection of provisionally accepted candidates.

WORKS-IN-PROGRESS ABSTRACTS

Development of the telephone interview involved creation of an interview protocol containing a series of questions designed to allow applicants to demonstrate their language proficiency, and an evaluation rubric designed to evaluate each applicant's overall English proficiency and level of motivation. This "work in progress" session describes how the protocol and rubric were developed; discusses issues that were identified; and provides an informal evaluation of the efficacy of such an assessment tool. The research questions we ask concern the relationship between the phone interview ratings and the fact of being selected for further testing (through formal OPIs), the subsequent OPI scores, and the fact of final selection into the program. The degree of the relationship between the phone interview ratings and the candidates' further performance is tested using correlation analysis. We hypothesize that the ratings of the phone protocol would generally predict applicants' further success in the selection process. The findings of our research seem to support our initial hypothesis; however, data analysis also uncovered additional considerations to be taken into account for future uses of the protocol and rubric.

Obtaining an opportunity to present this work in progress would be highly desirable for our project, as the feedback from the LTRC audience would help us investigate answers to some crucial questions uncovered in the process of data analysis and finalize our findings accordingly.

Our presentation provides a small-scale, but comprehensive, illustration of the theme of the conference. Developing a phone interview protocol and rubric to be used in the EHLS program was a highly collaborative task focused on increasing levels of transparency and accountability behind the participant selection process. Previous research and practices utilized by the EHLS program provided a basis for further learning and suggested directions of potential improvements for the program's candidate selection and testing processes.

Computer-based and internet-delivered College English Test in China: IB-CET in progress

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

This paper briefly introduces the process of designing, developing and trialing of the computer-based and internet-delivered college English test system in China. It consists of five parts with part I--an introduction to reformed CET, part II--the theoretical bases and principles for design and development, part III--the elaboration of its components, part IV--the testing environment and part V the implementation and its social effect. It concludes that the computer-based language testing, as the large-scale test CET, will become more practical in China and will stand a promising future. The paper also list some difficulties IB-CET are facing and hopefully seeks certain solutions from colleagues in the international level.

WORKS-IN-PROGRESS ABSTRACTS

Reform from within: A collaborative effort at transparency, reliability and validity in assessment

Claudia Kunschak, *Shantou University* (claudia.kunschak@azalumni.com)

With changes in curriculum come changes in assessment. In the case of an English Language Center (ELC) at a medium-sized public university in Southern China, a 7-level integrated program focusing on communicative competence and based on a North-American textbook series adapted for the Chinese market was instituted in 2003. Teachers were trained in using the textbook and in collaborating in designing quizzes, midterms and finals to measure achievement. They also collaborated in assembling and editing a student companion series to the textbook. However, few teachers at this center have a background in testing and the high turnover of part of the faculty - the international teachers who constitute about 30-40% percent of overall teaching staff - have consistently posed challenges to the consistency of program delivery across and within levels. The resulting student complaints about tests and scores lead the presenter to design a project in order to investigate the issue. Unlike large-scale commercial testing centers, which can afford field testing and regular validity and reliability checks, a small-sized university language center cannot afford a separate evaluation of its testing program. Thus, several faculty formed an assessment interest group and conducted an action research project to harmonize assessment practices at the center. Both test design and rater behavior were examined in order to check and if necessary improve reliability and validity of the respective instruments, processes and results. The presentation will introduce the main components of the project (research design, instruments, team members), its effectiveness in influencing attitudes to, knowledge of, and compliance with good testing and rating practice at ELC, as well as any changes in score distribution. Emphasis will be placed on the bottom-up nature of the project which lends itself to being replicated in other contexts, the cyclical nature of the process if it is to have a long-term effect, and the multiple factors that have contributed to the project's success.

ESL writing assessment: Does the selection of rating scale matter?

Chih-Kai (Cary) Lin, *Georgetown University* (kaicary@gmail.com)

This study aims to discuss the effectiveness of two rating scales designed to assess ESL (English as a second language) writing in a pedagogical context. Comparison between holistic rating scale and multiple-trait rating scale is made in terms of inter-rater reliability and perspectives elicited from raters. All raters are ESL practitioners at a large Midwest university in North America. Both rating scales are implemented for diagnostic purposes on which subsequent ESL writing units would be based.

An argument for a pedagogical preference of multiple-trait rating scale over holistic rating scale is drawn from various sources, such as the distinctiveness of L2 writing, the extent to which criteria exemplified in both rating scales resemble writing tasks that students are asked to do in the specific institutional context and the degree to which

WORKS-IN-PROGRESS ABSTRACTS

both rating scales inform teaching and learning. In addition, a statistical analysis reveals that the multiple-trait rating scale is more likely to yield a higher inter-rater reliability than the holistic rating scale.

One preliminary implication is that by adopting multiple-trait rating scales in grading L2 writing, language teachers could minimize the possible bias of placing greater scoring weight on one aspect of writing over another. In other words, if teachers could give discrete assessment to each writing component desired by their institutional demands, their ratings are more likely to adequately address the distinctive features of L2 writing. Discussion will also touch on a potential study examining the interaction between the two rating scales and novice/experience raters.

The Internet-Based TOEFL and test user beliefs

Margaret E. Malone, *Center for Applied Linguistics* (meg@cal.org)

Megan Montee, *Center for Applied Linguistics* (mmontee@cal.org)

The internet-based Test of English as a Foreign Language (iBT) presents examinees with integrated tasks requiring the synthesis of multiple skills. In doing so, it endeavors to simulate the requirements of English-medium universities. However, because the iBT has been in use for only a short time, little research has been conducted to determine whether this intention is borne out in test users' self-reported beliefs. This work-in-progress session will describe a project which investigates this further.

Data will be collected from students at the undergraduate and graduate levels, teachers, and administrators on their beliefs about the iBT as a measure of academic language ability. Research methods include focus groups, surveys and stimulated recalls, and participants include individuals in Germany, Korea, Saudi Arabia, and the United States. This session will present the preliminary results from focus group participants (N=66) and Internet-based surveys (N= 1,250), and how these results will inform the design of the stimulated recall methodology. The session will also include information about the purpose of the study, background literature, the research questions, a description of the study participants, data collection procedures, and preliminary results and data analysis.

The intent of this study is to provide insight into the extent to which users understand and agree with the construct of academic language that underlies iBT tasks (e.g., what users believe the test is testing), and the extent to which the beliefs of users abroad differ from those of users in the United States. It will extend prior research to a wider group of educators by collecting information from participants in Germany, Korea, and Saudi Arabia and by including U.S. university administrators in the study. During the session, presenters will solicit feedback on the design and implementation of the study, preliminary data, and the implications of the project.

WORKS-IN-PROGRESS ABSTRACTS

Developing new lexical measures for diagnostic assessment

John Read, *University of Auckland* (ja.read@auckland.ac.nz)

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

The current interest in diagnosis in language assessment (eg, Alderson, 2005) is giving fresh impetus to the development of measures of language knowledge, which have been neglected to some degree over the last 30 years through the dominance of the communicative paradigm. Lexical measures in particular have been shown to function well not only to produce estimates of vocabulary size but also as general indicators of proficiency level, for purposes such as placement of learners in a language teaching program and for vocabulary acquisition research. The simplest kind of vocabulary measure is the Yes/No format, in which test-takers indicate whether they know each of a sample of target words. A proportion of the items are not real words in the target language, so that the scores of those who tend to overestimate their vocabulary knowledge can be adjusted accordingly. Several recent research studies have investigated the appropriate scoring procedures for the format (Beeckmans et al., 2001; Huijbregtse et al., 2002; Mochida & Harrington, 2006). The project to be reported here extends the Yes/No research in three ways. First, it presents the target words in spoken as well as written form. Secondly, it explores how the addition of two types of sentence-based context influences performance on the Yes/No task. The third innovation is to investigate whether reaction time adds a significant dimension to the measurement of vocabulary knowledge with this test format. The sample of target words has been drawn from the word frequency counts of the British National Corpus. Several computer-based forms of the test have been developed to address the various research questions. In this session, the presenters will give a rationale for this approach to testing and discuss the design and development of the various versions of the test. Results from trials with learners of English at the university level in Japan will also be available.

Relationship Among the test, curriculum, and teacher content representations in an EFL setting

Sultan Turkan, *University of Arizona* (sultant@email.arizona.edu)

This study aimed to describe the comparability among the content represented on an English as a Foreign language (EFL) university entrance exam, on an English language high school curriculum, and teacher content representations. The comparability considerations among the test, curriculum, and teacher content representations in classrooms were framed according to the premises of test validity and/or instructional validity (Liu & Fulmer, 2008). Mainly, this study highlights comparability as a fairness issue on the basis of the condition that the test and curriculum standards do not match and the teachers do not adequately cover the content domains of the standards; the validity of inferences about the extent to which students have mastered the standards will be diluted.

WORKS-IN-PROGRESS ABSTRACTS

The comparability between the curriculum and the test was examined according to both language functions organized under particular goal statements and cognitive levels. A panel of three EFL teachers served to examine the content represented on the EFL test and the national high school curriculum. To get at the extent of alignment between the test and curriculum, EFL teachers first worked on the curriculum and mapped the national 11th grade English language functions organized under broad curriculum goal statements onto the given cognitive levels. Once the organization of these language functions was complete, teachers were asked to think of these language functions against 5 cognitive levels (knowledge, comprehension, application, analysis, synthesis). They were then asked to code the cognitive level of each language function or skill that falls under a particular goal statement within a particular content area (or 'modality', say, writing). Once the coding was complete, I then took sub-totals of rows and columns and converted the ratings into proportions. Test items were also coded vis a vis the cognitive levels and language functions organized under particular goal statements. Conversion of the raw numbers into proportions was done both for the curriculum table and test table. Once each conversion was complete, I took the discrepancy (subtraction) between each cell value of the curriculum table and that of the test table. Afterwards, the total absolute discrepancy was attained by summing the absolute values of these differences (Liu & Fulmer, 2008, p. 3). The overall alignment index tells the degree of comparability between test cells and curriculum cells.

To get at how English as a Foreign Language (EFL) teachers represent tested content, I utilized Doyle's (1985) concept of content representation defined as "the ways in which the curriculum is made concrete in the classroom tasks teachers define for students" (p. 3). Through conducting bi-weekly teacher logs and teacher interviews across five high schools in Antalya in the Mediterranean province of Turkey, I inquired what academic tasks are designed in the content representations of five EFL teachers. More specifically in line with Doyle's categorization of academic task (1984), I recorded the products, operations, and resources that form the tasks teachers described in their daily logs.

As far as the preliminary findings are concerned, another independent panel of EFL teachers, who judged the comparability between the content represented on the test and teachers' daily content representations, reported that the test does not register tasks that teachers design in such content domains as speaking, listening, and writing. The panel listed some of the teacher designed tasks that misalign with the content represented on the test as follows: Paraphrasing a text about technology; writing a complaint letter to a consumers' magazine; listening to a text and guessing the meaning of unknown words related to computers; oral description of a funny scene.

Vocabulary knowledge and its use in EFL speaking and writing test performance

Viphavee Vongpumivitch, *National Tsing Hua University*
(viphavee@mx.nthu.edu.tw)

WORKS-IN-PROGRESS ABSTRACTS

The aim of this research project is to investigate vocabulary use in standardized speaking and writing tests by Taiwanese EFL learners of different levels of proficiency. The project asks whether participants who perform better on the speaking and writing tests show a more advanced level of vocabulary knowledge, as defined by their higher vocabulary tests' scores and more uses of academic words in their free production. The project also asks whether the learners' ability to use their vocabulary knowledge is influenced by the different speaking and writing task types. Finally, the project examines the relationship between receptive and productive vocabulary knowledge as measured by the different vocabulary tests used in this study.

The speaking and writing tests used in this study are the practice versions of ETS' TOEFL iBT and Taiwan's LTTC's General English Proficiency Test (GEPT) high-intermediate level. The two tests' speaking sections cover a wide range of tasks, from short-answer questions, picture description, opinion expression, to integrated speaking tasks. Similarly, the writing sections of these two tests include different tasks, namely, argumentative writing and an integrated writing task based on reading and listening materials. It is hypothesized that different task types will require the learners to use their vocabulary knowledge to different degrees.

Vocabulary tests that have already been validated in the literature are used in this study. Receptive vocabulary knowledge is assessed through Vocabulary Levels Test (Schmitt, Schmitt, and Clapham, 2001). Productive vocabulary knowledge is assessed through the controlled productive test (Laufer & Nation, 1999) and Meara and Fitzpatrick's (2000) word association test. Finally, the participants' spoken and written data are analyzed for free productive vocabulary knowledge through Lexical Frequency Profile (Laufer & Nation, 1995).

At the time of LTRC (March, 2009), it is anticipated that approximately sixty EFL Taiwanese learners would have been involved in this study. Participants are high-school, undergraduate, and graduate students who are familiar with both the TOEFL iBT and GEPT tests. Data collection is done individually, with the participants first taking the speaking and writing tests, then the vocabulary tests; data collection ended with individual interviews. Participants' levels of EFL proficiency are categorized into low/mid/high based on their performance on the TOEFL iBT speaking/writing test tasks. In this work-in-progress presentation, preliminary results of statistical analyses, e.g., correlations, factorial ANOVAs, and regressions will be presented. The presenter will ask the audience for feedback regarding interpretation of the interview data which will cover participants' background information, perception of the role of vocabulary on speaking and writing ability, and their vocabulary learning motivation. Plans for further data collection and analysis will also be discussed.

Video discourse completion tasks for the testing of L2 pragmatic competence

Elvis Wagner, *Temple University* (elviswag@temple.edu)

Tina Hu, *Temple University* (tina.hu@temple.edu)

WORKS-IN-PROGRESS ABSTRACTS

The objective of this study is to explore the use of different types of instruments to assess non-native (L2) speakers of English pragmatic competence, especially how the use of videotexts might affect the measurement of that competence. There is a large amount of L2 pragmatics research that has examined the use of discourse completion tasks (DCTs) in assessing L2 pragmatic competence, but the research investigating how varying the characteristics of the tasks might affect the measurement of pragmatic competence is inconclusive. Consequently the current study will use a quasi-experimental design to investigate how: (a) the inclusion of the visual channel (through the use of videotexts) might affect test-takers' responses on the DCTs, and (b) how native speakers' responses to video DCTs compare with non-native speakers' responses.

Both advanced ESL speakers and native speakers will take a pragmatics competence test that consists of 9 DCTs that will serve as a pre-test measure of L2 pragmatic competence on a variety of speech acts (requests, refusals, and apologies). They will take this test on a computer, where the written contextual information and the actual prompt will be presented on the computer monitor, and the test-taker will then respond orally into a microphone connected to the computer.

Participants will then be assigned to either the experimental group or the control group, and they will then take another pragmatics competence test consisting of 12 DCT items designed to measure a learner's knowledge of refusals in English. The participants in the control group will receive a written-input DCT in which the written contextual information and written prompt appear on the computer. Those in the experimental group will receive a written and video-input DCT, in which the written contextual information and a video providing contextual information (the speaker and background setting) are shown, and then the written prompt will appear on the monitor. The participants in both groups will speak their response into a microphone connected to the computer. After completing the test, the participants will complete a questionnaire about their test-taking experience.

The test data will then be analyzed in a number of ways. A comparison of the average length of the responses for the two groups will be conducted by comparing the average number of words for each video DCT versus each written DCT. A second analysis will involve having five native speakers rating the non-native speaker DCT responses using an analytic rating scale. A third analysis will involve analyzing the nature of the responses themselves, comparing the types and quantity of strategies that the two different elicitation types elicited in the responses of the participants. In addition, the questionnaire will be examined to investigate the test-takers' perceptions of the use of the video versus written-only discourse completion tasks.

Using integrative task-based assessment to examine the effectiveness of task-based language teaching

Jing Wei, *UMCP* (jwei1@umd.edu)

Cheng-Chiang (Julian) Chen, *UMCP* (jucchen@umd.edu)

WORKS-IN-PROGRESS ABSTRACTS

This quasi-experimental study intends to investigate whether ESL students taught in a task-based language teaching (TBLT) class would outperform those in a grammar-driven class in linguistic competence and the ability to accomplish real-life task. According to TBLT advocates (Long & Crookes, 1992), TBLT is more effective than grammar-driven teaching because the former is an analytical approach that argues that language development can be better achieved through social interactions embedded in authentic and communicative tasks whereas the latter is a synthetic approach that requires that learners to integrate all the isolated linguistic elements in order to meet the communicative needs of real-world tasks (Pica, Kanagy and Falodun, 1993). The significance of our study is to bridge the gap that existing research comparing the effectiveness of TBLT and grammar-driven teaching is still in paucity (Chaudry, 1991; Deen, 1991). Even among the studies that compare the two approaches, few of them use both the task completion and linguistic competence as integrative criteria to assess the learning outcome (Robison, 1996b). Two research questions are raised: 1) Are students in a TBLT class more able to meet the ability requirements of target tasks than those in a traditional grammar-driven class? 2) If they can perform better in task completion than their peers in the traditional class, do they also achieve better linguistic competence than the latter?

The participants are 40 Hispanic students enrolled in a low intermediate ESL class at a U.S. community college. Cluster random assignment is conducted to divide participants into two groups, i.e., TBLT vs. traditional teaching method. "Job interview" is selected as the lesson topic for both groups based on the needs analysis in our pilot study. The TBLT lesson plan incorporates prototypical discourse samples collected from genuine job interviews, whereas the control group uses a linguistically based textbook focusing on job-related topics. Drawing on Robinson's (1996b) assessment framework, we will use task-based assessment (TBA) in our pre- and post-test to measure learners' second language ability and performance. Both system-referenced and performance-referenced types will be integrated in our design of pre-and-post-tests to achieve high face validity and generalizability. Each type will be evaluated based on its own criteria determined by two teams of three assessors: researcher, teacher and domain expert (i.e., a genuine employer from the job that the participants intend to apply for). The criteria used for performance-referenced approach include the minimal requirements for the success in completing target tasks. System-referenced approach will assess learners' linguistic competence measured by language complexity, accuracy and fluency (Yuan and R. Ellis, 2003). Both pre-and-post-tests will use job interview as the prototypical task. In both experimental and control groups, a simulated job interview will be conducted by two teams of three assessors to attain inter-rater reliability among multiple experts.

In resonance with previous TBLT studies, we also anticipate that 1) TBLT instruction will yield better performance in target task completion than grammar-driven instruction; 2) participants' linguistic competence in TBLT group will make comparable progress to the grammar-driven group in language complexity, accuracy and fluency. If the hypothesis is confirmed, our study will support the TBA argument that integrating the criteria in both task completion and linguistic competence to assess the

WORKS-IN-PROGRESS ABSTRACTS

learning outcome will generate results that are of value to both TBLT and grammar-driven advocates.

Applying protocol analysis in analyzing language test validity: A case study

Huijie Xu, *Zhejiang University of Technology* (zgdXu@zjhu.edu.cn)

Ying Zheng, *Queen's University* (2yz13@queensu.ca)

Heying Lou, *Zhejiang University of Technology* (lou_heyingsina@sina.com)

In understanding language test constructs and test performance, the primary focus is usually on the relationship between the two. The cognitive process that is situated in between these two ends of the continuum, however, is an aspect that is normally overlooked in language testing. In response to this situation, a desire to understand the cognitive processes that mediate test constructs and test performances is the rationale of employing protocol analysis in analyzing test validity. Protocol analysis can also provide a useful vehicle in comparing what the test constructs are designed to measure and what the students who write the test perceive of the test constructs.

This study examined 10 Chinese College English test-takers with varying degrees of proficiency levels. The purpose of this study is to investigate students' information processing and their affective and cognitive strategies in writing the test. College English Test (CET) is a large-scale high-stakes English proficiency test taken by Chinese university students.

Method

Protocol analysis employs test-takers' verbal reports as a direct source of information to access their covert mental processing. This approach can examine what tests actually test and whether they are valid reflections of what is assumed to be measured. In addition, using this approach, information with regards to students' utilization of test-taking strategies can also be evaluated apart from the more traditional way of using strategy questionnaires to collect this sort of information (e.g., Cohen, 1984; Nevo, 1989; Wijgh, 1996).

Preliminary Findings

The research questions involved the investigation of test-takers' employment of both linguistic and non-linguistic knowledge in performing on the test items as well as the investigation of the effects of different test formats on their test performance. The test-takers verbal reports revealed that less proficient test-takers depended more on the compensatory function of non-linguistic processing, while more proficient test-takers employed more facilitating functions of non-linguistic processing. The findings indicated that the MC format allowed uninformed guessing, which may cause the test-takers' selection of the right answer for the wrong reasons or for no reason. Further, an investigation into the test items showed that vocabulary may partially turned the reading part of the question into a test of vocabulary, and therefore threatened the construct validity.

WORKS-IN-PROGRESS ABSTRACTS

The preliminary findings of this study also showed that both high proficiency test-takers and low proficiency test-takers were aware of certain reading and test-taking strategies, and both claimed to employ some of them when taking a MC reading comprehension test in the CET. For example, identifying key words in question stems, and using context to guess the meaning of difficult sentences, which requires comprehensive use of linguistic knowledge. What leads to their differential test performance is their respective utilization of these strategies. High proficiency test-takers can correctly identify key words in question stems, while low proficiency test-taker may fail to find the key words. In addition, high proficiency test-takers use more intralingual cues, i.e., cues based on their knowledge of English, lexical and semantic cues, while low proficiency test-takers overuse interlingual cues, i.e., cues based on the first language other than English.

The study also showed that inadequate linguistic knowledge as well as test-taking strategies may lead low proficiency test-takers to make wrong choices in the test. When low proficiency test-takers use context to guess the meaning, they may focus on the one or two sentences close to the key words they presumed, based on their understanding of question stems. They could hardly connect the information conveyed in the preceding paragraph or the information three or four sentences away from the key words. Therefore, it would be very meaningful to equip students with more linguistic knowledge of English to improve their reading proficiency.

Educational Importance of the Study

Findings from the use of protocol analysis in language test validity are applicable to the research on gaining a better understanding of different test-takers' cognitive processes when they are taking a test. Knowledge about students' information processing in writing a test, especially students speaking different languages, coming from different backgrounds or at different proficiency levels, may have important implications with regards to test fairness and test bias studies. Clearly, using protocol analysis is a useful means to get answers to the "why" questions. i.e., why students differ in their test performance, why students answer a particular question in a particular way.

We aim to recruit more test-takers in the next phase of the study, mainly focusing on comparing test-takers with more varying characteristics (e.g. gender and major in university etc).

The Relationship of TOEFL scores to success in American universities: How high is high enough?

Brent Bridgeman, *Educational Testing Services* (bbridgeman@ets.org)

Yeonsuk Cho, *Educational Testing Services* (ycho@ets.org)

TOEFL iBT® was introduced in 2005 as an improved measure of academic language skills of non-native English speakers. The purpose of the study is to evaluate how well TOEFL iBT® can predict language skills necessary for academic tasks. About 2,880

WORKS-IN-PROGRESS ABSTRACTS

graduate and undergraduate international students are recruited from 30 universities in the U.S. with a high enrollment of international students. Participants are in the first or second year of their studies majoring in one of the following three disciplines where a majority of international students are concentrated: business, engineering and social sciences. They are asked to provide scores from TOEFL or IELTS, first year grades, and scores from GRE, GMAT, SAT or ACT if available, and complete an on-line questionnaire about their own academic language skills. For graduate students only, in addition to student self assessment, faculty ratings of the student's language skills are collected using an on-line questionnaire. Relationships between test scores and criterion variables such as course grades and ratings of academic language skills will be presented graphically in expectancy tables. More conventional regression-based estimates may also be possible after making adjustments for the probable non-linearity in the raw scores. Results of the study are expected to provide useful information to test users.

INDEX OF PRESENTERS

A

Ahn, Seongmee · 68
Aitken, Avril · 69
Assias, Nathalie · 37

B

Baffi-Bonvino, Melissa · 87
Baker, Beverly · 68
Banerjee, Jayanti · 32
Bauman, Jim · 40
Beck, Sara W. · 61
Bolli, Giuliana Grego · 36
Bridgeman, Brent · 101
Brooks, Rachel · 85
Brown, Annie · 70
Buck, Gary · 32
Bunch, Michael B. · 33

C

Cai, Hongwen · 46
Chen, Cheng-Chiang (Julian) · 98
Chi, Youngshin · 86
Cho, Yeonsuk · 101
Clark, Martyn · 70
Colby-Kelly, Christian · 86
Consolo, Douglas Altamiro · 87
Cumming, Alister · 38

D

Davis, Larry · 71
DeJong, Nivya · 59, 62
Downey, Ryan · 58

F

Florijn, Arjen · 62
Fujita, Tomoko · 72

G

Galaczi, Evelina · 45
Gebriel, Atta · 53
Grabe, William · 55
Gu, Xiangdong · 83

Guodong Jia, Renmin · 91

H

Halleck, Gene · 72
Han, Sumi · 88
Hetherington, Anne · 69
Hill, Yao · 54
Hsieh, Ching-Ni · 90
Hu, Tina · 97
Huang, Heng-Tsung Danny, · 73
Huang, Ting · 76
Hulstijn, Jan · 62

I

In'nami, Yo · 65
Isaacs, Talia · 57

J

Jiang, Xiangying · 55
Jones, Mariam · 76

K

Kadessa Abdul Kadir · 44
Kadir, Kadessa Abdul · 44
Kang, Okim · 64
Kanza, Tzahi · 37
Kenyon, Dorry · 60
Knoch, Ute · 47
Koizumi, Rie · 65
Kunschak, Claudia · 92

L

Lazaraton, Anne · 71
Lee, Jiyeon · 52
Lee, Yong-Won · 38
Leung, Constant · 50
Lin, Chih-Kai (Cary) · 92
Liu, Ou Lydia · 54
Llosa, Lorena · 61
Loomis, Summer · 74
Lou, Heying · 99
Lu, Lu · 75
Luo, Kaizhou · 76

INDEX OF PRESENTERS

M

MacGregor, David · 40, 60, 77
Malone, Margaret E. · 93
McNamara, Tim · 34
Montee, Meg · 41
Montee, Megan · 93

N

Nordby Chen, Natalie · 32

O

Ockey, Gary · 51

P

Pan, Yi-Ching · 43
Papp, Szilvia · 36
Park, Ok-Sook · 68
Pickering, Lucy · 64
Plakans, Lia · 53
Poehner, Matt · 50

Q

Quinlan, Thomas · 38

R

Read, John · 94
Rea-Dickins, Pauline · 50
Reed, Daniel · 68
Rocca, Lorenzo · 36
Rubin, Don · 64
Rumhild, Anja · 60

S

Saville, Nick · 34, 36
Sawaki, Yasuyo · 37, 38
Schoonen, Rob · 62
Shepard, Lorrie A. · 31
Shiotsu, Toshihiko · 78, 94
Shohamy, Elana · 34, 37, 50
Spokane, Abbe · 40
Spotti, Max · 35

Steinel, Margarita · 62

T

Taylor, Lynda · 50, 66
Templin, Jonathan · 32
Thomson, Ron · 57
Turkan, Sultan · 95

U

Urmston, Alan · 78

V

Van Avermaet, Piet · 35
Van Moere, Alistair · 58
van Naerssen, Margaret · 79
Vongpumivitch, Viphavee · 96

W

Wagner, Elvis · 97
Walters, JoDee · 80
Wei, Jing · 98
Weigle, Sara Cushing · 39
Wolfersberger, Mark · 39
Wright, Laura · 40
Wu, Weiping · 90

X

Xu, Huijie · 99

Y

Yang, Hui-chun · 81
Yang, WeiWei · 39
Yoshizawa, Kiyomi · 82
Yu, Guoxing · 37, 40

Z

Zhang, Jie · 48
Zhao, Cecilia G. · 61
Zheng, Ying · 99
Zheng, Yujing · 83