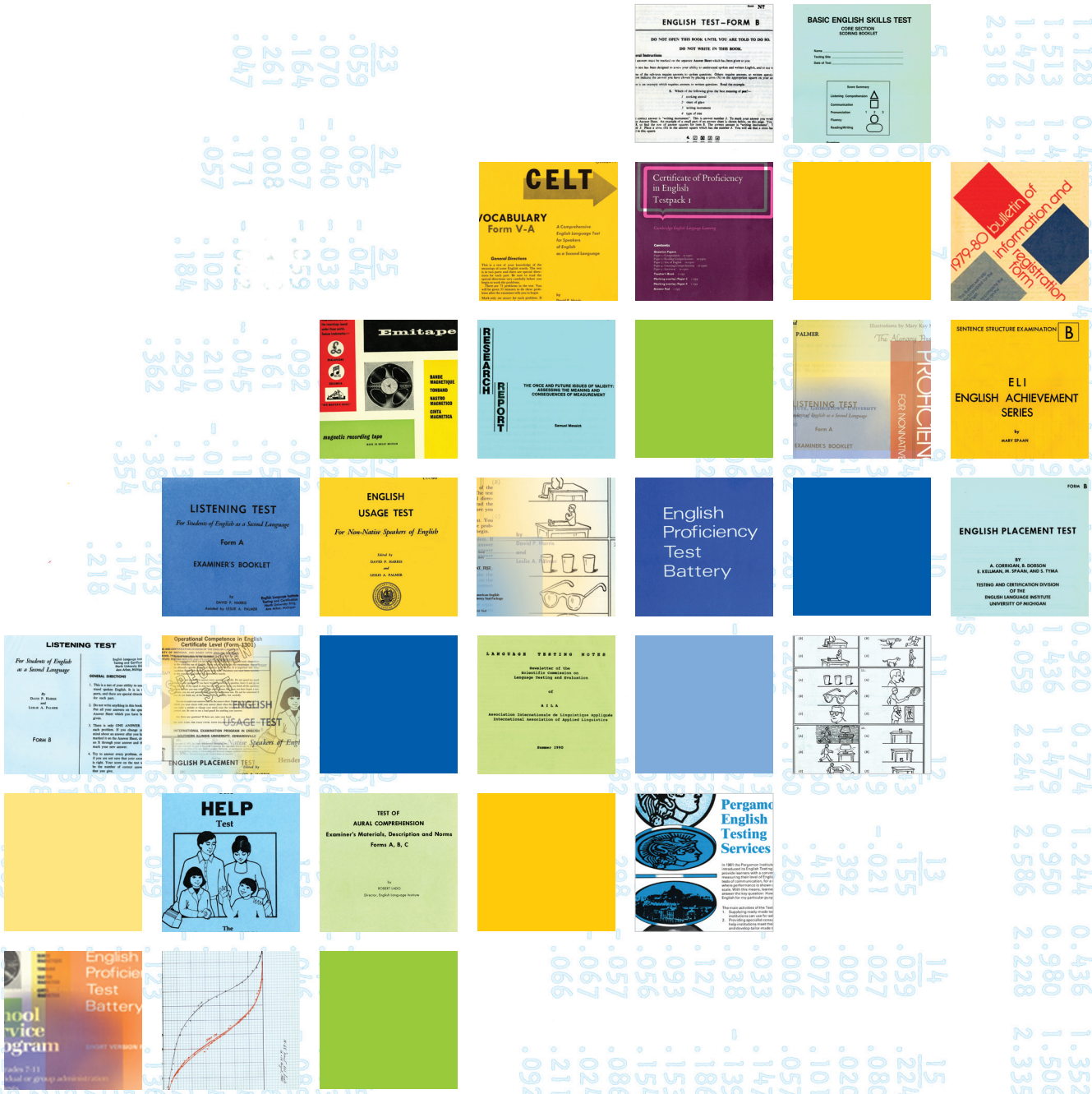
LTRC 2011

33rd Language Testing Research Colloquium

University of Michigan

June 23–25

[preconference workshops June 21 & 22]



Half a Century of Language Testing

It's our mission.

For more than 60 years, nonprofit Educational Testing Service (ETS) has built a reputation as the global leader in English-language learning. Administering more than 50 million tests — including the TOEFL® and TOEIC® tests — in more than 180 countries each year, ETS helps advance quality and equity in education worldwide.

To learn more about ETS's commitment to global education, visit **www.ets.org**.

ETS — *creator of the TOEFL® and TOEIC® tests*

ETS®

Listening. Learning. Leading.®

Table of Contents

Cambridge Michigan Language Assessments

Cambridge–Michigan Language Assessments
is delighted to be hosting LTRC 2011

As a collaboration between University of Cambridge ESOL Examinations and the University of Michigan, we share a commitment to using research to benefit language teaching and learning worldwide.

We welcome you all to Michigan for a unique opportunity to share knowledge and ideas with other experts in assessment from around the globe.

www.CambridgeMichigan.org

UNIVERSITY *of* CAMBRIDGE
ESOL Examinations

UNIVERSITY OF MICHIGAN

Welcome

Message from the ILTA President

It is my great pleasure to welcome you all to this 33rd Annual Language Testing Research Colloquium on behalf of the parent organization, the International Language Testing Association (ILTA).

For those of us in language testing and assessment, coming together at LTRC is a time to be treasured for a whole variety of reasons. It is the major international conference on language testing research, and we are proud of the consistently high standard of papers that have been presented at LTRC over the years. The conference also provides a venue for ILTA to hold its annual business meeting and a face-to-face meeting of the executive board, and for the major journals in our field to convene meetings of their editorial boards. In addition, we take the opportunity to celebrate the high achievers among us through the range of awards that are presented at the banquet, which is always a highlight of the conference program. We must not forget the social side of LTRC as well. In my experience, there is a distinctive quality of friendship and bonding among participants that has been a feature of the colloquium since its earliest years. It is one reason that many of us not only keep coming back to the conference year after year but also find ourselves forming warm friendships and professional collaborations with people we have met at LTRC.

I want to give a special welcome to those of you who are attending LTRC for the first time, and especially graduate students. Like many other spheres of contemporary society, our field has been dominated for years by us baby boomers and, as we turn grey and contemplate retirement, we are conscious of the need to nurture a new generation of language testers. We hope that you will experience the true spirit of LTRC here in Ann Arbor and it will encourage you to make a long-term commitment to the field of language testing and assessment.

After being a purely North American event for its first fifteen years, LTRC has been held elsewhere in the world on average every second year since 1993, in keeping with its international character. This year the colloquium is in a sense coming home, both to the U.S. and more specifically to the University of Michigan, which has a long history as a major center of language testing research and development. The theme of LTRC this year reflects our conventional notion that language testing as a discipline had its beginning in 1961, with the publication of Robert Lado's book *Language Testing* and John B. Carroll's presentation of an influential paper at the Washington conference that led to the development of TOEFL. Since Lado's book drew on his work here in Ann Arbor in the 1950s, it is highly appropriate for us to recognize the Michigan testing program as having a formative influence on his thinking.

I am reasonably confident that none of the participants in the Washington conference in 1961 is still active in the field. However, a number of us veterans had the privilege of participating in LT+25, a symposium organized by our ILTA affiliate in Israel, ACROLT, in 1986 to mark the 25th anniversary of the seminal contributions by Lado and Carroll, who were both there for the occasion. Regrettably, they are no longer with us, but we certainly honour their memory 50 years on.

Speaking of anniversaries, it is 20 years since a meeting was held during LTRC in Princeton to form an interim board which would establish the International Language Testing Association. We will be returning to Princeton for LTRC next year, but in the meantime I am happy to report that ILTA has grown substantially since those early days and can be said to have achieved a new level of maturity in recent years. There is always a limit on how much a purely voluntary organization can achieve, but we have sought to put our operations on a more professional basis by extending the terms of the vice president, president, and immediate past president from one year to two; by employing a management company; and by documenting our standard procedures more systematically than in the past.

More importantly, we are seeking to expand our influence beyond our direct membership by promoting best practice in language testing and assessment through a variety of means. We have seen a steady increase in our affiliates around the world, with new associations in Canada and Australia / New Zealand expected to seek affiliation soon. Each year we award two grants for workshops and meetings in parts of the world where expertise and good practice in testing are lacking. We have a project underway to translate the ILTA Code of Ethics and the Guidelines for Practice into a number of other languages. And another project we are actively engaged in is the scheme being developed by the International Civil Aviation Organization to endorse the quality of aviation English tests.

A significant new initiative since the last LTRC is the establishment of a Task Force on Quality Assurance for Language Testing, Evaluation, and Assessment, following the vigorous debate on the LTEST-L email list last year about how to address unethical and unprofessional practices in testing and to encourage a better understanding of what good quality language assessment involves. With Dan Douglas as chair, the task force has made great progress this year and they will hold two open sessions during the conference, on Thursday lunchtime and Saturday afternoon. This is your opportunity to hear what they are proposing and to give your input.

In addition to the task force sessions, we will have the ILTA Annual Business Meeting during the lunch hour on Saturday, and I urge all members to attend so that we on the executive board can inform you of our current activities and hear your views on the future direction of the association.

Finally, I want to express my great appreciation to India Plough, the organizing committee, and everyone else involved in preparing for this LTRC. It is always a major commitment to organize a meeting of this size and complexity, with all kinds of challenges and hitches along the way, right until the last minute. We will have the chance to thank them more fully at the closing banquet, but at this point I want to acknowledge the incredible amount of work they have put in and their absolute dedication to delivering a wonderful conference in the LTRC tradition. We can confidently look forward to a most successful meeting.

John Read

President, International Language Testing Association
2011–12

Samuel J. Messick Memorial Lecture

John Michael Linacre, PhD
University of the Sunshine Coast, Australia

Constructing Valid Performance Assessments — The View from the Shoulders of the Giants

After a brief survey of the activities of some of the giants of validity theory, the lecture focuses on five practical considerations aimed at obtaining valid measures of judge-intermediated rated performances.

1. The criterion for a "good judge" is defined precisely and coherently in advance of judging.
2. The judging plan is sufficiently crossed for all the measures to be estimated in one frame-of-reference, or so that data analysis can be constrained in a reasonable way to obtain the same result.
3. The data analysis begins as soon as ratings start to be collected, in order that flaws in the judging plan, failures in its implementation, and unintended behavior by the judges can be detected and remedied while judging is underway.
4. Measures are reported in an additive frame-of-reference.
5. When a problem arises, ask yourself "How would a physicist solve this problem?"

These five considerations are reinforced with anecdotal examples featuring intellectual giants past and present.

Background

Dr. Linacre is the research director of winsteps.com. He was formerly director of the MESA Psychometric Laboratory at the University of Chicago and has held many positions in academia. He has taught the principles and practice of Rasch measurement since the 1980s and authored over 200 journal articles and conference papers.

The significant influence that the work of Dr. Linacre has had on psychometrics and within the field of language testing in particular has been far-reaching. In 1989 Dr. Linacre extended Rasch models for dichotomous data, Likert scales, and partial credit data to accommodate measurement needs of assessment situations not addressed by earlier Rasch models. To operationalize the many-facet Rasch model, Dr. Linacre also created the computer program FACETS, which is regularly used by language testers to investigate aspects of performance assessments that affect examinee scores, such as rater and task characteristics.

EIKEN

Test in Practical English Proficiency

You spoke.

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

We responded.

STEP is pleased to announce the launch of our **new English website**, featuring:

- Overviews of the seven EIKEN grades
- Free downloads of test booklets and audio files
- Summaries of recent research
- Interactive registration system for schools recognizing EIKEN

英検

stepeiken.org

Stop by. Discover EIKEN.

Sponsors

Ballard & Tighe
www.ballard-tighe.com

Lidget Green
www. lidgetgreen.org

Cambridge Michigan Language Assessments
www.CambridgeMichigan.org

U-M College of Literature, Science, and the Arts
www.lsa.umich.edu

U-M Center for Japanese Studies
www.ii.umich.edu/cjs

Michigan State University
elc.msu.edu

Educational Testing Service (ETS)
www.ets.org/toefl

Second Language Testing, Inc. (SLTI)
www.2lti.com

EIKEN
www.stepeiken.org

University of Cambridge
ESOL Examinations
www.cambridgeesol.org

U-M International Institute
www.ii.umich.edu

Language Learning Journal
www.blackwellpublishing.com/journal.asp

Conference Exhibitors

- University of Cambridge ESOL Examinations
- Educational Testing Service (ETS)
- University of Michigan Press
- Learning Pyramid
- Taylor & Francis

Conference Support

Organizers

Pamela Bogart
University of Michigan

Sarah Briggs
Cambridge Michigan
Language Assessments

Mark Chapman
Cambridge Michigan
Language Assessments

Katie Coleman
Cambridge Michigan
Language Assessments

Barbara Dame
Cambridge Michigan
Language Assessments

Renee Dean
Cambridge Michigan
Language Assessments

Barbara Dobson
Cambridge Michigan
Language Assessments

Christine Feak
University of Michigan

Caitlin Gdowski
Cambridge Michigan
Language Assessments

Ching-Ni Hsieh
Cambridge Michigan
Language Assessments

Eric Lagergren
Cambridge Michigan
Language Assessments

Cindy Lin
University of Michigan

Fabiana MacMillan
Cambridge Michigan
Language Assessments

Meg Malone
Center for Applied Linguistics

Amanda McConville
Cambridge Michigan
Language Assessments

Aaron Ohlrogge
Michigan State University

Spiros Papageorgiou
Cambridge Michigan
Language Assessments

India C. Plough
Cambridge Michigan
Language Assessments

Robin Stevens
Lidget Green

Ildi Porter-Szucs
Cambridge Michigan
Language Assessments

Paula Winke
Michigan State University

Barbara Wood
Cambridge Michigan
Language Assessments

Xiaoming Xi
ILTA, ETS

Volunteers

Allison Piippo
Eastern Michigan University

Dan Hopper
Eastern Michigan University

Edward Getman
Teachers College,
Columbia University

Erik Voss
Iowa State University

Heejeong Jeong
University of Illinois
at Urbana-Champaign

Hyejin Yang
Iowa State University

Jing Xu
Iowa State University

Jinsong Fan
Shanghai Jiao Tong University

Megan Montee
Georgia State Unviersity

Michael Collins
Eastern Michigan University

Michelle Raquel
University of Tasmania

Rika Tsushima
McGill University

Sara Okello
Eastern Michigan University

Sarah Goodwin
Georgia State University

Sawako Matsugu
Northern Arizona University

Shangchao Min
Zhejiang University

Soo Hyon Kim
Michigan State University

Stéphanie Gaillard
University of Illinois
at Urbana-Champaign

Tsang Hoi Ka
Institute of Education
University of London;
Hong Kong Polytechnic University

Xiaomei Song
Queen's University

Xu Liu
Guangdong University
of Foreign Studies

Yeon-Sook Yi
University of Illinois
at Urbana-Champaign

Yeting Liu
University of Pennsylvania

Yoo-Ree Chung
Iowa State University

Yu-Chen Tina Lin
Indiana University, Bloomington

Program Design

Kate Boyd
Cambridge Michigan
Language Assessments

Eric Lagergren
Cambridge Michigan
Language Assessments

Barbara Wood
Cambridge Michigan
Language Assessments

Signage & Info Booklet

Sarah Erlewine
Cambridge Michigan
Language Assessments

ILTA 2011

ILTA Executive Board 2011

President
John Read
University of Auckland

Vice President
Gillian Wigglesworth
University of Melbourne

Secretary
Margaret (Meg) Malone
Center for Applied Linguistics

Treasurer
Sara Cushing Weigle
Georgia State University

Immediate Past President
Carolyn Turner
McGill University

Members at Large

Nathan Carr
California State University, Fullerton

Lia Plakans
University of Iowa

Yasuyo Sawaki
Waseda University

Lorena Llosa
New York University

ILTA Nominating Committee 2011

Chair
Jayanti Banerjee
Cambridge Michigan
Language Assessments

Luke Harding
Lancaster University

Sari Luoma
Ballard & Tighe

India C. Plough
Cambridge Michigan
Language Assessments

Task Force on Quality Assurance in Language Testing, Evaluation, and Assessment

Chair
Dan Douglas
Iowa State University

Jamie Dunlea
STEP EIKEN

Liz Hamp-Lyons
University of Bedfordshire

Yasuyo Sawaki
Waseda University

Diane Schmitt
Nottingham Trent University

Bernard Spolsky
Bar-Ilan University

Lynda Taylor
University of Cambridge
ESOL Examinations

Awards

Robert Lado Memorial Award for Best Graduate Student Paper

2011 Selection Committee

Chair
Spiros Papageorgiou
Cambridge Michigan
Language Assessments

Ofra Inbar-Lourie
Tel Aviv University & Beit Berl College

Carol Moder
Oklahoma State University

Jessica Wu
Language Training and Testing
Center, Taiwan

Student Travel Awards

Chair
Gillian Wigglesworth
University of Melbourne

Nathan Carr
California State University, Fullerton

Yasuyo Sawaki
Waseda University

Christine Doe
Queen's University
Validation of a large-scale assessment for diagnostic purposes across three contexts: Scoring, teaching, and learning

Yujie Jia
University of California, Los Angeles
Justifying score-based interpretations from a second language oral test: Multi-group confirmatory factor analysis

Best Article Award

Chair
Xiaoming Xi
Educational Testing Service

Ari Huhta
University of Jyväskylä

Lia Plakans
University of Iowa

Rob Schoonen
University of Amsterdam

Elvis Wagner
Temple University

Glenn Fulcher and **Fred Davidson** (2009). Test Architecture, Test Retrofit. *Language Testing*, (26) *1*, 123–144.

Workshops and Meetings Award

Chair
John Read
University of Auckland

Lorena Llosa
New York University

Lia Plakans
University of Iowa

Kateřina Vlasáková
Kateřina Hlinová
Relating examinations to language frameworks especially in the Czech Republic

Ana Maria Ducasse
Javier Menéndez
Juan Robisco
Testing Spanish as a foreign language in secondary schools in the Philippines

Jacqueline Ross TOEFL Dissertation Award

Kirby C. Grabowski
Teachers College, Columbia University

Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking

Dissertation advisor
Professor James Purpura

2010 Caroline Clapham IELTS Masters Award

Thom Kiddle
Lancaster University

The effect of mode of response on a semi-direct test of oral proficiency

Supervisor
Dr. Judit Kormos

Abstract Review

Lyle Bachman
University of California, Los Angeles

Beverly Baker
McGill University

Jayanti Banerjee
Cambridge Michigan
Language Assessments

Vivien Berry
Hong Kong University

Sarah Briggs
Cambridge Michigan
Language Assessments

Annie Brown
Ministry of Higher Education
and Scientific Research

Gary Buck
Lidget Green

Nathan Carr
California State University, Fullerton

Micheline Chalhoub-Deville
University of North Carolina at Greensboro

Martyn Clark
University of Oregon

Christine Coombe
Higher Colleges of Technology

Deborah Crusan
Wright State University

Alister Cumming
University of Toronto

Fred Davidson
University of Illinois

John De Jong
Pearson

Barbara Dobson
Cambridge Michigan
Language Assessments

Dan Douglas
Iowa State University

Janna Fox
Carleton University

Evelina Galaczi
University of Cambridge
ESOL Examinations

Ardeshir Geranpayeh
University of Cambridge
ESOL Examinations

April Ginther
Purdue University

Kirby Grabowski
Columbia University

Anthony Green
University of Bedfordshire

Liz Hamp-Lyons
Hong Kong Polytechnic University

Luke Harding
Lancaster University

Claudia Harsch
University of Warwick

Talia Isaacs
University of Bristol

Dorry Kenyon
Center for Applied Linguistics

Ute Knoch
University of Melbourne

Antony Kunnan
California State University, Los Angeles

Yong-Won Lee
Seoul National University

Gad S. Lim
University of Cambridge
ESOL Examinations

Lorena Llosa
New York University

Sari Luoma
Ballard & Tighe

Meg Malone
Center for Applied Linguistics

Fumiyo Nakatsuhara
University of Bedfordshire

Spiros Papageorgiou
Cambridge Michigan
Language Assessments

Lia Plakans
University of Iowa

India C. Plough
Cambridge Michigan
Language Assessments

James Purpura
Columbia University

David Qian
Hong Kong Polytechnic University

John Read
University of Auckland

Dan Reed
Michigan State University

Nick Saville
University of Cambridge
ESOL Examinations

Rob Schoonen
University of Amsterdam

Toshihiko Shiotsu
Kurume University

Elana Shohamy
Tel-Aviv University

Charles Stansfield
Second Language Testing, Inc.

May Tan
McGill University

Lynda Taylor
University of Cambridge
ESOL Examinations

Randy Thrasher
Okinawa Christian University

Carolyn Turner
McGill University

Alistair Van Moere
Ordinate Corporation

Elvis Wagner
Temple University

Gillian Wigglesworth
University of Melbourne

Paula Winke
Michigan State University

Guoxing Yu
Bristol University

Workshops

Everything You Need to Know to Teach Statistics in Language Testing Classes

Pre-Conference Workshop • June 21–22, 2011 • 1500 North Quad (Lab 1)

Dr. J. D. Brown

You will walk away from this workshop with hundreds of ideas about what can be taught in a language testing course. You will also develop the skills necessary to decide what to teach in your language testing course based on thinking about who you are teaching (with a focus on issues of assessment literacy, testing class audience, innumeracy, etc.) and why you are teaching language testing to these particular folks. You will improve your statistics-teaching skills by learning: what topics to avoid, why you should use conceptual rather than short-cut formulas, ways to keep statistics simple, the importance of using humor, the value of hands-on practice, etc. Along the way, you will see and critique a demonstration lesson, and the next day, teach a ten-minute lesson on a narrowly defined topic to a small group who will give you feedback/suggestions on your mini-lesson.

I am not assuming any particular knowledge of statistics. In fact, I anticipate that participants will range widely in such knowledge. However, rudimentary background in descriptive statistics will help, as will minimal skills in using some version of the Exceltm spreadsheet program. I encourage you to bring your own laptop computer (though computers will be available on site) because I believe you will benefit from working on your own version of Excel on your own computer. If you already have it, please bring Brown, J. D. (2005). *Testing in language programs*. New York: McGraw-Hill. If not, I will make copies available on site. Also, feel free to bring along materials that you currently use for teaching language testing. The workshop will include talking-head sessions, group work, hands-on practice, and a good many laughs!

Tuesday, June 21

8:30–9:00	Workshop and conference registration (North Quad)
9:00–10:15	Introduction to workshop and participants. Why do we teach language testing? And, to whom? The questions of assessment literacy, testing class audience, innumeracy, etc.
10:30–11:45	What should we teach in our language classes?
12:45–2:00	What should we teach in our language classes? (cont.)
2:25–4:00	How should we teach language testing? Topics to avoid, keeping it simple, demonstration lesson/critique, and set up mini-lessons for day 2

Wednesday, June 22

9:00–10:15	How should we teach language testing? (cont.) Use conceptual rather than shortcut formulas, the importance of humor and hands-on practice in learning statistics
10:30–11:45	Participant statistics mini-lesson presentations/critiques
12:45–2:00	Participant statistics mini-lesson presentations/critiques (cont.)
2:25–4:00	Wrapping it up. What you concluded from this workshop, and evaluation

Each full day's session includes two breaks plus time for lunch.
All snacks, drinks, and lunches will be provided.

Participant Evaluation

Participants will be asked to give written comments on the content of the workshop, the presentation of the materials, and the usefulness of the hands-on activities. Additionally, they will be asked for possible suggestions for future workshops.

Workshops

Applying SEM for Language Testing Research

Pre-Conference Workshop • June 21–22, 2011 • 1500 North Quad (Lab 2)

Ardeshir Geranpayeh, PhD

This is an advanced course in language testing and is aimed at experienced and knowledgeable language testing professionals. Participants are expected to have some understanding of basic regression and analysis of variance. The workshop will use EQS (Multivariate Software), which is one of the most widely used Structural Equation Modeling programs, know for its ease of use and powerful features. The workshop will be offered in one of the University of Michigan's state-of-the art computer labs. All participants will receive a copy of the full version of the software to try out on their own PC for two weeks.

This workshop will introduce you to the conceptual principles underlying SEM and will allow you to quantify and test substantive theories. It provides you with the basic skills to build and evaluate your priori theory in various language testing/learning domains. The focus is on using quantitative methodology to investigate substantive theories; for example, in language acquisition, language

assessment, or the factors that may affect performance on language tests. At the end of the workshop, you should be able to build both measurement and structural models and evaluate them to test your theories. The workshop will have hands-on exercises on each day to allow you to work independently at your own pace.

Tuesday, June 21

8:30–9:00	Workshop and conference registration (North Quad)
9:00–10:30	Introduction to workshop and participants Conceptual Matters 1: SEM, Components of SEM, Different Models
10:50–12:15	Building Measurement Models
1:15–3:00	Conceptual Matters 2: Model Specification, Identification, Parameter Estimates, Fit Indices
3:15–5:00	Practical SEM investigation using EQS

Wednesday, June 22

9:00–10:30	Understanding SEM Models
10:45–12:15	Evaluating SEM Models
1:15–3:00	Practical examples using EQS
3:20–5:00	Advanced Q & A including a note on multi-group analysis

Each full day's session includes two breaks plus time for lunch.
All snacks, drinks, and lunches will be provided.

Participant Evaluation

Participants will be asked to give written comments on the content of the workshop, the presentation of the materials, and the usefulness of the hands-on activities. Additionally, they will be asked for possible suggestions for future workshops.

Conference Overview

Tuesday, June 21

8:30–9:00 **Workshop Registration**
North Quad *Room 1500*

9:00–5:00 **Workshops**
North Quad *Room 1500*

3:00–7:00 **Conference Registration**
Rackham Lobby

Wednesday, June 22

8:15–5:15 **Conference Registration**

9:00–5:00 **Workshops**
North Quad *Room 1500*

1:00–5:00 **ILTA Executive Board Meeting**
Rackham East Conference Room *4th Floor*

3:00–6:00 **ILTA Task Force Meeting**
Rackham West Conference Room *4th Floor*

6:30–8:00 **Opening Reception**
University of Michigan Museum of Art

Thursday, June 23

8:15–5:15 **Conference Registration**
Rackham Lobby

9:00–6:00 **Exhibitors**
Rackham Assembly Hall *4th Floor*

8:30–8:45 **Welcome**

8:50–10:00 **Messick Memorial Speaker Award and Lecture**
Rackham Auditorium *1st Floor*

10:00–10:15 **Break**

10:15–11:55 **Paper Session**
Rackham *1st Floor*

11:55–1:15 **Lunch**

11:55–1:00 **ILTA Task Force Open Meeting**
Rackham West Conference Room *4th Floor*

11:55–1:00 ***Language Assessment Quarterly*** **Editorial Board**
Rackham East Conference Room *4th Floor*

1:15–2:55 **Poster Presentations**
Rackham Lobby

1:15–2:55 **Paper Sessions**
Rackham *1st and 4th Floors*

2:55–3:10 **Break**

3:10–4:50 **Paper Sessions**
Rackham *1st and 4th Floors*

4:50–5:00 **Break**

5:00–6:40 **Paper Session**
Rackham *4th Floor*

5:00–7:00 **Symposium**

All **Symposia** are in Rackham Auditorium, 1st Floor
All **Breaks** are in Rackham Assembly Hall, 4th Floor

Friday, June 24

8:15–5:15 **Conference Registration**
Rackham Lobby

8:15–8:30 **Announcements**
Rackham Auditorium *1st Floor*

9:00–6:00 **Exhibitors**
Rackham Assembly Hall *4th Floor*

8:30–10:10 **Paper Sessions**
Rackham *1st and 4th Floors*

10:10–10:25 **Break**

10:25–12:05 **Paper Sessions**
Rackham *1st and 4th Floors*

12:10–1:30 **Lunch**

12:10–1:30 ***Language Testing*** **Editorial Board**
Rackham East Conference Room *4th Floor*

1:30–3:10 **Works in Progress**
Rackham East and West Conference Rooms *4th Floor*

1:30–3:10 **Paper Session**
Rackham *4th Floor*

2:05–3:10 **Paper Session**
Rackham *1st Floor*

3:10–3:25 **Break**

3:25–4:30 **Paper Session**
Rackham *1st Floor*

4:00–5:10 **Paper Session**
Rackham *4th Floor*

5:05–5:15 **Break**

5:15–6:20 **Paper Session**
Rackham *4th Floor*

5:15–7:15 **Symposium**

Saturday, June 25

8:15–12:00 **Conference Registration**
Rackham Lobby

8:00–8:15 **Announcements**
Rackham Auditorium *1st Floor*

9:00–6:00 **Exhibitors**
Rackham Assembly Hall *4th Floor*

8:15–9:55 **Paper Sessions**
Rackham *1st and 4th Floors*

9:55–10:10 **Break**

10:10–11:50 **Paper Sessions**
Rackham *1st and 4th Floors*

11:50–1:25 **Lunch**

11:55–1:25 **ILTA All Members Business Meeting**
Rackham Amphitheater *4th Floor*

1:30–3:10 **Paper Sessions**
Rackham *1st and 4th Floors*

3:10–3:25 **Break**

3:25–5:25 **Symposium**

5:05–5:35 **Paper Session**
Rackham *4th Floor*

6:30–9:30 **Closing Banquet**
Michigan League Ballroom

Conference Schedule

Tuesday, June 21				
Time	North Quad (Room 1500)			
8:30–9:00	**Workshop Registration**			
9:00–5:00	Workshop **Everything You Need to Know to Teach Statistics in Language Testing Classes** James Dean Brown	Lab 1	Workshop **Applying SEM for Language Testing Research** Ardeshir Geranpayeh	Lab 2
	Rackham (Lobby)			
3:00–7:00	**Conference Registration**			

Wednesday, June 22				
Time	Rackham (Lobby)			
8:15–5:15	**Conference Registration**			
	North Quad (Room 1500)			
9:00–5:00	Workshop **Everything You Need to Know to Teach Statistics in Language Testing Classes** James Dean Brown	Lab 1	Workshop **Applying SEM for Language Testing Research** Ardeshir Geranpayeh	Lab 2
	Rackham			
1:00–5:00	ILTA Executive Board Meeting			East Conference (4th Floor)
3:00–6:00	ILTA Task Force Meeting			West Conference (4th Floor)
	University of Michigan Museum of Art (UMMA)			
6:30–8:00	**Opening Reception**			APSE

Thursday, June 23		
Time	Rackham Auditorium (1st Floor)	Rackham Amphitheatre (4th Floor)
8:15–5:15	**Conference Registration** Rackham Lobby	
9:00–6:00	**Exhibitors** Rackham Assembly Hall (4th Floor)	
8:30–8:45	**Welcome**	
8:50–10:00	**Messick Memorial Speaker Award and Lecture** Constructing Valid Performance Assessments: The View from the Shoulders of the Giants J. M. Linacre Rackham Auditorium (1st Floor)	
10:00–10:15	**Break** Rackham Assembly Hall (4th Floor)	

Conference Schedule

Thursday, June 23

Time	Rackham Auditorium (1st Floor)	Rackham Amphitheatre (4th Floor)
10:15–10:45	Session Chair Ana Lado *Marymount University* *From test item to test task: A 50-year survey of item writing* Randolph Thrasher *Okinawa Christian University*	
10:50–11:20	*The way towards a code of practice: A survey of EFL testing in China* Jinsong Fan *Shanghai Jiao Tong University* Yan Jin *Shanghai Jiao Tong University*	
11:25–11:55	*The European survey on language competences: Constructing comparable multilingual tests aligned to the CEFR* Neil Jones *University of Cambridge ESOL Examinations* Karen Ashton *University of Cambridge ESOL Examinations*	
11:55–1:15	**Lunch**	
11:55–1:00	**ILTA Task Force Open Meeting** **Quality Assurance for Language Testing, Evaluation, and Assessment** Rackham West Conference Room (4th Floor) ***Language Assessment Quarterly* Editorial Board** Rackham East Conference Room (4th Floor)	
1:15–2:55	**Poster Presentations** Rackham Lobby	
1:15–1:45	Session Chair Spiros Papageorgiou *Cambridge Michigan Language Assessments* *English language teacher educators and their assessment practices* Elizabeth Ruiz-Esparza *University of Sonora* Sofia Cota *University of Sonora*	Session Chair Gad S. Lim *University of Cambridge ESOL Examinations* *Assessing academic presentation performance: Does the rater matter?* Yoonah Seong *Teachers College, Columbia University* Elizabeth Bottcher *Teachers College, Columbia University*
1:50–2:20	*Motivation, test anxiety, and test performance within and across contexts: The CAEL, CET, and GEPT* Liying Cheng, *Queen's University* Janna Fox, *Carleton University* Don Klinger, *Queen's University* Yan Jin, *Shanghai Jiaotong University* Christine Doe, *Queen's University* Jessica R. W. Wu, *The Language Training and Testing Center*	*The impact of rater speaking proficiency level and native language on speaking test scores* Rachel Brooks *Federal Bureau of Investigation*
2:25–2:55	*A longitudinal case study of CET washback on college English classroom teaching in China* Xiangdong Gu, *Chongqing University* Xiaohua Liu, *Chongqing University* Zhiqiang Yang *Chongqing University*	*Justifying score-based interpretations from a second language oral test: Multi-group confirmatory factor analysis* Yujie Jia *University of California, Los Angeles*
2:55–3:10	**Break** Rackham Assembly Hall (4th Floor)	

Conference Schedule

Thursday, June 23

Time	Rackham Auditorium (1st Floor)	Rackham Amphitheatre (4th Floor)
3:10–3:40	Session Chair John DeJong *Pearson* *Investigating the validity of integrated listening-speaking tasks: A discourse-based analysis of test takers' oral performances* Kellie Frost *The University of Melbourne*	Session Chair Ardeshir Geranpayeh *University of Cambridge ESOL Examinations* *The diagnosis of reading in a second or foreign language: What factors are involved?* Charles Alderson, *Lancaster University* Lea Nieminen, *Jyväskylä University* Ari Huhta, *Jyväskylä University* Riikka Ullakonoja, *Jyväskylä University*
3:45–4:15	*Investigating the criterion-related validity of speaking tests* Angeliki Salamoura *University of Cambridge ESOL Examinations* Hanan Khalifa *University of Cambridge ESOL Examinations*	*Bi-literacy assessment in rural South African schools: Implications for L1 and L2 reading interface* Leketi Makalela *University of Limpopo*
4:20–4:50	*Establishing evidence of construct: A case study* Dianne Wall, *Trinity College London / Lancaster University* Cathy Taylor, *Trinity College London* Barry O'Sullivan *Roehampton University*	*Effects of multidimensionality on estimates of reading ability in passage-based assessments* Youngsoon So *University of California, Los Angeles*
4:50–5:00	**Break** Rackham Assembly Hall (4th Floor)	
5:00–5:30	**Symposium** *5:00–7:00* Session Chair Natalie Nordby Chen *Cambridge Michigan Language Assessments* **Language assessment literacy: Communicating theories of language testing to users** **Organizer** Ofra Inbar-Lourie *Tel-Aviv University / Beit Berl College* **Discussant** Lynda Taylor *University of Cambridge ESOL Examinations* *Assessment literacy as LSP* Cathie Elder *University of Melbourne* April Ginther *Purdue University* *Operationalizing assessment literacy* Glenn Fulcher *University of Leicester* *The essentials of assessment literacy in a post-Messick world: Contrasts between testers and users* Margaret E. Malone *Center for Applied Linguistics* *Language assessment literacy among prospective language teachers: From novice to proficient* Ofra Inbar-Lourie *Tel-Aviv University / Beit Berl College*	Session Chair Micheline Chaloub-Deville *University of North Carolina at Greensboro* *The use of source texts in integrated writing assessment tasks* Lia Plakans *The University of Iowa* Atta Gebril *The American University in Cairo*
5:35–6:05		*Investigating the comparability of TOEFL® iBT integrated writing task prompts* Yeonsuk Cho *Educational Testing Service* Jakub Novak *Educational Testing Service* Frank Rijmen *Educational Testing Service*
6:10–6:40		*Assessing authorial voice strength in L2 argumentative writing* Cecilia Guanfang Zhao *Shanghai International Studies University*

Conference Schedule

Friday, June 24		
Time	Rackham Auditorium (1st Floor)	Rackham Amphitheatre (4th Floor)
8:15–5:15	**Conference Registration** Rackham Lobby	
8:15–8:30	**Announcements** Rackham Auditorium (1st Floor)	
9:00–6:00	**Exhibitors** Rackham Assembly Hall (4th Floor)	
8:30–9:00	Session Chair April Ginther *Purdue University* *Balancing practicality and construct representativeness for IEP speaking tests* Sawako Matsugu *Northern Arizona University* Anthony Becker *Northern Arizona University* Mansoor Al-Surmi *Northern Arizona University*	Session Chair Jayanti Banerjee *Cambridge Michigan Language Assessments* *Extending the scope of the yes/no vocabulary test format* John Read *University of Auckland* Toshihiko Shiotsu *Kurume University*
9:05–9:35	*Investigating the construct validity of a listening-to-retell test in national matriculation English test in China* Fenxiang Cheng *Guangdong University of Foreign Studies*	*Modeling vocabulary knowledge: A mixed model approach* Wen-Ta Tseng *National Taiwan Normal University*
9:40–10:10	*Variables influencing the communicative competence of second/foreign language speakers* Thomas Dinsmore *University of Cincinnati / Clermont College*	*The "I don't know" option in vocabulary size test* Xian Zhang *Penn State University*
10:10–10:25	**Break** Rackham Assembly Hall (4th Floor)	
10:25–10:55	Session Chair Lynda Taylor *University of Cambridge ESOL Examinations* *Developmental considerations for employing task-based paired assessment among young learners of foreign language* Yuko Butler *University of Pennsylvania* Wei Zeng *University of Pennsylvania* Yeting Liu *University of Pennsylvania*	Session Chair Randy Thrasher *Okinawa Christian University* *Building a theoretical foundation for test impact research* Dayong Huang
11:00–11:30	*Assessing interaction skills: Evidence from Finnish and Swedish L2 national exams from grade 9* Marita Härmälä *University of Jyväskylä* Outi Toropainen *University of Jyväskylä*	*Building a construct validity argument for the GEPT reading test: A confirmatory factor analysis approach* Hsinmin Liu *University of California, Los Angeles*
11:35–12:05	*Classroom-based assessment of the effects of instruction and children's strategy use on success in ESL* Pamela Gunning *McGill University and Concordia University*	*Item features and construct of academic English* Shu Jing Yen *Center for Applied Linguistics* Ying Zhang *Center for Applied Linguistics* David MacGregor *Center for Applied Linguistics*

Conference Schedule

Friday, June 24

Time	Rackham Auditorium (1st Floor)	Rackham Amphitheatre (4th Floor)
12:10–1:30	***Language Testing* Editorial Board** Rackham East Conference Room (4th Floor)	
	Lunch	
1:30–3:10	**Works in Progress** Rackham East and West Conference Rooms (4th Floor)	
1:30–2:00		Session Chair Deborah Crusan *Wright State University* ***Assessing written English skills for business communication using time-constrained tasks*** Alistair Van Moere, *Knowledge Technologies, Pearson* Masanori Suzuki, *Knowledge Technologies, Pearson* Ryan Downey, *Knowledge Technologies, Pearson* Mallory Klungtvedt, *Knowledge Technologies, Pearson*
2:05–2:35	Session Chair Anthony Green *University of Bedfordshire* ***Language aptitude, first language ability, and second language proficiency: A meta-analytic investigation*** Kathryn Nelson, *SWA Consulting Inc.* Amy DuVernet, *North Carolina State University* Eric Surface, *SWA Consulting Inc.* Milton Cahoon, *SWA Consulting Inc.*	***The importance of discourse-based research in LSP test validation*** Gene Halleck *Oklahoma State University* Carol Moder *Oklahoma State University*
2:40–3:10	***Past, present, and future of language assessment courses: Are we headed into the right direction?*** Heejeong Jeong *University of Illinois at Urbana Champaign*	***Cognitive validity of an ESP test: Cognitive processing during reading comprehension*** Ivana Vidakovic *University of Cambridge ESOL Examinations* Hanan Khalifa *University of Cambridge ESOL Examinations*
3:10–3:25	**Break** Rackham Assembly Hall (4th Floor)	
3:25–3:55	Session Chair Ching-Ni Hsieh *Cambridge Michigan Language Assessments* ***Aligning internal and external validation of a high-stakes ESL exam*** May Tan *McGill University* Carolyn Turner *McGill University*	Session Chair Tineke Brunfaut *Lancaster University* ***The generalizability of scoring keys for the computer automated scoring of web-based language tests*** Nathan Carr *California State University, Fullerton*

Conference Schedule

Friday, June 24

Time	Rackham Auditorium (1st Floor)	Rackham Amphitheatre (4th Floor)
4:00–4:30	*AFL: Focus on formative assessment in an L2 classroom* Christian Colby-Kelly *McGill University*	*Investigating the dependability of analytic scoring for a computer-based oral test* Yujie Jia *University of California, Los Angeles* Alan Urmston *The Hong Kong Polytechnic University* Felicia Fang *The Hong Kong Polytechnic University*
4:35–5:05		*Is computer literacy construct-relevant in a language test in the 21st century?* Yan Jin *Shanghai Jiaotong University* Jiang Wu *Shanghai Jiaotong University* Ming Yan *Heilongjiang University*
5:05–5:15	**Break** Rackham Assembly Hall (4th Floor)	
5:15–5:45	**Symposium** *5:15–7:15* Session Chair Sarah Briggs *Cambridge Michigan Language Assessments* **Rating scales for writing/speaking: Issues of development and use** **Organizers** Evelina D. Galaczi *University of Cambridge ESOL Examinations* Gad S. Lim *University of Cambridge ESOL Examinations* **Discussant** Sara Weigle *Georgia State University* *From framework to scale: Integrating the CEFR into an operational rating scale* Anthony Green *University of Bedfordshire* *Meeting multiple validation requirements in rating scale development* Gad S. Lim *University of Cambridge ESOL Examinations* *Stakeholder engagement in test design and rubric development* India C. Plough *Cambridge Michigan Language Assessments* *Raters as stakeholders: Uptake in the context of diagnostic assessment* Janna Fox *Carleton University* Natasha Artemiva *Carleton University*	Session Chair Yong-Won Lee *Seoul National University* *Do tests promote changes in receptive skills? A comparative study in a Taiwanese EFL context* Yi-Ching Pan *National Pingtung Institute of Commerce*
5:50–6:20		*Policy and practice in EFL contexts: Investigating assessment in oral communication courses in Japan* Rika Tsushima *McGill University*
6:25–7:15		

Conference Schedule

Saturday, June 25

Time	Rackham Auditorium (1st Floor)	Rackham Amphitheatre (4th Floor)
8:15–12:00	**Conference Registration** Rackham Lobby	
8:00–8:15	**Announcements** Rackham Auditorium (1st Floor)	
9:00 –6:00	**Exhibitors** Rackham Assembly Hall (4th Floor)	
8:15–8:45	Session Chair Ildi Porter-Szucs *Cambridge Michigan Language Assessments* *EIKEN program testing context analysis* James Dean Brown *University of Hawai'i at Mānoa* Keita Nakamura *Society for Testing English Proficiency*	Session Chair Elvis Wagner *Temple University* *Effects of presenting question stems, answer options or neither on multiple-choice listening comprehension tests* Gary Ockey *Kanda University of International Studies* Angela Sun *Kanda University of International Studies* Dennis Koyama *Kanda University of International Studies*
8:50–9:20	*Relationship of TOEFL scores to success in American universities* Brent Bridgeman *Educational Testing Service* Yeonsuk Cho *Educational Testing Service*	*The validity of using integrated tasks to assess listening skills* Ying Zheng *Pearson* John H.A.L. De Jong *Pearson / VU University Amsterdam*
9:25–9:55	*Time lag as a mediator in determining language proficiency: Graduate vs. undergraduate levels* Kate Kokhan *University of Illinois at Urbana-Champaign*	*EAP listening task difficulty: The impact of task variables, working memory and listening anxiety* Tineke Brunfaut *Lancaster University* Andrea Revesz *Lancaster University*
9:55–10:10	**Break** Rackham Assembly Hall (4th Floor)	
10:10–10:40	Session Chair Paul Winke *Michigan State University* *A survey of methodological approaches employed to validate language assessments: 1999–2009* Ikkyu Choi *University of California, Los Angeles* Jonathan Schmidgall *University of California, Los Angeles*	Session Chair Toshihiko Shiotsu *Kurume University* *Investigating the use of academic vocabulary and its effect on test taker performance* Kirsten Ackermann *Pearson*
10:45–11:15	*Validation of a large-scale assessment for diagnostic purposes across three contexts: Scoring, teaching, and learning* Christine Doe *Queen's University*	*Testing spelling: An investigation into non-native test takers' spelling errors* Catherine Hayes *Birkbeck College & Pearson Language Tests*
11:20–11:50	*Validity issues in the placement testing of language minority students in community colleges* Lorena Llosa *New York University* George Bunch *University of California, Santa Cruz*	*Assessing lexical richness in spontaneous speech data* Andrea Hellman *Missouri Southern State University*
11:50–1:25	**Lunch**	
11:55–1:25	**ILTA All Members Business Meeting** Rackham Amphitheater (4th Floor)	

Conference Schedule

Saturday, June 25

Time	Rackham Auditorium (1st Floor)	Rackham Amphitheatre (4th Floor)
1:30–2:00	Session Chair May Tan *McGill University* *Computer assisted English speaking testing: A case study of CEOTS, China* Xin Yu *University of Bath*	Session Chair Natalie Nordby Chen *Cambridge Michigan Language Assessments* *Assessment literacy: A call for more teacher education reform and professional development* Masoomeh Estaji *Allameh Tabataba'i University*
2:05–2:35	*Examinee perceptions of automated scoring of speech and validity implications* Xiaoming Xi, *Educational Testing Service* Yuan Wang, *Educational Testing Service* Jonathan Schmidgall, *University of California, Los Angeles*	*Classifying and reporting language proficiency patterns to inform integrative language teaching* Huan Wang *CTB / McGraw Hill*
2:40–3:10	*Examining the validity of an elicited imitation instrument to test oral language in Spanish* C. Ray Graham *Brigham Young University*	*Impact of rater fatigue on the scoring of speaking responses* Ling Guanming, Pamela Mollaun, and Xiaoming Xi *Educational Testing Service*
3:10–3:25	**Break** Rackham Assembly Hall (4th Floor)	
3:25–5:05	**Symposium** *3:25–5:25* Session Chair Elana Shohamy *Tel Aviv University* **Workforce language assessment in the 21st Century knowledge economy: A quality management perspective** **Organizer** Kathleen M. Bailey *Monterey Institute of International Studies* **Discussant** Jean Turner *Monterey Institute of International Studies* *Surveying English language assessment practices in international pluralingual organizations* Kathleen M. Bailey *Monterey Institute of International Studies* Ryan Damerow *The International Research Foundation for English Language Education* Courtney Pahl *Monterey Institute of International Studies* *Using quality management to improve language testing* Michael Milanovic *University of Cambridge ESOL Examinations* Nick Saville *University of Cambridge ESOL Examinations* *Issues in assessing workplace ESL instructional programs* Mary Ann Christison *University of Utah*	**ILTA Task Force Open Meeting** *Quality Assurance for Language Testing, Evaluation, and Assessment* Dan Douglas, Jamie Dunlea, Liz Hamp-Lyons, Yasuyo Sawaki, Diane Schmitt, Bernard Spolsky, Lynda Taylor
5:05–5:35		Session Chair India C. Plough *Cambridge Michigan Language Assessments* *A discourse analysis of the pragmatic meanings generated on a role-play speaking test* Kirby C. Grabowski *Teachers College, Columbia University*
6:30–9:30	**Closing Banquet and Awards Ceremony** Michigan League Ballroom	

Randolph Thrasher
Okinawa Christian University

Thursday, June 23
10:15–10:45
Rackham Auditorium (1st Floor)

From test item to test task: A 50-year survey of item writing

The publication of Robert Lado's *Language Testing* was the start of an ever-increasing flow of books discussing the sorts of items considered appropriate for testing the various language skills. Since Lado's book is a mixture of theory and practical advice to test writers, both primarily theory oriented and "how-to" books from this outpouring are sampled. This paper surveys the selected books in an attempt to determine what we have learned about item writing in the last half century and what we have yet to fully understand. The survey, of course, reflects the various bandwagons that we have jumped upon, but, at the same time, it demonstrates a deepening and broadening of our understanding of what various item types can and cannot do. It also shows the increasing recognition of the role of testing theory in determining the sorts of test tasks deemed appropriate. The survey reveals several trends such as the growing awareness of the necessity of explicitly stated item specifications and a widening of the statistical procedures used in item analysis. The impact of technological change as it can be seen in the books sampled is discussed. This survey also provides the opportunity to reflect on the elements of Lado's theory that have been largely abandoned and those that have been retained. But the author believes that the greatest contribution of such a survey is that it gives us the chance to see what our understanding of item writing has been, to determine the degree to which we have a consensus today, and what the issues are that need to be resolved as we move into the future.

Jinsong Fan
Shanghai Jiao Tong University

Yan Jin
Shanghai Jiao Tong University

Thursday, June 23
10:50–11:20
Rackham Auditorium (1st Floor)

The way toward a code of practice: A survey of EFL testing in China

This study reports a survey of the large-scale standardized EFL tests in China with respect to test development, administration, and use. Adopting a three-phase design and synthesizing the views from both test developers and the primary stakeholders, the study is intended to paint a representative picture of the current EFL testing practices in China. The participants of this study are the six predominant EFL examination boards in China and the primary stakeholder groups including 145 EFL teachers and 280 students from different regions of the country. Phase 1 is document analysis, intended to investigate the availability of test-related information to the general public and the nature and amount of the accessible information. Phase 2 is an interview with the six predominant EFL examination boards in China, with a view to finding out at the general level their practices in developing, administering, and validating their tests and their views about the uses and misuses of their tests. Phase 3 involves three questionnaires, administered to the six EFL examination boards, the 145 EFL teachers, and 280 students respectively. Both descriptive and inferential statistical analysis of the data are performed. The findings of the data indicate that examination boards on the whole follow their own quality control procedures in developing, administering, and validating their tests. But the validity of these procedures is open to question. Misuses of EFL tests and test results are identified as rampant, which constitute a serious threat to test validity. Students and teachers express both positive and negative feelings about the EFL tests. The study awakens China's language testers to the importance and urgency of developing a code of practice that is applicable to the China's EFL testing context and also calls for more communication between test developers and stakeholders.

Neil Jones
University of Cambridge
ESOL Examinations

Karen Ashton
University of Cambridge
ESOL Examinations

Thursday, June 23
11:25–11:55
Rackham Auditorium (1st Floor)

The European survey on language competences: Constructing comparable multilingual tests aligned to the CEFR

The European Survey on Language Competences (ESLC) is a groundbreaking new European initiative, intended to further the aims of the Lisbon Strategy and promote the goal of a multilingual Europe. A field trial was successfully completed in 2010, and when LTRC meets, the main data collection for the survey will be complete. The ESLC will provide an indicator of the foreign language proficiency of school pupils across Europe, in the five most widely studied languages: English, French, German, Italian, and Spanish. It is intended to inform policy makers, teachers, and practitioners in improving language teaching methods and raising levels of achievement. The survey is technically challenging, involving the administration of targeted tests at three levels, in both computer- and paper-based modes. This presentation will briefly review the project and then discuss the major challenges of the survey from a language testing point of view: the development of a valid approach to constructing closely comparable tests in five languages, testing the skills of reading, listening and writing, and the alignment of these to a common standard in relation to the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) levels. Despite the prominence of the CEFR and the finalization of a Council of Europe manual for aligning exams to it, very little work to date has explicitly focused on the validity of cross-language comparisons or addressed the recognized issue that understandings of proficiency levels are to an extent culturally determined. It is our intention that the ESLC language tests and the methodology developed for their validation should contribute usefully to making progress in this area.

Elizabeth Ruiz-Esparza
University of Sonora

Sofía Cota
University of Sonora

Thursday, June 23
1:15–1:45
Rackham Auditorium (1st Floor)

English language teacher educators and their assessment practices

This questionnaire-based study was carried out in a bachelor of arts in English language teaching program in Mexico, in a state bordering the United States. The focus of the study was to find out: (a) the teachers' educational background and studies about testing, (b) their perceptions about knowledge and skills in testing and, (c) the frequency of use of a variety of assessment practices. The practices under investigation were adapted and narrowed from a list identified in Zhang's and Burry-Stock's (2003) research. The sample of the study was the entire cohort of academics working in the program. The response rate was a 100 percent. The SPSS program for Windows was used to analyze the teacher responses. Results evidenced that the majority of the teacher educators had theoretical knowledge about assessment. However, one of the salient findings was that more than half of these educators reported they were unfamiliar with and not skilled in assessment practices that have to do with ensuring that their assessments provide valid information. The study also evidenced that there are assessment practices that teachers consider of utmost importance over other practices. The implication of the study is the need for a series of professional development courses on assessment for the teachers to develop confidence in designing, applying and grading their own instruments. This research aims to contribute to the literature about the importance of teacher preparation in testing in regards to the issues of validity and reliability. Moreover, the study also presents information about a context that has been largely unexplored, that of a bachelor of arts in English language teaching in Mexico.

Yoonah Seong
Teachers College,
Columbia University

Elizabeth Bottcher
Teachers College,
Columbia University

Thursday, June 23
1:15–1:45
Rackham Amphitheatre (4th Floor)

Assessing academic presentation performance: Does the rater matter?

Several studies have found differences in teacher and nonteacher raters' attention and scoring behavior (e.g., Brown, 1995; Elder, 1993; Lumley, 1998). The findings all indicate that raters do not always interpret student performance and scoring criteria a similar way. This study, which examines academic presentation performances, yields similar results. The oral presentation was a final project requirement for two English for Academic Purposes (EAP) courses at two different American universities. One class was an elective course for intermediate ESL learners in a nondegree intensive ESL program. The other class was for advanced-level international students already matriculated in the university. Thus, the second class was a mandatory course that counted towards the students' English language requirements. The presentations were videotaped and scored by two raters: (1) the teacher who taught both classes and (2) an ESL instructor with no EAP teaching experience. Using FACETS based on the Rasch model, the examinees' performance on this test task, the relative difficulty of the rubric components (content, organization, visual, delivery, language) upon which the students were measured, and most importantly, the raters' influence on the scoring process were examined. This study shows that raters' bias towards certain examinees and contradicting patterns of rater interaction with classes existed. Postanalysis discussion between the raters revealed that the raters' awareness of the class context (e.g., class configuration, test purpose, and students) could have affected the raters' scoring behavior and pattern. Although there is a need for further investigation, the results of this study raise important questions regarding the validity of teacher-based assessment in a classroom context.

Liying Cheng
Queen's University

Don Klinger
Queen's University

Christine Doe
Queen's University

Janna Fox
Carleton University

Yan Jin
Shanghai Jiaotong University

Jessica R. W. Wu
The Language Training
and Testing Center

Thursday, June 23
1:50–2:20
Rackham Auditorium (1st Floor)

Motivation, test anxiety, and test performance within and across contexts: The CAEL, CET, and GEPT

Research suggests that test takers' motivation and test anxiety are significant cognitive factors associated with their test performance (Gardner, 1985; Horwitz, 2001; MacIntyre, 2002). However, the interaction of these two factors has yet to be investigated in relation to language test performance in high-stakes contexts. Previous research (e.g., Horwitz, 2001; Horwitz & Young, 1991; MacIntyre, 2002; Noel, et al, 2000) has only examined these two factors separately as individual cognitive processes in relation to a single context and among one group of test takers (i.e., test takers who share a common first language). This study examined test takers' motivation, test anxiety, and language test performance within and across a range of social and educational contexts on three high-stakes language tests—the Canadian Academic English Language (CAEL) Assessment in Canada, the College English Test (CET) in mainland China, and the General English Proficiency Test (GEPT) in Taiwan. A questionnaire was issued to test takers across these testing contexts (CAEL = 255; CET = 495; GEPT = 533). Test takers' demographic information, motivational orientations, test anxiety, and perceptions of the uses and stakes of the tests were collected. Questionnaire responses were then linked to test takers' respective language test performance. The results illustrate complex interrelationships amongst test takers' motivation, test anxiety, and test performance within each testing context and across the three social and educational contexts where the testing took place. There were differences in motivation and test anxiety based on the test contexts in relation to both personal variables (gender) and social variables (test importance to stakeholders and test purposes). Further, motivation and test anxiety along with personal factors (gender and age) were associated with test performance. Examining the interrelationships within and across the three testing contexts begins to address a gap in the research literature, as motivation and test anxiety have typically been examined in isolation and in relation to a single test context.

Rachel Brooks
Federal Bureau of Investigation

Thursday, June 23
1:50–2:20
Rackham Amphitheatre (4th Floor)

The impact of rater speaking proficiency level and native language on speaking test scores

Language testing researchers have conducted studies on rater characteristics that significantly impact test scores. Research to date has indicated that raters' scores are affected by rater attributes such as age, gender, occupation, international experience, personality, cultural background, and opinion (Barnwell, 1989b; Eckes, 2008; Galloway, 1980; Kang, 2008; Ludwig, 1982; Reed & Cohen, 2001). Previous research has overlooked other aspects of raters' language abilities that may affect rating reliability: speaking proficiency level and the relationship between the rater's first language and the language being evaluated. This study investigated the impact of these two rater variables on an integrated skills test, the Speaking Proficiency Test (SPT). Participants were FBI SPT raters (n = 95), including 15 English raters and 80 raters of approximately 40 other languages. These raters evaluated four SPTs, including examinees of various proficiency levels and linguistic backgrounds. Rater notes, scores, justifications, and post-test discussions were coded for analysis. Raters were divided into four groups by their speaking proficiency ratings for the first analysis, and two groups by language family (same as of different from examinee) for the second. Data were analyzed using MANOVAs of final scores and subscores across six rating categories (based on language elements such as grammar, vocabulary, pronunciation, cultural appropriateness, and fluency), along with qualitative analyses of rater comments by language element. Power analysis revealed that 95 subjects divided in four groups with 4 repetitions, an effect size of f = 0.2, and an alpha of 0.05 had a power of 0.985. Initial results reveal a significant effect of SPT score on rater comparability and a limited effect based on language family relationship. Results support the selection of raters using speaking proficiency score rather than a demographic variable, such as native speaker. Moreover, uneven results in the relationship between rater and examinee language family warrants further investigation.

Xiangdong Gu
Chongqing University

Zhiqiang Yang
Chongqing University

Xiaohua Liu
Chongqing University

Thursday, June 23
2:25–2:55
Rackham Auditorium (1st Floor)

A longitudinal case study of CET washback on college English classroom teaching in China

College English (CE) is a compulsory course for all non-English major undergraduate students in China. A new teaching syllabus for the course was issued in 2004. In accordance with the change of the teaching objective in the new syllabus, the National College English Test (CET) launched an innovation in 2005. The present study aims to explore the washback of the innovated CET on CE classroom teaching in China by revisiting three CE teachers' classrooms in 2009. The three teachers' classrooms were observed both in 2003 and 2009 by coding observation scheme and taking field notes, accompanied with video recordings and interviews. By comparing their classroom teaching in 2009 with that in 2003, the present study outlines similarities and differences between CE classroom teachings before and after the CET innovation, and then endeavors to identify main factors contributing to such similarities and differences. Research findings show that the essential mode of CE classroom teaching remained the same over the years (teacher-dominated), but obvious changes have also been found in teaching plans, teaching content, and teaching methods. The CET innovation seems to be one of the main factors contributing to such changes. Other factors, however, such as teaching objective in the new teaching syllabus, materials selected in the course books, administrators' course design, teachers' individual differences, and students' proficiency and classroom participation also play roles in such changes. Due to the small sample of subjects observed and interviewed, whether the research results can be generalized remains to be validated. However, the present study is significant in that it enriches and expands theoretical models in the field of washback study by way of constructing a triangle model, a two-hierarchy model, and a main factor model illustrating the complexity of washback studies. [This paper is part of the research project "A Longitudinal Study of the CET Washback" supported by the National Philosophy and Social Science Foundation of China (07BYY030) and by the National Research Centre for Foreign Language Education (MOE Key Research Institute of Humanities and Social Sciences at Universities), Beijing Foreign Studies University.]

Yujie Jia
University of California,
Los Angeles

Thursday, June 23
2:25–2:55
Rackham Amphitheatre (4th Floor)

Justifying score-based interpretations from a second language oral test: Multi-group confirmatory factor analysis

This study used Bachman and Palmer's (2010) Assessment Use Argument to investigate the claim that rating-based interpretations from an ESL oral test are meaningful and impartial across different groups of test takers. Candidates' oral responses to five integrated speaking tasks in this computer-based oral test were rated on five dimensions. To provide backing for the meaningfulness of interpretations, confirmatory factor analysis (CFA) was used to analyze 537 candidates' scores. Several CFA models were tested and compared. Higher-order trait-uncorrelated method model was selected as the final one because it could explicate the relationships among five dimensions and overall speaking proficiency. The trait factors were found to have larger factor loadings on the analytic scores than the method factors, providing evidence to support the claim that the test scores could be meaningfully interpreted as indicators of five dimensions. The large factor loadings of the higher-order speaking ability factor on five trait factors also provided backing for reporting of one composite score. To investigate the impartiality of interpretations, two multi-group CFAs were conducted to examine the extent to which the factor structure of the test varied across gender and disciplines. The means of the five trait factors and the higher-order factor were compared between males and females by specifying factorial invariance between these two groups. The findings indicated that females performed significantly better than males on five dimensions and overall speaking ability. The multi-group analysis of business and nonbusiness students showed that the factor structure for these two groups was significantly different and the oral test may measure nonbusiness students' five dimensions and overall speaking ability better than business students. Task analysis of the oral test was conducted to better interpret the results of the statistical analysis. The theoretical and practical implications for the design and development of oral tests were discussed.

Kellie Frost
The University of Melbourne

Thursday, June 23
3:10–3:40
Rackham Auditorium (1st Floor)

Investigating the validity of integrated listening-speaking tasks: A discourse-based analysis of test takers' oral performances

Performance on integrated tasks requires candidates to engage skills and strategies beyond language proficiency alone, in ways that can be difficult to define and measure for testing purposes. While it has been widely recognized that stimulus materials impact test performance, our understanding of the way in which test takers make use of these materials in their responses, particularly in the context of listening-speaking tasks, remains predominantly intuitive, with little or no base in empirical evidence. Limited studies to date on integrated speaking tests have highlighted the problems associated with content related aspects of task fulfilment (Brown et al.(2005), TOEFL Monograph Series MS-29; Lee (2006), Language Testing, 23: 131), but little attempt has been made to operationalize the way in which content from the input material is integrated into speaking performances. Using discourse data from a trial administration of the new Oxford English Language Test, this paper investigates the way in which test takers integrate stimulus materials into their speaking performances on an integrated listening then speaking summary task, and examines if test scores reflect real differences in the quality of oral summaries produced. An innovative discourse analytic approach was developed to analyze content-related aspects of performance in order to determine if such aspects represent an appropriate measure of the construct of speaking ability. Results showed that the quantity and quality measures devised to operationalize content, such as the number of key points included from the input text, and the accuracy with which information from the input text was reproduced or reformulated, effectively distinguished participants according to their level of speaking proficiency, indicating that these discourse-based measures have the potential to be applied to other integrated tasks and in other assessment contexts.

Charles Alderson
Lancaster University

Ari Huhta
Jyväskylä University

Lea Nieminen
Jyväskylä University

Riikka Ullakonoja
Jyväskylä University

Thursday, June 23
3:10–3:40
Rackham Amphitheatre (4th Floor)

The diagnosis of reading in a second or foreign language: What factors are involved?

The ability to read in a foreign language (FL) is of growing importance in a globalized world and therefore diagnostic language testing is potentially an important area of language test development and research. However, what are claimed to be diagnostic tests are often little more than placement tests, and are very rarely based on a theory of language learning, or a theory of diagnosis. Moreover, several recent studies have merely attempted to retrofit diagnostic information to tests that have been developed as proficiency tests. This paper reports on the first study in a 4-year (2010–2013) research project into the diagnosis of reading in L2, which is studying learners of English in three different age groups: Grade 4, aged 10–11, Grade 8, aged 14–15, Second Year of upper secondary school (aged 17–18). This first study explores the diagnostic potential of a range of cognitive and psycholinguistic measures (e.g., working memory, phonological processing, ability to process nonwords) as used for detecting L1 dyslexia. These were delivered in informants' first and second languages, in order to examine their applicability for L2 diagnosis. We also included measures of L2 vocabulary, motivation and background information on the informants, and we examine the relationship of all variables to measures of first and foreign language reading abilities. The results have major implications for the methodology of the second and third studies, which are longitudinal and interventionist in nature, as well as for the development of a theory of second language reading development and diagnosis.

Angeliki Salamoura
University of Cambridge
ESOL Examinations

Hanan Khalifa
University of Cambridge
ESOL Examinations

Thursday, June 23
3:45–4:15
Rackham Auditorium (1st Floor)

Investigating the criterion-related validity of speaking tests

Criterion-related validity refers to the extent to which a test correlates with a suitable external measure of performance (Anastasi, 1988:145, Messick, 1989:16), e.g., an older, established test or a framework which is believed to be a measure of the same construct. Criterion-related validity is, in other words, concerned with a posteriori and external validity evidence based on data. Although a number of language testers acknowledged its significance early on and argued that external validation based on data is always to be preferred over other types of validity (Anastasi, 1982; Davies, 1983), the concept of criterion-related validity has been given a new lease of life lately. As a result of the multiplication of tests on offer and the growing influence of external standards, such as the Common European Framework of Reference (CEFR; Council of Europe, 2001), great importance is currently attached to criterion-related validity and test comparability issues by both test users and providers (Khalifa and Weir, 2009). Against such a background, this presentation will discuss the practices adopted by a large examinations provider in establishing and, critically, maintaining criterion-related evidence within the context of its speaking tests. Following Weir's (2005) evidence-based model of test validation, we will examine criterion-related validity issues from three perspectives, illustrating with case studies: (i) cross-test comparability with other providers' tests, (ii) equivalence with different versions of the same test, and (iii) comparability with external standards, such as the CEFR. Although criterion-related validity has traditionally been perceived as a posteriori test validation, the presentation will argue that within the context of an international examination board, it would better be approached not as a one-off exercise but as an integral part of the lifecycle of a test—from inception to the construction of parallel versions and future revisions. Implications of such an approach will be discussed.

Leketi Makalela
University of Limpopo

Thursday, June 23
3:45–4:15
Rackham Amphitheatre (4th Floor)

Bi-literacy assessment in rural South African schools: Implications for L1 and L2 reading interface

A plethora of assessment studies on literacy development show that the majority of South African learners cannot read at the expected proficiency levels commensurate with their grade levels both in English and in their home languages. While a considerable amount of research on the role of L2 proficiency and L1 in the development of L2 reading in the western countries has been substantially carried out, there is a dearth of similar studies that assessed the effects of a reading interface between two distal language systems in Africa. This paper reports on a continuous assessment measures that sought to determine reading proficiency levels based on vocabulary, reading comprehension, and reading speed equivalents in both English and Sepedi (a South African language) from a data pool of 950 systematically (random) sampled grade 6 learners. Using both MANOVA and multiple regression analysis of various achievement scores, the results show that (i) the learners are at least four years below their expected proficiency in English and in home language, and that (ii) there is a statistically significant higher reading variance in English than in the home language. These rather surprising results further revealed that there are differential cognitive language proficiency levels between English and the African language, with English having high gains due to its position of prestige as the language of learning and teaching in the education system. Interpreted within both the linguistic interdependence hypothesis and the linguistic threshold hypothesis to arrive at a systematic account for such a variance, several implications for long-term cognitive language and literacy development in African languages are drawn for comparison with related contexts. Finally, useful recommendations on testing for language proficiency and literacy development, respectively, in marginal languages and other comparable situations are highlighted at the end of the paper.

Dianne Wall
Trinity College London,
Lancaster University

Barry O'Sullivan
Roehampton University

Cathy Taylor
Trinity College London

Thursday, June 23
4:20–4:50
Rackham Auditorium (1st Floor)

Establishing evidence of construct: A case study

This presentation will describe the approach taken by a European testing body to gather empirical evidence of the construct underlying its main speaking test. Although the testing body considers that the speaking test assesses communicative ability, it has not until recently made explicit the model of language underlying the test or systematically investigated whether the tasks set for the different levels of the test and the criteria used for marking performance work together to support the claims made by the test developers. The approach taken was based on a model of validation developed by Weir (2005) and refined by O'Sullivan and Weir (2010). A series of tables was created in which criterial parameters associated with the test taker, test task, and scoring system were identified. These tables were used by a panel of internal and external language testing specialists to analyze the tasks used at all levels of the speaking test. The panel members studied the test specification and user handbooks, and scrutinized a number of test performances at each of the test levels in order to inform their decisions. The approach proved successful in confirming the construct underlying the speaking test, and it revealed how different aspects of the construct are assessed as the tasks change at different test levels. Tasks at the lower levels require only a limited amount of linguistic competence, while tasks at higher levels demand not only more linguistic ability, but also sociolinguistic, discourse and strategic competences. The presentation will present the findings of this validation process and comment on the strengths of the approach and some areas in need of improvement. We will also discuss some of the implications of the findings, including ideas for further research into the expansion of the construct as candidates progress through the test levels.

Youngsoon So
University of California,
Los Angeles

Thursday, June 23
4:20–4:50
Rackham Amphitheatre (4th Floor)

Effects of multidimensionality on estimates of reading ability in passage-based assessments

This study investigated the effects of multidimensionality on estimates of reading ability in passage-based assessments. The data analyzed in the study came from the 4,858 test takers' responses to the reading paper of Certificate in Advanced English (CAE), administered in December 2002. A previous study conducted by the author (2010) indicated that the data can be best explained by a multidimensional factor structure, more specifically by a bi-factor model, in which one trait factor and a set of passage factors explain the inter-item relationships on the test. As an extension of this previous study, the present study applied two different item-response theory (IRT) models to estimate test takers' reading ability—a traditional unidimensional model and a bi-factor model. Comparisons of the estimates from the two IRT models clearly demonstrated that applying unidimensional IRT models to multidimensional test data yield biased estimates of both item parameters and test takers' ability. The finding that test takers were ranked differently by different models calls for a particular attention to making use of test scores in order to make high-stakes decisions about test takers. This study addresses an issue that is of considerable importance both theoretically and practically for language assessment. It is of theoretical interest in that it tested the viability of current measurement models for estimating test takers' ability levels in a passage-based comprehension assessment on the basis of sets of item scores that violate the mathematical assumptions of these models. This study is also of great practical significance, because virtually all high-stakes large-scale tests of ESL reading comprehension utilize the method of presenting test takers with a passage and then asking them comprehension questions based on these passages. In conclusion, the findings provide implications for choosing a measurement model that can take into account the factor structure of the assessment data.

Lia Plakans
The University of Iowa

Atta Gebril
The American University
in Cairo

Thursday, June 23
5:00–5:30
Rackham Amphitheatre (4th Floor)

The use of source texts in integrated writing assessment tasks

Integrating skills is a current trend in language assessment but has been a part of the language testing landscape since Lado's book *Language Testing*. The study detailed in this paper considers an important issue in integrated writing tests—the role of the texts provided for reading and listing in test taker's written performance. How writers integrate content from the source texts in their writing holds implications for development, score interpretation, and test use. Drawing on writing from 480 reading-listening-writing iBT TOEFL tasks, source use was analyzed in four ways. First, the integration of key ideas from the source texts was investigated by ranking the ideas in each text and calculating their appearance in written responses. Then writing was analyzed to find which source text ideas originated from, the reading or listening text. The style of integrated was considered to determine if writers used the source content implicitly by paraphrasing or summarizing or explicitly by quotation or verbatim borrowing of words. Related to the integration style, an analysis was conducted for verbatim source use, defined as borrowing of three or more words in sequence from either source text. These four issues in source text use were considered across five score levels on two tasks using analysis of variance and follow-up pair-wise comparisons. The results indicate significant differences in how low scoring writers utilized sources compared to higher scoring writers. Low scoring writers showed more reliance on the reading source texts then other writers and borrowed words from the source texts more than others. The implications for the integration of skills in assessment include the differential utilization of reading and listening source content and the challenge for test takers to successfully incorporate source materials. These issues should be considered in scoring and using assessments that draw on multiple skills.

Yeonsuk Cho
Educational Testing Service

Jakub Novak
Educational Testing Service

Frank Rijmen
Educational Testing Service

Thursday, June 23
5:35–6:05
Rackham Amphitheatre (4th Floor)

Investigating the comparability of TOEFL®, iBT™ integrated writing task prompts

This study examined the comparability of the TOEFL, iBT integrated writing task across test forms by investigating the relationship between prompt characteristics and average scores of integrated writing tasks. The integrated writing task requires examinees to provide a written response summarizing and synthesizing information from an academic reading text and a lecture. While the integrated writing task is carefully designed and developed to ensure task comparability, the average scores of the integrated writing tasks showed more variation across forms than those of the TOEFL iBT independent writing tasks. This observation suggests the need to look into factors contributing to the observed variation in the average test scores of the integrated writing tasks. The study focused on the following two prompt characteristics: linguistic characteristics and task difficulty. Linguistic characteristics referred to textual features of an integrated writing task prompt (e.g., the number of high frequency words). Task difficulty measures included individuals' perceptions of the level of difficulty in completing a writing task in response to a particular integrated writing prompt. One hundred seven previously administered TOEFL iBT integrated writing prompts were analyzed for prompt characteristics. Multilevel modeling analyses were conducted to examine the relationship between prompt characteristics and the average writing scores while controlling for test takers' average English ability that varied across test forms. Results indicated that differences in average English ability and in some of the prompt characteristics contributed to the variation in the average test scores of the integrated writing task across forms. The integrated writing prompts that were perceived to be relatively easy or difficult were reviewed to substantiate the results of the quantitative analysis. The findings of both the quantitative and qualitative analyses will be presented and the implications for test development and validity will be discussed.

Cecilia Guanfang Zhao
Shanghai International Studies University

Thursday, June 23
6:10–6:40
Rackham Amphitheatre (4th Floor)

Assessing authorial voice strength in L2 argumentative writing

Although a construct commonly addressed in writing textbooks and a key component in various writing rubrics, voice remains a concept that is only loosely defined in the literature and mystically assessed in practice. Few attempts have been made, in the field of language assessment in general and writing assessment in particular, to formally investigate whether an authorial voice in written texts can be reliably measured and how the strength of an authorial voice may affect the assessment of the overall quality of writing. Using a mixed method approach, therefore, this study (1) developed and preliminarily validated an analytic rubric that measures voice strength in argumentative writing, and (2) formally investigated the relationship between voice and overall writing quality in the context of a high-stakes L2 writing assessment. Results from the rubric development and validation phase of the study offered an alternative conceptualization of voice that sees voice as being realized through four different dimensions: (1) presence and clarity of ideas and contents, (2) manner of idea presentation, (3) explicit writer and reader presence, and (4) direct interactions between the writer and the readers. Analysis of the relationship between voice and writing quality showed that overall voice strength was a statistically significant predictor of L2 argumentative writing quality. Of the identified voice dimensions, the content-related dimension most strongly predicted writing quality. Results also showed that writer background variables had limited impact on the realization of voice in L2 writing, and that they did not affect the relationship between voice and writing quality in any significant way. Finally, implications of such results for L2 writing assessment and instruction are addressed, together with the directions for future research.

Sawako Matsugu
Northern Arizona University

Anthony Becker
Northern Arizona University

Mansoor Al-Surmi
Northern Arizona University

Friday, June 24
8:30–9:00
Rackham Auditorium (1st Floor)

Balancing practicality and construct representativeness for IEP speaking tests

Since a rather high level of speaking ability is required for success in mainstream university courses, it is important for intensive English programs (IEPs) to make certain that international students have sufficient levels of speaking ability (Luoma, 2004). Doing so necessitates the use of multiple tasks to provide a comprehensive picture of the speaking ability construct. To ensure that we have a better understanding of students' speaking abilities, our IEP tests typically include a combination of four types of speaking tasks (i.e., narrative, independent, paired, and integrated). These tasks factor into determining students' overall speaking ability, as each task is likely to contribute a unique element to the speaking ability construct (Iwashita, McNamara, & Elder, 2001; Powers, 2010). However, one potential drawback of including four different speaking tasks is that the development, administration, and scoring of these tasks might not always be practical. This is particularly true for our IEP, where there has been an increasing number of incoming students over the past five semesters. We believe that it is important to investigate how well these four tasks account for examinees' speaking ability, as the possibility for using fewer tasks could help the IEP in using less human, material, and time resources (see Bachman & Palmer, 1996), while still maintaining high-quality speaking tests that sufficiently measure students' speaking ability. Using multiple regression analysis, this study evaluated how well four types of speaking tasks on IEP proficiency and achievement tests accounted for the speaking ability of our IEP students (N = 130). The findings indicated that independent and integrated tasks uniquely contributed to the speaking ability construct, while there was also considerable overlap between narrative and paired speaking tasks. This study has implications for the importance of balancing practicality issues with construct representativeness for speaking ability.

John Read
University of Auckland

Toshihiko Shiotsu
Kurume University

Friday, June 24
8:30–9:00
Rackham Amphitheatre (4th Floor)

Extending the scope of the yes/no vocabulary test format

Vocabulary tests were a core component in Lado's approach to language testing, and lexical measures are again taking a prominent role in language assessment. The actual procedures for vocabulary testing have not changed that much, with the general tendency being to adapt existing item types rather than to invent new formats. One case of an established format is the Yes/No test, which has been used not only as a research tool for measuring vocabulary size but also for purposes such as estimating learner proficiency levels in DIALANG (Alderson, 2005) and placement in classes at a language school (Harrington & Carey, 2009). In its standard form, a Yes/No test presents written words in isolation, with nonwords included to control for guessing. This paper reports on an investigation of two extensions to the Yes/No format: first, oral presentation of the target words; and secondly, presentation of the words in two types of sentence context. The main participants in the study were 270 tertiary-level EFL learners in Japan, who took two versions of Yes/No test varying in test words (Form A/B), in contrasting input modalities (written/oral) or context conditions (none/syntactic/semantic), while 75 others received two forms (Form A/B) in a common modality and context condition. Of the latter group, 50 also completed a translation task. The paper will focus initially on the translation results as a validation measure and then on the effects of input modality and context on Yes/No item difficulty. Contrary to expectations, comparisons of Rasch-based item difficulty measures indicated that providing context, and particularly semantic context, could make the same Yes/No items more difficult, especially if the input was in oral modality. The paper will discuss the implications of this finding and, more generally, the potential of oral input and contextual presentation to extend the scope of the Yes/No format.

Fenxiang Cheng
Guangdong University of Foreign Studies

Friday, June 24
9:05–9:35
Rackham Auditorium (1st Floor)

Investigating the construct validity of a listening-to-retell test in national matriculation English test in China

Very few studies have been done to investigate the validity of integrated tasks. To fill some of these gaps, this research intends to explore the validity of a newly developed listening-to-retell task in a high-stakes test—National Matriculation English Test (NMET)—in China employing Bachman's (2010) assessment use argument (AUA) approach. To be more specific, the present study addresses the following research questions: (a) What is the relationship between listening-to-retell ability and independent measures of listening and speaking ability? (b) Does the listen-to-retell task reflect the three aspects of the construct, namely, the listening aspect, the speaking aspects and the reformulating aspects? (c) What are the differences among the student of different proficiency levels in retelling performance? The participants were 300 target students from three high schools of different levels. A retelling test, a listening test, a speaking test, and a post-test questionnaire were administered to them. The results showed that both their listening and speaking ability significantly contributed to their retelling scores, even though speaking contributed more. Factor analysis, discourse analysis, and think-aloud data analysis results revealed that the task manifested the three aspects of the construct: the listening aspect, the speaking aspect, and the reformulating aspect. Discourse analysis results revealed that the students from the three different schools differed significantly in complexity, accuracy, and fluency measures. Differences in the reformulation process were also detected.

Wen-Ta Tseng
National Taiwan Normal University

Friday, June 24
9:05–9:35
Rackham Amphitheatre (4th Floor)

Modeling vocabulary knowledge: A mixed model approach

Empirical studies have shown that the relationship between breadth and depth of vocabulary knowledge appears to be positively correlated. However, there is still a debate over the usefulness and validity of the distinction between breadth and depth of vocabulary knowledge. To fill the gap in the literature, a structural equation modeling approach was taken, and three rival models were proposed to answer the research question. Rival model 1 posited that size and depth of lexical knowledge were two entirely distinct constructs and no relationship was formulated between the two contructs. Rival model 2 posited that size and depth of lexical knowledge were two separate but interrelated constructs. Rival model 3 posited that size and depth dimensions were distinct and loaded on a superordinate, higher-order construct, which was lexical knowledge per se. The 2000, 3000, and 5000 levels of Vocabulary Levels Test were used as the indicators of vocabulary size construct, whereas the vocabulary depth construct was measured by a polysemy test, a collocation test, and a written form test. The participants were from Chinese university students majoring in a wide range of disciplines. In total, 301 college students joined the study. The results of pilot study showed that the test instruments were both reliable and valid via Rasch model analysis. The results of confirmatory factor analysis revealed that both rival model 2 and rival model 3 could fit the data equally well, whereas rival model 1 fit the data very poorly. It is argued that the preference should be given to rival model 3 due to a number of theoretical and empirical underpinnings. An important theoretical implication is that the mental processing of size and depth of vocabulary knowledge may function in parallel initially but can eventually be explained by a more general latent trait.

Thomas Dinsmore
University of Cincinnati &
Clermont College

Friday, June 24
9:40–10:10
Rackham Auditorium (1st Floor)

Variables influencing the communicative competence of second/foreign language speakers

This study investigates the influence pronunciation on the oral communicative competence of nonnative speakers of English (NNS). NNSs' performance of oral tasks (conversational questions, explanation of a syllabus, and definition of a term of the NNS's field of study) is rated on a scale with ratings ranging from 1 (low proficiency) to 4 (high proficiency) for the following variables: pronunciation, grammar, vocabulary, listening comprehension, and organization. The NNSs are also rated on the above scale for an overall communicative competence (intelligibility) score. These ratings are the mean of three evaluators' ratings on each of the tasks for each of the variables and the overall communicative competence score. Through the use of multiple regression, each of the variables' influence on the communicative competence score is evaluated. From the results, pronunciation is the variable that has a significant influence on the communicative score rating. Subsequent to this finding, a sampling is drawn from the pool of NNSs and the evaluators. The evaluators are then asked to evaluate the NNSs, focusing on their pronunciation. Follow-up interviews are then conducted with the evaluators, and they are also asked to write short narratives to explain the rationale behind assigning the score for the NNS's score on pronunciation. From these follow-up interviews and narratives, it appears that suprasegmental features are having a greater effect on the determination of the score, rather than segmental features. The results are then discussed in light of their implications for the oral assessment foreign/second language.

Xian Zhang
Penn State University

Friday, June 24
9:40–10:10
Rackham Amphitheatre (4th Floor)

The "I don't know" option in vocabulary size test

The current study examined whether including the "I don't know" choice in the Vocabulary Size Test (VST) by Beglar (2010) will reduce guessing effect that can overstate test takers' vocabulary size. One hundred and fifty first year students in a university of China took part in the study. They were randomly assigned into three different groups. The first group took the original VST. The second group took the VST with the option of "I don't know" and one additional instruction that asked the test takers to choose "I don't know" option for unknown items or else final scores would be deducted for each wrong answer. The third group took the VST with the choice of "I don't know" but without the additional instruction. All the three groups were required not to guess the answer. Immediately after the VST, a follow-up section, which participants did not know, was administered to participants requesting them to write down the meanings for each tested lexical item either in English or Chinese. This section was to examine whether participants were making guesses on each test item. In the second day after the VST, the Vocabulary Level Test by Schmitt et al. (2001) was given to all participants for concurrent validation on the estimated vocabulary size. Rasch analysis was run for data analysis. Results indicated that the "I don't know" choice significantly reduced the guessing effects and the total time for conducting the test. It is therefore believed that the "I don't know" option improves the performance of the VST.

Yuko Butler
University of Pennsylvania

Wei Zeng
University of Pennsylvania

Yeting Liu
University of Pennsylvania

Friday, June 24
10:25–10:55
Rackham Auditorium (1st Floor)

Developmental considerations for employing task-based paired assessment among young learners of foreign language

Despite the popularity of foreign language (FL) education at elementary school, it is still not clear how best to evaluate young learners' foreign language abilities. The assessment needs to be matched with the learners' cognitive, linguistic, and affective developmental levels. The present study aimed to understand developmental differences in interaction among students and how such interaction was evaluated by their teachers as well as the students themselves. By doing so, the study aimed to address possibility and limitation of introducing task-based paired-assessment among young learners of FL. The participants were 32 fourth-grade and 32 sixth-grade students and their teachers in a FL program in China where task-based instruction is promoted to teachers. The students were engaged in two sets of tasks (an information-gap task and an open-ended problem-solving task) with their peers. After each task, the teacher evaluated each student's performance and the students self-assessed their performance. It was found that widely used Storch's (2002) model of dyadic interaction captured the sixth graders' interaction well but not the fourth graders'. One of the binary features of the model, equality, was not useful primarily due to the students' similarly limited English proficiency and background knowledge to complete the task. The other feature, mutuality, also often had limited application because of the following "pre-interactive" characteristics of the 4th graders' interaction: (1) mechanical turn-taking; (2) difficulty with taking the partners' perspective (e.g., failed to use reference points when giving information to the partner); (3) minimal mutual topic development; and (4) limited use of reasoning and back-channel behavior. The potential for eliciting a wide range of interactive skills and dynamics of interaction, which is the primary advantage of introducing paired-assessment, was not present among the fourth-grade dyads. Evaluation of their interactional skills, accordingly, was difficult for both the teachers and the students.

Dayong Huang

Friday, June 24
10:25–10:55
Rackham Amphitheatre (4th Floor)

Building a theoretical foundation for test impact research

Test impact is an important aspect of testing. It is taken as the consequential aspect of test validity and the first consideration in test use and test development. Though a large amount of research has been done to investigate the impact of language testing, especially the washback effects, there is still a lack of systematic theories to guide future research. Based on the analysis of existing studies on the topic, the paper summarizes four types of theoretical explorations in these studies: the work to conceptualize the notion, the attempts to include test impact into validation framework, the efforts to explore the working mechanisms of test impact and the application of different theories to interpret test impact. Based on these explorations, three new models, including a concept model, a mechanism model, and an interpretation model, are proposed as the theoretical frameworks for further empirical research on the impact of language testing.

Marita Härmälä
University of Jyväskylä

Outi Toropainen
University of Jyväskylä

Friday, June 24
11:00–11:30
Rackham Auditorium (1st Floor)

Assessing interaction skills: Evidence from Finnish and Swedish L2 national exams on grade 9

How do participants in a peer-peer interaction employ their linguistic and interactional resources and construct identities for themselves and others (Young 2008)? In what way do discursive practices vary across topics, task types, and proficiency levels? In interaction, meaning is co-constructed (Kramsch 1986; McNamara 1997) and this, in turn, poses problems in a testing situation. For the purposes of the study a sample consisting of 12 oral pairs in Swedish and 17 oral pairs in Finnish was chosen on the basis of pupils' test score, sex, and task set. The performances were analyzed qualitatively to explore what kinds of microlevel functions and interactive schemata the pupils employed in order to complete the tasks, and how did the present (and absent) participants affect the nature of interaction. The original data consists of the most recent national assessments of learning outcomes done by the Finnish National Board of Education in 2008 and 2009. In these assessments, 795 sample pupils did the speaking test in Swedish and 360 pupils in Finnish. The audio-recorded oral performances were double rated by language teachers using a 10-level proficiency level scale (A1–B2) based on the Common European Framework of Reference. The results showed that discursive practices varied between proficiency levels, e.g., on A1-level simple question-answer patterns and monologues prevailed. In addition to the two test takers, the role of the teacher and the absent item writer influenced the pupils' performance. The study has implications for teaching interactive skills in the classroom as well as for creating empirically based assessment criteria for spoken interaction especially on the lower levels of language proficiency.

Hsinmin Liu
University of California,
Los Angeles

Friday, June 24
11:00–11:30
Rackham Amphitheatre (4th Floor)

Building a construct validity argument for the GEPT reading test: A confirmatory factor analysis approach

One issue that has been raised about the General English Proficiency Test (GEPT) is the lack of theory specification for the construct to be measured (Roever and Pan, 2008). In addition, Weir (2005) has argued that "can-do" statements that are provided for score interpretations are not sufficient for what is being measured because behavior does not equate to the ability engaged in a target context. In order to establish the inferential link between the test construct and the target language use situation, a theory-based construct validity argument for the test must first be articulated and backed by evidence. This study built and supported such an argument for the GEPT high-intermediate reading test based on Bachman and Palmer's (2010) "Assessment Use Argument" framework. It examined randomly sampled data from the target test population using confirmatory factor analysis (CFA) with item-level responses. A tetrachoric correlation matrix of the dichotomous item response data was analyzed. In order to collect all relevant evidence to support the warrants that score interpretations are meaningful with respect to a general theory of reading ability, two complementary models, the higher-order factor model and the bi-factor model were tested. Both models supported the theory-based construct validity of the test, thus weakening the rebuttal that the test lacks "theory specification" for the construct to be measured. However, it was also found that the test does not reflect enough depth of that reading construct at the item level. In other words, a good CFA model fit does not guarantee the test to be a sufficient measure of the construct; low item-level variances need to be taken in consideration and interpreted with great caution.

Pamela Gunning
McGill University &
Concordia University

Friday, June 24
11:35–12:05
Rackham Auditorium (1st Floor)

Classroom-based assessment of the effects of instruction and children's strategy use on success in ESL

To date, many studies have examined learning strategy use (Oxford, 2001) but few have assessed the effects of strategy instruction and children's strategy use on success in ESL, in classroom-based research. The assessment of children's strategies offers particular challenges to researchers, as traditional methods are not always appropriate (Gu, Hu, & Zhang, 2005a). In Québec, Canada, an education reform instigated a curriculum which advocates a seamless link among teaching, learning, and assessment (Policy on the Evaluation of Learning, 2003), and makes the teaching and assessment of strategies in the performance of tasks integral to the ESL curriculum. This paper reports a quasi-experimental strategy intervention study which assessed the impact of strategy use on success in oral interaction tasks, among Québécois sixth graders, N = 56. Two intact groups of participants served as an experimental group and a control group. Innovative techniques were devised for assessing children's strategies as they executed ESL tasks. Qualitative and quantitative data, including questionnaires and videotapes of classroom proceedings, provided seven sources of evidence to support the findings of this mixed-methods investigation. Findings of the statistical analyses indicate that (a) the strategy intervention group showed statistically significant gains in oral interaction from pre- to post-test; and (b) the strategy intervention group outperformed the control group in a planned comparison of post-test oral interaction results. This study is of interest and importance to the field of language assessment as it provides insight into strategy assessment among children, and contributes to the bank of age-specific research instruments. This paper is related to the LTRC conference theme because classroom-based assessment of children's use of strategies to facilitate language learning and task performance represents an innovative facet in the evolution of half a century of assessment theory.

Shu Jing Yen
Center for Applied Linguistics

Ying Zhang
Center for Applied Linguistics

David MacGregor
Center for Applied Linguistics

Friday, June 24
11:35–12:05
Rackham Amphitheatre (4th Floor)

Item features and construct of academic English

The purpose of this study is to investigate how item features of the WIDA ACCESS Reading assessment items interact with the construct of academic English. This work seeks to extend previous work by Romhild, Kenyon, and MacGregor (2008, 2010) that provided empirical support for the presence of factors related to the academic language of content areas beyond the general academic English language proficiency factor for higher levels of academic English proficiency. It would be interesting to examine whether the factor structure identify by Romhild and her colleagues can be explained by the item features associated with the items. In particular, the degree to which levels of vocabulary knowledge and linguistic complexity overlap more with the language of the content areas at higher proficiency level tests than lower proficiency level tests can explain why factors related to the academic language of content areas are more salient in higher proficiency level tests. That is, vocabulary demands and linguistically complexity may serve as some kind of mediator variables in explaining the relationship between the English language factor and factors related to the academic language of content areas. This study proposes to introduce endogenous variables which characterize item features of WIDA ACCESS Reading assessment and incorporate these variables in the structural equation model framework where these variables serve as mediators in explaining the latent relationship between the English language factor and factors related to the language of academic content areas. Item features introduced in the model include but are not limited to measures of word frequencies using tools such as WORDSIFT, measures of genre, and measures of linguistic syntactic complexity. This study draws data from the 2010 reading test using a sample of 50,000 high school students (grade 9–12) and three test forms measuring language at low, mid, and high proficiency levels.

Alistair Van Moere
Knowledge Technologies, Pearson

Ryan Downey
Knowledge Technologies, Pearson

Masanori Suzuki
Knowledge Technologies, Pearson

Mallory Klungtvedt
Knowledge Technologies, Pearson

Friday, June 24
1:30–2:00
Rackham Amphitheatre (4th Floor)

Assessing written English skills for business communication using time-constrained tasks

A survey of widely used tests of written communication skills for workplace contexts reveals that they share similar design features: 2–3 writing prompts, 45–60 minutes in length, a 2–3 week wait to receive scores, and a single overall score. But a needs analysis conducted with employers and HR managers (10 multinational corporations in five countries and online survey responses from 157 companies) revealed that users wanted tests with: shorter test time, immediate score turnaround, workplace-relevant tasks, and multiple subscores. Therefore, a new computer-based test of written English workplace skills was developed, consisting of five tasks. The presentation reports on the rationales for the selection and development of the tasks based on the results of the needs analysis. Data provided by Rasch, G-theory, rating scale analysis, and language content analysis were applied during test construction to inform test design. Special attention in this paper is given to two item types in the test: Passage Reconstruction and Email Writing. Responses elicited from these two tasks are scored by machine using LSA (latent semantic analysis) techniques and evaluated on traits such as organization, voice, grammar, and vocabulary. In Passage Reconstruction, candidates are given 30 seconds to read a short passage, after which the passage disappears, and the candidates then reconstruct the passage in 90 seconds. In Email Writing, candidates have 9 minutes to compose an email responding to a certain situation described in the prompt. Data is provided from over 2,300 tests showing that the tasks are highly reliable at discriminating among proficiency levels. A design characteristic of the test is to elicit frequent yet comparatively short written performances under strict time constraints in order to reflect the time pressures of office work and productivity required by employers. It is argued that time-constrained tasks are an authentic yet underutilized tool in the assessment of writing.

Kathryn Nelson
SWA Consulting Inc.

Eric Surface
SWA Consulting Inc.

Amy DuVernet
North Carolina State University

Milton Cahoon
SWA Consulting Inc.

Friday, June 24
2:05–2:35
Rackham Auditorium (1st Floor)

Language aptitude, first language ability, and second language proficiency: A meta-analytic investigation

Organizations have a growing need for global communication capabilities (Warschauer, 2000), resulting in the need for more individuals with second language (L2) proficiency and higher levels of proficiency. Organizational training and university foreign language education programs need to identify individuals who are most likely to succeed or struggle in learning a L2, in order to leverage instruction resources and optimize learning. Identifying individuals who may struggle in learning a language is necessary in order to provide targeted learning interventions that will increase chances of training success. Many factors have been studied to predict success in foreign language learning. Our study focuses on two of the most common—language aptitude and first language (L1) ability. Language aptitude measures are commonly used to predict L2 learning success (e.g., Gardner & Smythe, 1976). L1 knowledge and ability measures are also used for this purpose (e.g., Mueller & Wiersma, 1963). The goal of this study is to examine the how well these two types of measures predict overall foreign language proficiency outcomes across four proficiency skill modalities (i.e., reading, writing, speaking, and listening) using a meta-analytic strategy. Additionally, we investigate the influence of different sample, setting, and training characteristics using moderator analysis. Meta-analysis combines the results of previous research to statistically determine the average relationship across all included studies. Unlike primary research studies, meta-analyses provide an understanding of the relationships between variables that is less susceptible to sampling error (Hunter & Schmidt, 2004). The current study includes 181 studies examining the relationship between L1 measures and L2 proficiency outcomes and 236 studies examining the relationship between language aptitude measures and L2 proficiency outcomes. Moderating factors—including previous foreign language experience, language difficulty, training setting, and trainee age—are included. Findings and implications for future language training and research will be discussed.

Gene Halleck
Oklahoma State University

Carol Moder
Oklahoma State University

Friday, June 24
2:05–2:35
Rackham Amphitheatre (4th Floor)

The importance of discourse-based research in LSP test validation

When designing LSP tests, test designers must consider what aspects of language proficiency stakeholders need in order to function successfully. Tests must reflect actual workplace language use and be representative of the tasks commonly undertaken. This paper outlines the steps that we undertook (beginning with attempting to specify the relevant domain characteristics so that categories of performance could be sampled appropriately) in our test design projects. We focus on two specific domains: Aviation English and International Teaching Assistants, and discuss why it is important to conduct a thorough needs analysis in order to determine the relative importance of various events and functions (McNamara 1996; Douglas, 2000). We describe how we undertook such a needs analysis, beginning with a complete "Domain Analysis," and "Domain Modeling" (Mislevy, Steinberg, & Almond, 2003) collecting samples of authentic discourse. This was done from actual recordings of pilot/ATC communication (for Domain 1) and from classroom observation of professors and graduate student instructors in different academic departments (for Domain 2) to determine what roles and tasks should be included in the test items that we were developing. In addition we describe how stakeholders' understanding of how the test constructs relate to the domain was sometimes problematic and how we had to deal with stakeholders who did not understand testing and how testing samples the domain. In addition we suggest that in Domain 1 we were initially outsiders, but in Domain 2 we were part of the domain, and how that impacted our research. We also discuss the difficulties encountered when we got advice from subject matter experts regarding the representativeness of certain tasks and our ability to evaluate performance on them. We conclude by suggesting how the discourse analysis revealed important distinctions about the characteristics of these otherwise well-defined domains.

Heejeong Jeong
University of Illinois at Urbana Champaign

Friday, June 24
2:40–3:10
Rackham Auditorium (1st Floor)

Past, present, and future of language assessment courses: Are we headed into the right direction?

Even before the field of language testing emerged as an independent field in applied linguistics, language assessment courses had been taught in various forms and by different instructors. Currently, these courses are being taught by professionals who have majored in the area of language testing (LTs) but also by others who come from different majors (non-LTs). This study seeks to investigate the effect instructors bring in shaping the characteristics (i.e., content, structure) of language assessment courses and to what extent the student-teachers (ST) are satisfied with the course. To get the full picture, the characteristics and satisfaction of the courses are researched from four different lenses; instructors who teach the course (language testers vs. non-language testers) and different grade levels of language STs (adult vs. K–12) who have taken the course. A large-scale online survey, in-depth follow-up phone interviews, and syllabi document review have been done for the study. A total of 380 instructors and STs completed the online survey (instructors n = 140, student-teachers n = 240). Based on the survey results an in-depth phone interview has been conducted with 13 instructors from 5 different countries and 5 student-teachers. Survey findings show there are significant differences in the content of the course depending on the instructors' background in six areas; test specifications, test theory, basic statistics, classroom assessment, rubric development, and test accommodation. Interview results confirm non-LTs are less confident in teaching technical assessment skills compared to LTs and have a tendency to focus on more classroom assessment issues. Student-teachers are overall satisfied with the course but wanted more activities that were directly related to the course. The study ends by predicting the future of language assessment courses and why it is important for both LTs and non-LTs to communicate and work actively to develop a better course that fulfills the needs of the student-teachers.

Ivana Vidakovic
University of Cambridge
ESOL Examinations

Hanan Khalifa
University of Cambridge
ESOL Examinations

Friday, June 24
2:40–3:10
Rackham Amphitheatre (4th Floor)

Cognitive validity of an ESP test: Cognitive processing during reading comprehension

In the teaching and assessment of reading comprehension in the 1980s, the focus was on text and test scores (Berhnardt 1986). Since the 1990s, there has been a growing focus on mental processes in reading comprehension (Anderson 1991, Cohen & Upton 2006). However, studies mostly drew on skill/strategy inventories rather than a full-blown model of reading comprehension. In the ESP field, much research is still devoted to the influence of subject-matter knowledge on scores (Krekeler 2006). Clearly, the score as the endpoint of the reading comprehension process cannot reveal the different levels of cognitive processes test takers go through to reach a particular level of proficiency in reading. This line of enquiry is vital for investigating cognitive validity (Weir 2005) of a test by determining whether a test elicits the intended mental processes and behavior, and consequently, whether it can be deemed a valid predictor of ability to perform real-life reading tasks. Using a reading comprehension model adopted from Khalifa and Weir (2009) and reworking it into an introspection checklist, this paper investigates the cognitive processes and reading types activated by ten test takers while taking an ESP reading test: the International Legal English Certificate (ILEC). We also report on a follow-up focus-group interview with the test takers. The findings based on quantitative and qualitative data will be presented to discuss the agreement between actual cognitive processes and reading types employed by test takers and the ones anticipated by test developers. Task features which elicited a more authentic interaction with texts and encouraged test takers to draw on legal knowledge will also be discussed. We will situate the findings within the broader ESP debate and argue for the need for a systematic and comprehensive approach to gathering evidence to support validity of LSP tests, in particular, and language tests in general.

May Tan
McGill University

Carolyn Turner
McGill University

Friday, June 24
3:25–3:55
Rackham Auditorium (1st Floor)

Aligning internal and external validation of a high-stakes ESL exam

In recent years, there has been increasing interest in aligning classroom-based assessment (CBA) and high-stakes testing (assessment external to the classroom) to maximize student learning (Pellegrino et al., 2001). In language testing, there have been calls for more research and discussion on links between pedagogy and the traditional measurement paradigm (Davison & Leung 2009). In the latter, test developers, concerned with validity, distance themselves from test takers to reduce construct-irrelevant variance. In the CBA paradigm, teachers know their students and tailor assessment for their needs. In the Quebec context, these two paradigms converge: the high-stakes Secondary ESL exit writing exam developed by the Education Ministry (MELS) is administered and corrected by classroom teachers. Teacher feedback is regularly solicited during test development. Teachers also receive training concerning task requirements and evaluation rubrics. Although student proficiency varies, one consistent factor is that all teachers assess their own students. How does this additional task characteristic impact the results? How should these results be interpreted? To answer these questions, this study uses a mixed methods approach. Participants include 11 classroom teachers, 11 MELS-trained raters, 2 MELS officials and 593 students. Data consists of student scores rated first by their classroom teachers and again by three raters (who are also teachers); interviews with teachers, raters, and officials; and rater survey information. Rasch analysis was conducted on rater behavior and content analysis on the interviews and survey. Results to date indicate that, in terms of task validation, this unique procedure increased validity and alignment. Teacher involvement throughout the testing process improved task validity, enabling more valid teacher judgments concerning students' writing. It helped align classroom teaching, learning, and assessment to external testing. These results contribute to the evolving concept of validity (Lado, 1961; Messick, 1996), and the ongoing dialogue concerning CBA and high-stakes testing.

Nathan Carr
California State University, Fullerton

Friday, June 24
3:25–3:55
Rackham Amphitheatre (4th Floor)

The generalizability of scoring keys for the computer automated scoring of web-based language tests

This study involves the automated scoring of limited-production reading comprehension questions. Clauser, Kane, and Swanson (2002) point out that previous studies of computer-automated scoring (CAS) have focused on the comparability of human- and computer-produced scores. Two additional studies (Clauser, Harik, & Clyman, 2000; Clauser, Swanson, & Clyman, 2000) further note important concerns over how representative the raters used in such comparability studies might be, and how representative the scoring criteria or algorithms used for CAS might be. To date, however, no studies appear to have been performed looking at this question in the context of language assessment, or in the area of limited-production tasks. This is surprising, as the generalizability of scoring keys is a crucial prerequisite for a test's construct validity. This study uses scoring keys written by seven pre- and in-service ESOL teachers and 253 student responses to an academic reading and listening comprehension test. The key authors wrote their keys independently, and subsequently worked together in small teams to arrive at consensus versions of their scoring keys. The resulting keys were then used to address the following research questions regarding the comparability of scoring keys written by different authors: (1) How consistent are the scores that result when a test is scored using CAS keys written by different authors? (2) How consistent are the scores that result when a test is scored using CAS keys written by different teams of authors? (3) When training teachers to write scoring keys, what issues need to be addressed in order to obtain more generalizable results? The study addresses these issues by comparing the dependability of scores resulting from each key, examining the proportion of test score variance accounted for by the author (or team) facet, and comparing test scores to see whether the different keys result in significant differences.

Christian Colby-Kelly
McGill University

Friday, June 24
4:00–4:30
Rackham Auditorium (1st Floor)

AFL: Focus on formative assessment in an L2 classroom

Assessment for Learning (AFL), a finely tuned formative assessment approach used in general education, asks teachers to be inventive and learners to take greater responsibility for their learning. AFL has seldom been applied or researched in L2 classroom settings. Having a metacognitive focus, AFL fits well with calls in the language testing literature for research into alternative assessment (Brookhart, 2005; Fox 2009; McNamara 2001; Poehner and Lantolf 2005; Turner, 2009). This paper describes an innovative application of L2–AFL in pre-university ESL classes. Using a grammatical element (would and will usage) to investigate L2–AFL, the paper reports on a quasi-experimental, mixed-methods study incorporating "treatment" and "control" groups. For the study, L2–AFL teacher training was developed, as were L2–AFL pedagogical materials—computer-assisted language learning (CALL) exercises, an online concept mapping exercise, and group and teacher-class concept mapping exercises. Their effect on student learning was investigated, and evidence was sought of the assessment bridge (the area where classroom assessment, teaching, and learning intertwine). The data collection instruments included the concept maps produced, observation field notes, transcribed classroom discourse, and teacher and student questionnaires, which were analyzed through mixed methods (qualitatively through content analysis and quantitatively through frequency counts). Pre- and post-treatment tests were also administered to indicate trends. The triangulated results show strong teacher and student support for L2–AFL, suggest that learner metacognition was increased, and also that the L2–AFL may have enhanced these students' learning. Evidence of the assessment bridge was also found. Based on the results, a call is made for a research agenda for further study of L2–AFL in the LT community. The present study's L2–AFL application attempted to help learners notice a grammatical interlanguage element in contrast with target usage through individual and interactive formative assessments, reflecting the progression from Lado's theory and practice to those of today.

Yujie Jia
University of California, Los Angeles

Alan Urmston
The Hong Kong Polytechnic University

Felicia Fang
The Hong Kong Polytechnic University

Friday, June 24
4:00–4:30
Rackham Amphitheatre (4th Floor)

Investigating the dependability of analytic scoring for a computer-based oral test

This study investigated the dependability of analytic scoring for a computer-based oral test with five integrated tasks. The test measures five dimensions of test takers' performance: task fulfillment and relevance (TFR), clarity of presentation (CoP), grammar and vocabulary (GV), confidence and fluency (CoFlu), and pronunciation (Pron). Univariate and multivariate generalizability and decision studies were conducted to analyze the analytic score profiles of 300 candidates on the five tasks. The complex rating scheme in this study featured an r: (p x t) design for each dimension, with persons (p) crossed with tasks (t) and raters (r) partially nested within persons and tasks. Subdividing method was employed to analyze the dataset since it could yield more accurate estimates of score dependability (Chiu & Wolf, 2002; Xi, 2006). The results revealed high score variance on TFR both between raters and across tasks. On the other four dimensions, scores varied between raters but were relatively stable across tasks. In addition, disattenuated correlations among the analytic scores by task were very high but those between TFR and the other four dimensions were generally lower, consistent with Chalhoub-Deville's (2003) claim that while communicative competence is to some extent stable, some components may be local and dependent on the contexts in which the interactions occur. The study suggested that there would be little to gain from analytic score reporting for this computer-based oral test, as opposed to the composite scores currently given. It is hoped that the findings can offer useful information for the test developers to consider when making decisions about the scoring of the test. The findings also have theoretical and practical implications for assessing the quality of large-scale performance tests, rater monitoring, and rater training. The study is expected to contribute to the use of generalizability theory for analyzing data with complex rating schemes.

Yan Jin
Shanghai Jiao Tong University

Jiang Wu
Shanghai Jiao Tong University

Ming Yan
Heilongjiang University

Friday, June 24
4:35–5:05
Rackham Amphitheatre (4th Floor)

Is computer literacy construct-relevant in a language test in the 21st century?

Since the introduction of computers into language testing in the 1990s, the use of technology has been a source of concern for its potential threat to test validity resulting from construct-irrelevant difficulties (Messick 1996). Studies in this area have focused on establishing the equivalence between paper and computer-based tests, and have often evidenced that the assessment of language ability was confounded by the test taker attribute of computer literacy. In this study, analyses of 3,984 test takers' scores in the paper and Internet-based College English Test (IB-CET), together with data on test-taker computer literacy and their evaluation of the tests, identified a statistically significant relationship between test takers' computer familiarity and their perceptions of and performances on the tests. Analyses of the effect size (Cohen 1988) confirmed practical significances of the effects (d = 0.33~0.94). However, analyses of 100 test takers' writing scripts on the two versions showed that significantly lengthier and more complex sentences were produced in the IB-CET, irrespective of their level of computer literacy. Further analyses of the cognitive processes involved in writing indicated a better use of planning and revision strategies in the IB-CET. The IB-CET, which engages test takers in language activities performed on the computer, is not its paper version equivalent because they do not measure the same thing. It is argued that the construct of a language test in the twenty-first century needs to be reconceptualized by incorporating computer literacy into its definition. In an era of extensive application of computers and the Internet, target language use situations have changed in fundamental ways. Language tests with a narrow construct definition are neither feasible nor desirable. A research agenda should be set out for a clearer definition of computer literacy for language use and better ways of engaging test takers' computer literacy to facilitate test performance.

Yi-Ching Pan
National Pingtung Institute of Commerce

Friday, June 24
5:15–5:45
Rackham Amphitheatre (4th Floor)

Do tests promote changes in receptive skills?: A comparative study in a Taiwanese EFL context

According to Ross (2005, p. 462) most test washback studies focus on stakeholder opinions of various test consequences without evaluating actual test score changes. In response to that criticism this study compares how the listening and reading test scores of two groups of students changed over a nine-month period. A control group of 140 Taiwanese university students at a school without any English proficiency certificate graduation requirement was compared with an experimental group of 136 similar students from a school requiring students to pass an English certification test in order to graduate. An independent sample t-test showed no statistically significant difference in scoregains between these two groups. Moreover, 89% of the respondents also completed pre-test and post-test questionnaires to ascertain their motivation, methods of study, and other factors that may have influenced their learning outcomes. The questionnaire findings suggest that students at the school with exit requirements did concentrate on listening skills more than their peers. However, the lack of significant scoregains between the two groups suggests that a test-driven policy such as the exit certificate requirement discussed herein cannot by itself significantly improve students' English skills.

Rika Tsushima
McGill University

Friday, June 24
5:50–6:20
Rackham Amphitheatre (4th Floor)

Policy and practice in EFL contexts: Investigating assessment in oral communication courses in Japan

Research has reported that the occurrence of washback effects—the impact of high-stakes exams on classroom practice and activity—is often partly attributed to teachers (Cheng et al., 2004; Turner, 2009). In Japanese secondary schools, due to the strong influence of university entrance exams in the society, it is often argued that in practice, speaking-focused courses titled Oral Communication (OC) courses do not focus on interactive language learning activities but on grammar exercises to prepare students for the high-stakes exams (Kikuchi & Browne, 2009). This paper, using mixed methods, examines the current status of OC courses in relation to the national educational policy from the perspective of Japanese teachers of English (JTE). As findings, quantitative data from a teacher survey (N = 87) revealed that washback effects were more evident in the assessment of the courses than in teaching, suggesting that classroom teaching and assessment were not congruent with each other or with the course objectives. Thematic analyses of guided interviews with nine JTE provided insight into their opinions on their grammar-oriented teaching practice. Along with washback, a lack of confidence in assessing students' speaking stemmed from JTE's anxiety as nonnative English speakers; it was an influential factor that hindered JTE from implementing the course objectives. Moreover, the results suggest that the high-stakes exams tend to be more influential than the educational policy in this context, especially in academically inclined schools. Finally, the paper discusses the importance of the inclusion of a speaking component in high-stakes exams as well as the necessity of speaking assessment guidelines specifically designed for nonnative language teachers that will contribute to the improvement of EFL pedagogy.

James Dean Brown
University of Hawai'i at Mānoa

Keita Nakamura
Society for Testing English Proficiency

Saturday, June 25
8:15–8:45
Rackham Auditorium (1st Floor)

EIKEN program testing context analysis

This testing context analysis was conducted on the Eiken EFL testing program in 2008. Attending to the concepts of stakeholder-friendly and defensible testing described in Brown (2008), we were able to investigate the relevance/utility, values implications, and social consequences described in Messick's (1989) framework for validity. This paper describes how mixed qualitative and quantitative research methods were used to evaluate the attitudes of the following four stakeholder groups toward various aspects of the Eiken testing program: (a) test takers, (b) test takers' parents, (c) teachers, and (d) staff members of Eiken program. Initially, qualitative data were gathered using interviews and observations. Then based on these qualitative data, questionnaires were developed to gather primarily quantitative data on functional, political, and economic aspects of testing. The questionnaires were administered at various sites in Japan to all four groups of stakeholders; a total of 1,518 test takers, 521 test takers' parents, 211 teachers, and 33 Eiken staff members responded. To investigate common variance structures on the questionnaires within and across different stakeholder groups, the data for the first three groups were analyzed using separate principle components analyses. Four distinct components were found. To investigate differences in attitudes as measured on the questionnaire among the four groups, ANOVA procedures were used. The analyses focus on comparing the views of the four stakeholder groups, and the discussion provides concrete examples (based on specific questionnaire items) of similarities and differences in their views. The implications are discussed in terms of constructing broader validity arguments for Eiken test score use and decision making based on test score relevance/utility, values implications, and social consequences. In the process, suggestions are made for improving the Eiken tests and testing program.

Gary Ockey
Kanda University of International Studies

Dennis Koyama
Kanda University of International Studies

Angela Sun
Kanda University of International Studies

Saturday, June 25
8:15–8:45
Rackham Amphitheatre (4th Floor)

Effects of presenting question stems, answer options or neither on multiple-choice listening comprehension tests

Multiple-choice (MC) formats have been a popular technique for assessing listening comprehension for decades, but little research is available to specifically guide how MC formats should be put into practice. Some researchers have argued that to have a purpose for listening, test takers must be provided with the questions prior to listening to the assessed input (Buck, 1995; Sherman, 1997) while others argue that allowing test takers to preview the questions and answer options may decrease the authenticity of the task by changing the way test takers process the input (Hughes, 2003). For instance, when test takers are allowed to preview questions and answer options, research suggests a "lexical matching strategy" in which test takers, particularly less proficient ones, listen for key words they see in the answer options rather than for general comprehension (Yanagawa & Green, 2008). Stratified random sampling techniques were employed in which an equal number of more and less proficient Japanese university-level English learners (N = 210) was assigned to take one of three test formats. The formats included: preview of question stems and answer options, preview of question stems only, and no preview. The test included item types that were distinguished by whether or not they were designed to include key words from the input in one or more of the answer options. The researchers will report relative item difficulty and reliability of the three formats as well as their effects for the different item types and ability levels of the test takers. The effects of test format will be discussed in relation to a listening construct.

Yeonsuk Cho
Educational Testing Service

Brent Bridgeman
Educational Testing Service

Saturday, June 25
8:50–9:20
Rackham Auditorium (1st Floor)

Relationship of TOEFL scores to success in American universities

The study investigated the validity of TOEFL iBT test scores for predicting nonnative English speaking students' academic performance at American universities. Academic records of 2,594 undergraduate and graduate students were collected from ten universities in the U.S. with a high enrollment of international students. The data from each school consisted of TOEFL iBT scores, other admission-related test scores (e.g., GRE and GMAT), demographic information, and GPAs in specific departments for graduate students and in specific course clusters (e.g., arts and humanities, business, science and engineering) for undergraduate students. Correlation and regression analyses were performed to examine the extent to which TOEFL iBT scores could predict academic success defined as GPA. The relationship between TOEFL iBT scores and academic success was further explored descriptively using expectancy graphs. Specifically, within each graduate department, or undergraduate course cluster, the top and bottom quartiles in terms of TOEFL iBT scores were compared with the top and bottom quartiles in terms of GPA. Analyses were conducted for subgroups by academic status and by academic disciplines. Results indicated modest correlations between TOEFL iBT scores and academic performance, and small increments in the multiple correlations when other admissions test scores were entered prior to the TOEFL iBT scores, possibly due to the restricted range in test scores and the lack of variation in GPA. Nevertheless, strong evidence for the predictive validity of TOEFL iBT was observed in the expectancy graphs. In the top quartile of students in terms of TOEFL iBT scores, there were twice as many students in the top grade quartile as in the bottom quartile, while in the bottom quartile of TOEFL iBT scores there were twice as many students in the bottom grade quartile as in the top quartile. Results also indicated some variation across subgroups and will be presented in detail.

Ying Zheng
Pearson

John H.A.L. De Jong
Pearson / VU University Amsterdam

Saturday, June 25
8:50–9:20
Rackham Amphitheatre (4th Floor)

The validity of using integrated tasks to assess listening skills

This research provided an evaluation and validation of the listening section of Pearson Test of English Academic, which contains 11 item types assessing academic listening skills either alone or in combination with other skills (e.g., reading, writing). This presentation reports on results from a task analysis of how the 11 integrated item types performed as a measure of academic English listening ability. The findings were further explored statistically to examine the effectiveness of item types, the underlying listening constructs, and the relationship between listening performance and demographic backgrounds. First, task analysis helped identify skills important for listening comprehension in academic settings. More specifically, aspects analyzed included the purpose of assessment tasks, skills/constructs assessed, tasks, task stimuli, and speech events employed in PTE Academic. These aspects were compared to those specified in Powers (1986) and Buck (2001), which provide comprehensive overviews of academic English listening skills. The findings indicated that modern technologies enabled PTE Academic, a computer-based test, to facilitate the assessment of students' academic listening abilities in real time using the integration of multi-modal sources. The statistical validation consisted of three stages. First, item scores on different listening skills were subjected to Rasch analysis using CONQUEST. Difficulties of the item types were estimated and the effectiveness of these item types was evaluated by calculating the information function of each item type. Second, exploratory factor analysis was performed with a sample of over 1,000 students who took PTE Academic, to examine the underlying listening constructs as measured by the 11 item types. Further, differential score patterns were compared using ANOVA to ascertain if any of the score differences were associated with students' demographic background (i.e., gender, ethnicity, first language). The study has implications for test developers and test users regarding the interpretation of student performance on listening assessments.

Kate Kokhan
University of Illinois
at Urbana-Champaign

Saturday, June 25
9:20–9:55
Rackham Auditorium (1st Floor)

Time lag as a mediator in determining language proficiency: Graduate vs. undergraduate levels

According to ETS, TOEFL iBT measures the ability of international students to use and understand English in the academic environment. The University of Illinois at Urbana-Champaign (UIUC) had established a campuswide requirement that all new international students with the TOEFL iBT score of 102 and below and with the TOEFL PBT score of 610 and below have to take the EPT (the English Placement Test) before their first semester at the university. However, currently there are no established university guidelines on interpreting older test scores. In this work, I explored the changes in the predictive power of TOEFL iBT scores on ESL placement over the time between both tests. I collected self-reported and official TOEFL iBT scores for all EPT test takers at UIUC over the last five years (N = 1560) and calculated correlations for these students grouped by the time between taking TOEFL iBT and the EPT. I found no significant change in correlation over the period of time between TOEFL iBT and the speaking part of the EPT. However, there is a distinct correlation pattern between the time of taking TOEFL iBT and the written EPT: the correlation goes down and reaches zero at about the 50th week from the date of taking TOEFL but, surprisingly, it goes up after week 50. This pattern is especially pronounced for undergraduate students. The results of this research may be used to establish new campus guidelines for dealing with older test scores in ESL placement and admission decisions. Possible explanations of this phenomenon are discussed.

Tineke Brunfaut
Lancaster University

Andrea Revesz
Lancaster University

Saturday, June 25
9:25–9:55
Rackham Amphitheatre (4th Floor)

EAP listening task difficulty: the impact of task variables, working memory and listening anxiety.

This paper reports on a study investigating the effects of a group of task factors on EAP listening test performance related to various EAP proficiency levels, and whether their effects are mediated by test takers' working memory capacity and listening anxiety. Although a number of studies exist that have examined the impact of text and response characteristics on listening test difficulty, the novelty of our research lies in that we have explored these links in relation to individual test-taker characteristics. In particular, we investigated how (1) the linguistic complexity, speech rate, and explicitness of the texts, and (2) the linguistic complexity and comprehension of response options affect item difficulty for test takers at different EAP proficiency levels. Additionally, the study explored how phonological short-term memory, verbal working memory capacity, and foreign language listening anxiety may modulate these relationships. The participants were 100 EAP students at a UK university at proficiency levels ranging from A2–B2. They completed thirty versions of a task, which involved listening to a short passage and selecting an appropriate ending for the passage (multiple-choice). The passages and answer options were analysed by means of Praat v5.0.25, Cohmetrix v2.0, and/or Web VocabProfiler v3. Test takers' comprehension of response options was assessed through a translation task. We used a digit span test to assess test takers' phonological short-term memory, and a backward digit span test to gauge their verbal working memory capacity (Gathercole, 1999). The Foreign Language Listening Anxiety Scale (Elkhafaifi, 2005) was administered to obtain a measure of test takers' listening anxiety. We estimated the difficulty of the various task versions by the means of Rasch analysis, and regression analyses were used to examine the links between the task and individual difference factors. After presenting and discussing the results of the study, the implications for EAP listening test task design will be considered.

Ikkyu Choi
University of California, Los Angeles

Jonathan Schmidgall
University of California, Los Angeles

Saturday, June 25
10:10–10:40
Rackham Auditorium (1st Floor)

A survey of methodological approaches employed to validate language assessments: 1999–2009

The process of validation has been one of the primary concerns within educational measurement and within the smaller field of language assessment. Examining research in this field over a period of time within a framework for validation may help identify well-supported methodological approaches as well as areas in which further attention is needed (Kunnan, 1998). Using Messick's framework of validity to categorize sixteen years of research (1980–1996), Kunnan (1998) reviewed empirical studies in the field of language assessment to provide an overview of which areas of the framework were being well researched. Hamp-Lyons and Lynch (1998) performed a similar analysis focusing on sixteen years of LTRC abstracts. Researchers in the field of educational measurement (Kane, 2001) and within language assessment (Bachman & Palmer, 2010) have continued to propose frameworks for validity research. Bachman and Palmer's Assessment Use Argument (AUA) expands to include validity research on the Decisions made based on score interpretations as well as Consequences of those decisions. Our survey investigates how the field of language assessment has evolved in its approach to validation since Kunnan (1998) and Hamp-Lyons and Lynch (1998), particularly in reference to methodological approaches. To facilitate this overview, we reviewed empirical validation studies published between 1999–2009 in relevant journals (e.g., *Language Testing*, *Language Assessment Quarterly*, etc), research reports (e.g., ETS, Cambridge), and doctoral theses. In phase one, studies were categorized according to the focus (criterion-related, content, consequential, etc.), methods, and objects of validation. In phase two, we related these facets to a recent framework for test validation, Bachman and Palmer's (2010) AUA. In addition to producing an overview of methodological approaches employed in the field of language testing over the last decade, the utility of recent validation frameworks (e.g., the AUA) for capturing the diversity of research practice will be discussed.

Kirsten Ackermann
Pearson

Saturday, June 25
10:10–10:40
Rackham Auditorium (1st Floor)

Investigating the use of academic vocabulary and its effect on test-taker performance

The purpose of this study is to investigate the extent to which test-taker performance on writing tasks on an academic English proficiency test at two different ability levels was affected by the use of academic vocabulary; i.e., academic words and academic collocations. As a prerequisite for this investigation, an academic word list and collocation list were compiled from a corpus comprising over 25 million academic words from textbooks and journal articles. The corpus covers 28 subjects and four academic disciplines including humanities, social science, natural and formal science, professions and applied science. In this study, 500 test-taker responses to two different item types—i.e., summarizing a written text, and writing an essay—were analyzed in relation to the two word lists using RANGE and WordSmith Tools 5.0. In addition, the relation between the proportion of academic vocabulary in the item prompts and the test-taker responses was determined to answer the questions to which extent these item prompts elicit academic vocabulary. Preliminary findings suggest that while both test-taker groups use academic vocabulary, there are difference in quality, frequency, and range of the academic words and collocations employed between the more proficient test-taker group and the less proficient one. There is evidence that the use of academic vocabulary systematically decreases as the scores descend, and that collocations are used for differing purposes by the two test-taker groups. The paper shortly introduces the academic word list and the academic collocation list. It discusses the differences in the quality, frequency, and range of academic vocabulary employed in test-taker responses of different levels of proficiency and the effect on the overall item score and specific trait scores.

Christine Doe
Queen's University

Saturday, June 25
10:45–11:15
Rackham Auditorium (1st Floor)

Validation of a large-scale assessment for diagnostic purposes across three contexts: Scoring, teaching, and learning

Large-scale assessments are increasingly being used for more than one purpose, such as admissions, placement, and diagnostic decision-making (e.g., Jang, 2009), with each additional use requiring validation regardless of previous studies investigating other purposes (Fulcher & Davidson, 2009). Despite this increased multiplicity of test use, there is limited validation research on adding diagnostic purposes—with the intention of directly benefiting teaching and learning—to existing large-scale assessments designed for high-stakes decision-making. A challenge with validating diagnostic purposes is to adequately balance investigations into the score interpretations and the intended beneficial consequences for teachers and students (Alderson, 2005; 2007; Davies & Elder, 2005; Fox, 2009). The Assessment Use Argument (AUA) makes explicit these internal and consequential validity questions through a two-stage validation argument (Bachman & Palmer, 2010). This research adopted the AUA, to validate the Canadian Academic English Language (CAEL) Assessment for diagnostic purposes, by forming a validity argument that asked, to what extent did the CAEL essay meet the new diagnostic scoring challenges from the rater perspective, and a utilization argument centered on teacher and students' use of the diagnostic information obtained from the assessment. This study employed three research phases at an English for Academic Purposes (EAP) program in one Canadian university. Data collection strategies included interview and verbal protocol data from two raters (phase one), interview and classroom observation data from one EAP course instructor (phase two), and interview and open-ended survey data from 50 English language learners (phase three). A multifaceted perception of CAEL for diagnostic purposes was observed: raters noted the greatest diagnostic potential at higher score levels, and teacher and student perceptions were largely influenced by previous diagnostic assessment experiences. This research emphasized the necessity of including multiple perspectives across contexts to form a deeper realization of the inferences and decisions made from diagnostic results.

Catherine Hayes
Birkbeck College &
Pearson Language Tests

Saturday, June 25
10:45–11:15
Rackham Amphitheatre (4th Floor)

Testing spelling: An investigation into non-native test takers' spelling errors

This research examined a corpus of misspellings made by native and nonnative English speakers during a gap-fill dictation item-type on a major international test of academic English. Although there is a wealth of literature on L1 spelling development and a few key studies into EFL spelling errors (see, e.g., Cook, 1999, Ibrahim 1978, Okada, 2005), there has been a lack of investigation into how spelling is treated in large scale language testing. This research drew on Lado's contrastive analysis hypothesis, error analysis theory, and modern day spell-checker research to examine three questions:

- What kind of errors do L2 learners make, and how do they compare to native speaker errors?
- How does spelling relate to overall English ability?
- Do learners from different L1 backgrounds have different error rates?

4,019 error tokens were collated from 605 test takers with 96 background languages. A scaled spelling score was created for each candidate. Spearman's rho correlation coefficients showed that spelling accuracy had a significantly strong relationship with overall English ability (rs = 0.819), and with writing skill test scores (rs = 0.843). ANOVA results showed that the orthographic type of the L1 influenced the number of spelling errors made on this task-type, in particular, test takers with a logographic language background had a significantly lower mean spelling score than the alphasyllabic, Roman alphabet and native-speaker groups. The 2,276 unique misspellings types in the corpus were examined and categorized. The types of errors made by EFL test takers were found to be similar to those made by the native speaker control group, although morphological errors accounted for the greatest number of single-spot error types. This study has implications for the teaching and testing of spelling.

Lorena Llosa
New York University

George Bunch
University of California, Santa Cruz

Saturday, June 25
11:20–11:50
Rackham Auditorium (1st Floor)

Validity issues in the placement testing of language minority students in community colleges

Community colleges represent the first point of access to public higher education for many language-minority students in the U.S. In California, for example, these students represent over 25% of the 2.5 million community college student population. In community colleges, the language testing and placement process is one of the first aspects of higher education encountered by language-minority students transitioning from U.S. high schools. This process represents high stakes for students' academic trajectories, as it determines whether students have access to credit-bearing English courses or will be assigned to developmental English or ESL courses that often do not earn credits toward a degree or transfer. Yet, despite the stakes involved in community college placement testing, little research has examined the placement process, the language tests that are at its core, and how this process affects language minority students. Through a systematic content analysis, we examined the six most commonly used, commercially available ESL and English placement tests in order to comparatively investigate their characteristics and the assumptions about language proficiency that underlie them. One of our findings was that these placement tests provide a very limited picture of what a student can do with language since the only skills assessed are students' ability to read and understand relatively short passages and their knowledge of grammar. This is particularly problematic for U.S.-educated language minority students (sometimes referred to as Generation 1.5) whose language profiles are different than those of traditional ESL students and monolingual students. This and other findings raise concerns as to whether the current system of placement testing in California's community colleges is appropriate for ensuring that language minority students are successful in pursuing academic pathways. The study raises important issues regarding the nature and validation of placement tests, a topic seldomly discussed despite the prevalence of placement tests in educational contexts.

Andrea Hellman
Missouri Southern State University

Saturday, June 25
11:20–11:50
Rackham Amphitheatre (4th Floor)

Assessing lexical richness in spontaneous speech data

Spontaneous speech data reflect the lexical resources available to language learners; these data can provide a comprehensive measure of vocabulary skills. Assessing spontaneous speech data for lexical richness has been particularly problematic. The paper discusses the advantages and disadvantages of various measures for the assessment of lexical richness. Although a few measures have been widely used to index lexical richness, the validity and reliability of these measures for particular purposes have been in dispute. In addition, software exist for the analysis of various components of lexical richness; however, the measurements obtained by different software are difficult to compare given the technicalities involved in sampling, data cleaning, and data preparation for analysis. Therefore, the various measures of lexical richness are specific to the application that calculates them. This paper elaborates a system suggested by Read (2000) for understanding the components of lexical richness as lexical variation, sophistication, density, and appropriateness. I discuss the measurements available to assess these components detailing the sensitivity inherent in each of the measures. To explore the relatedness of various submeasures of lexical richness, I use corpus data from transcripts of extensive spontaneous speech samples from interviews with native speakers of English and adult-onset learners of English. I make recommendations for improving the validity, reliability, and ease of assessment of lexical richness in adult speech samples. The goal of the paper is to serve as a guide for readers who interpret lexical richness measures and for researcher who employ these measures to index the vocabulary attainment of language learners.

Xin Yu
University of Bath

Saturday, June 25
1:30–2:00
Rackham Auditorium (1st Floor)

Computer assisted English speaking testing: A case study of CEOTS, China

This research aims to evaluate the use of the College English Oral Test System (CEOTS): a computer-assisted test of spoken English used in Chinese universities. Although this is a study of a particular test in a particular context, the findings suggest some general points about the use value of computer-assisted spoken English tests that are generalizable at the theoretical level. There are two main foci, the first being people's perception of the CEOTS in comparison with conventional face-to-face (F2F) tests, and the second being a comparison of the language output obtained from each of these test types. The notion of usefulness is reviewed, together with six key concepts: validity, reliability, authenticity, interactiveness, practicality, and fitness for purpose. The students' language output is evaluated in terms of lexical density, lexical variation, and communicative competence. As regards to the human and computer interaction, four aspects were reviewed: interface design, functional usage, computer anxiety, and gender. For the study a mixed method research design is employed, with, first, a structured questionnaire being implemented in order to investigate students' general attitudes towards computer assisted speaking tests. Second, 12 students attended mock tests using CEOTS and face-to-face tests, and were subsequently interviewed, along with four English teachers and three test developers. The findings reveal that the time restriction, the background noise, the prompts part and the technical quality of the test delivery, introduced construct-irrelevant factors that affected the validity of the CEOTS. The missing nonlinguistic elements can be criticized as another factor that could reduce reliability. Moreover, the participants adopted more connective discourse markers than particle discourse markers when using the CEOTS, which resulted longer and more complex sentences. A higher use of the past perfect tense, subjective mood and rhetorical questions were also found in the case of the CEOTS, as compared with F2F.

Masoomeh Estaji
Allameh Tabataba'i University

Saturday, June 25
1:30–2:00
Rackham Amphitheatre (4th Floor)

Assessment literacy: A call for more teacher education reform and professional development

Assessment is a widespread—if not intrinsic—feature of most language teaching programs worldwide. This has resulted in a proliferation of various standardized tests as well as in the introduction of teacher-conducted assessments used as a basis for reporting learners' progress and achievement against national standards (Brindley, 1997). Classroom assessment is an integral part of the language learning process and a powerful informed decision-making tool. However, relatively little is known about how teachers are dealing with these demands and, more importantly, how such assessment practices impact on their daily teaching. Unfortunately, not many language teachers are trained to make assessment decisions that will engage and motivate students and, as a result, enhance learning. In addition, studies investigating classroom-based assessment practices within the ESL/EFL school context (Arkoudis & O'Loughlin, 2004; Breen et al., 1997; Cheng et al., 2004; Cumming, 2001; Davison, 2002; Davison & Leung; 2001, 2002) have stressed the need for further research as the picture is not yet complete. This study is an attempt to examine teachers' perceptions about language assessment and the way they use language assessment in their classroom. Moreover, the teaching experiences of the participants are examined for their influence on the level of assessment literacy. The findings suggest that there is a significant difference in the perceptions that teachers have depending on the level of training they have in language assessment. Results also show that graduate and postgraduate teachers enjoy higher assessment literacy than undergraduate teachers, and those with prior teaching experience demonstrate higher assessment literacy. Thus, this study highlights the importance of providing adequate training in language assessment for all prospective language teachers, discusses the need, and presents ways of how teachers can become more "assessment literate" and how testing experts and teacher trainers can help in this direction.

Xiaoming Xi
Educational Testing Service

Jonathan Schmidgall
University of California, Los Angeles

Yuan Wang
Educational Testing Service

Saturday, June 25
2:05–2:35
Rackham Auditorium (1st Floor)

Examinee perceptions of automated scoring of speech and validity implications

SpeechRaterSM, a computer system for scoring spontaneous speech, is used to provide quick score feedback on the speaking section of the TOEFL® Practice On-line (TPO) test, which helps examinees prepare for the TOEFL iBTTM test. A critical validity issue for using SpeechRater is examinee perceptions, a topic underexplored in automated scoring research. This study investigates examinees' general perceptions about computer speech scoring, and more specifically the impact of SpeechRater on their test-taking strategies, and perceptions, interpretations and uses of the SpeechRater scores. We administered a survey to 469 TPO users of various background characteristics. We then interviewed 36 users and conducted two focus groups in China as Chinese speakers are one of the largest TPO user groups and constituted over half of the survey sample. Although a large proportion of the survey participants showed some confidence in computer scoring for speaking, the majority perceived human scoring as more accurate and would still prefer human scoring. They also viewed computer scoring combined with human scoring more positively than computer scoring only, and this trend was much more prominent among the interview and focus group participants, who developed a better understanding of automated scoring. With regard to their specific reactions to SpeechRater, overall, the participants showed a good level of acceptance of using SpeechRater for scoring TPO despite some reservations. In addition, most did not change the way they responded to the practice test because of SpeechRater scoring although they would be more likely to "trick" the computer if computer scores were used alone for high-stakes decisions. Devising strategies to "trick" the computer when it is used along with human raters to score speaking was beyond their imagination although the possibility still existed. Implications of the results for the validity of automated scoring for different contexts and uses will be discussed.

Huan Wang
CTB / McGraw Hill

Saturday, June 25
2:05–2:35
Rackham Amphitheatre (4th Floor)

Classifying and reporting language proficiency patterns to inform integrative language teaching

An important factor contributing to successful integrative task-based language instruction is to choose tasks with appropriate difficulty levels. One way to decide the appropriate difficulty level is to refer to a student's overall English language proficiency level. However, language proficiency has long been regarded as multifaceted. Due to English language learners (ELLs)' potentially unbalanced linguistic development, students placed on the same overall proficiency level may actually have different areas of strengths and weaknesses. To ensure appropriate scaffolding and avoid student frustration, it is important to target instruction based on ELLs' proficiency patterns rather than their overall language proficiency. The present study examined and compared four methods of classifying proficiency patterns with the purpose of informing the design of an assessment report for ESL/EFL teachers. Method 1 utilizes the overall language proficiency levels; Method 2, the domain proficiency patterns, with four domains representing listening, speaking, reading, and writing; Method 3, cluster analysis of domain scale scores; and Method 4, patterns highlighting at-risk groups. The empirical test data used in the study were collected from 1,400 students who participated in an online English language test. The test had four modules that were intended to assess skills in reading, writing, listening and speaking, respectively. The comparison results suggest more instructional usefulness of Method 4 than Methods 1 and 3. In addition, Method 4 is more efficient than Method 2 in reducing the number of groups to a manageable level for a classroom teacher. The study revealed the complexity in designing an assessment report for facilitating integrative language teaching. Potential utilization of a dynamic web-based assessment report was discussed for greater flexibility in accommodating ESL/EFL teachers' various pedagogical needs.

C. Ray Graham
Brigham Young University

Saturday, June 25
2:40–3:10
Rackham Auditorium (1st Floor)

Examining the validity of an elicited imitation instrument to test oral language in Spanish

Bernstein, Moere, and Cheng (2010) present evidence to support the validity of the Versant, a fully automated test which uses elicited imitation (EI) as a principal strategy. The evidence that they present falls into three main categories: construct definition, construct representation, and concurrent evidence. They also present interpretive arguments that attempt to extrapolate from observed test behavior to language behavior beyond the test and from test scores to intended test use. In this paper we present evidence similar to that of Bernstein et al. regarding construct definition and representation, and we present the results of a study in which we show that our elicited imitation instrument has concurrent validity with the ACTFL OPI. We claim that even though the EI technique by itself may lack face validity, for the present such tests can be used as placement tests and as a predictors of students' readiness to take a more widely accepted and expensive high-stakes oral language test. The paper reviews the theory underlying EI as a method of examining linguistic competence and it presents the results of a study involving the administration of an 84-item elicited imitation instrument in Spanish concurrently with an official ACTFL OPI test to 94 learners of Spanish as a foreign language who ranged in ability from novice to superior. The test requires less than 15 minutes to complete and can be administered by computer in a lab. The correlation between the scores on the two instruments was r = 0.92. In the presentation we discuss the design of the instrument, the details of its administration, and the scoring results using both human judges and automatic speech recognition.

Ling Guanming
Educational Testing Service

Pamela Mollaun
Educational Testing Service

Xiaoming Xi
Educational Testing Service

Saturday, June 25
2:40–3:10
Rackham Amphitheatre (4th Floor)

Impact of rater fatigue on the scoring quality of speaking responses

Continuously rating for hours can negatively affect raters' judgment (Anastasi, 1979) and the rating quality of speaking responses. Even though standardized training and thorough monitoring of raters have helped maintain a high quality of scoring, greater knowledge of fatigue effect could inform further refinement of scoring practices, thereby improving the reliability of scores. This study investigated the fatigue effects by comparing 72 raters' performance on more than 5,000 speaking responses in four different shift schedules: 8–2, 8–4, 6–2 and, 6–3 (8 or 6 represents the shift length in hours, and 2, 3, or 4 represents the session length). The rating quality within each scoring hour was examined and compared among rater groups working on different shifts. An end-of-shift survey about raters' perceptions of fatigue was also delivered. The scores assigned within each hour in each shift were compared to the expert scores (assigned prior to this study) using three statistics: the proportion of exact agreement, the root mean square of deviations (RMSD), and the Kappa between the two sets of scores. The data based on the survey and the ratings were linked and analyzed as well. The results indicated that the overall scoring quality was satisfactory for the ratings provided in each of the four shifts. However, the rating quality measured by the three statistics varied substantially across hours and had a general increasing pattern for the eight-hour shifts. Ratings in the 6–2 shift were the most consistent and accurate, while those in the 8–4 shift were the least. This is the first study examining the fatigue effects on raters scoring speaking responses. The results suggest that there are spaces for improvement on the shift design and quality control for speaking response scored by human raters.

Kirby C. Grabowski
Teachers College,
Columbia University

Saturday, June 25
5:05–5:35
Rackham Amphitheatre (4th Floor)

A discourse analysis of the pragmatic meanings generated on a role-play speaking test

Research in applied linguistics has demonstrated that language used in the real world is often indirect, and misunderstandings in interaction often stem from a failure by the interlocutors to appropriately convey and/or interpret implied, or pragmatic, meanings. Given their relative importance in everyday communicative contexts, pragmatic aspects of language, including indirectness, have been integrated into current influential models of language ability. However, even with these models as a guideline for test development, pragmatic aspects have remained, in large part, noticeably absent in both large- and small-scale assessments. This makes it difficult to make inferences about a test taker's language ability as it pertains to real world contexts. For the current study, a test was developed to measure test takers' grammatical and pragmatic knowledge in the context of speaking. 102 test takers from three different levels of ability participated in four role-play tasks, representing real world situations, in which each test taker was paired with the same native speaker partner. The conversations were recorded and then scored by two raters. Although the data were analyzed statistically using both multivariate generalizability theory and many-facet Rasch measurement in a separate analysis, the purpose of the current paper is to show how raters were able to isolate and score the appropriateness of a range of pragmatic meanings in the responses. To accomplish this, a discourse analysis of pragmatic meanings was performed for one task across four ability levels, highlighting contextualization cues to uncover degrees of pragmatic success and failure in the performance samples. Findings from the statistical analysis will also be briefly discussed in order to support the qualitative analyses.

Language assessment literacy: Communicating theories of language testing to users

Thursday, June 23 • 5:00–7:00 • Rackham Auditorium (1st Floor)

Session Chair
Natalie Nordby Chen
Cambridge Michigan Language Assessments

Organizer
Ofra Inbar-Lourie
Tel-Aviv University & Beit Berl College

Discussant
Lynda Taylor
University of Cambridge ESOL Examinations

Paper Presenters
Cathie Elder
University of Melbourne

April Ginther
Purdue University

Glenn Fulcher
University of Leicester

Margaret E. Malone
Center for Applied Linguistics

Ofra Inbar-Lourie
Tel-Aviv University & Beit Berl College

The knowledge base required for performing assessment functions, hence assessment literacy, is currently the focus of research and discussion in the general assessment domain. Likewise, language testing has recently begun discussion of the language assessment literacy required to perform language assessment functions for different purposes in diverse contexts (Inbar-Lourie, 2008; Taylor, 2009). Questions arise as to the scope of knowledge required by different stakeholders, its attainment and application. This is especially relevant to the LTRC 2011 conference, which celebrates Lado's contribution to the field as it considers the components of language testing.

This symposium brings together research studies that examine language assessment literacy in different professional settings. The first paper argues that assessment literacy is context-specific and therefore treating it as a general body of knowledge to be transmitted to test users may have limited impact. The second paper examines the assessment literacy needs of test score consumers in higher education, and paper three investigates successful and unsuccessful ways that concepts of validity and reliability have been communicated to language educators. The last paper describes the process of acquiring language assessment literacy by prospective language teachers in two frameworks: a specialized assessment course and one directed at a more general education audience.

The duration of the symposium is two hours: following a short introduction (five minutes) each presenter will be allowed fifteen minutes for presentation, followed by five minutes for clarification questions and brief discussion (20 x 4). The discussant will then relate to the research studies and the emerging themes (ten minutes) followed by a general fifteen-minute discussion among the participants and audience. Likely themes for discussion will include the nature of the language assessment literacy knowledge base, difficulties in effectively conveying theoretical underpinnings, the process of acquiring language assessment literacy, and recommendations for implementation and further research.

Language assessment literacy: Communicating theories of language testing to users :: individual abstracts

Cathie Elder
University of Melbourne

April Ginther
Purdue University

Assessment literacy as LSP

Discussions of assessment literacy tend to emphasize the importance of training a range of stakeholders in what they need to know to be able to understand and use tests or assessment appropriately (Brindley, 2001; Inbar-Lourie, 2008; Taylor, 2009). However, the question of what knowledge is required varies widely according to the type of assessment and the context of its use.

This paper outlines a range of contexts in the authors' professional experience where lack of assessment literacy among particular stakeholders has proved to be (1) an obstacle to the implementation of an assessment project and/or (2) a potential threat to its appropriate use or interpretation of test outcomes. The contexts of score use include (a) an evaluation project assessing the efficacy of heritage language teaching in the school context, (b) the widespread use of TOEFL scores for international student admissions to North American institutions of higher learning, and (c) post-entry screening of oral English skills for prospective international teaching assistants. What is emphasized in these accounts is the different types of understandings needed in each context, the diverse backgrounds of score users, and the impact of the local policy context on how test information is construed. It is argued that treating assessment literacy as a general body of knowledge to be transmitted to test users through training courses or textbooks may have limited impact, and may also be impractical, given that policy makers and other test users may have little motivation and/or resources to acquire such knowledge. We conclude that test developers should consider the assessment literacy of test users at the design stage of any project, and that appropriate networks be set up to ensure that these users can access the expertise/resources needed to enable sound assessment practice and appropriate decisions and actions on the basis of test scores.

Glenn Fulcher
University of Leicester

Operationalizing assessment literacy

Language testing has seen unprecedented expansion during the first part of the 21st century, due in no small measure to the growing use of language tests for educational accountability and immigration control. As a result there is an increasing need for the language testing profession to consider more precisely what it means by 'assessment literacy' and to articulate its role in the creation of new pedagogic materials and programs in language testing and assessment to meet the changing needs of teachers and other stakeholders for a new age. This paper describes a research project in which a survey instrument was developed, piloted and delivered on the internet to elicit the assessment training needs of language teachers. The results were used to inform the design of new teaching materials and the further development of online resource materials that could be used to support program delivery. The paper will focus on the creation of new web-based materials to supplement text-based teaching materials, thus incorporating multimedia and web 2.0 technologies as a basis for learning, seminars, assignments and project work. It is argued that the development of rich content-open source materials can facilitate learning and teaching wherever language testing is studied.

Language assessment literacy: Communicating theories of language testing to users :: individual abstracts

Margaret E. Malone
Center for Applied Linguistics

The essentials of assessment literacy in a post-Messick world: Contrasts between testers and users

This paper investigates efforts to elicit language testers' beliefs about measurement basics and compare them to the related views of language educators to gauge the latters' level of language assessment literacy. Specifically, the paper investigates successful and unsuccessful ways that concepts of validity and reliability have been communicated to educators and examines new ways to approach it. Language assessment literacy refers to language instructors' familiarity with measurement practices and the application of this knowledge to classroom practices in general and specifically to issues of assessing language (Stiggins, 2001; Inbar-Lourie, 2008; Taylor, 2009). While widely agreed that classroom teachers must assess student progress (Schafer, 1993; NEA, 1983), many teachers and other test users have a limited understanding of assessment fundamentals (Popham, 2009).

As Messick's approach to validity (1989) has transformed the work of language testers during the past two decades, it is a challenge to convey such information to language instructors. This paper accomplishes four objectives: highlight some definitions of core knowledge of language testing deemed necessary for language instructors, describe three assessment literacy projects for language instructors, examine successes and challenges of these projects and identify foci for future research. First, the paper examines recommendations of language testing experts, elicited via focus groups and surveys (N = 24), on the necessary components of language assessment literacy and the critical need to convey Messick's concepts. Next, the paper demonstrates how such concepts were explained via three language assessment literacy projects for language educators: a self-access workshop, (N = 12), an e-learning course (N = 100), and a blended learning course (N=120). Analyses of the projects reveal the challenges of including technical information considered essential by testers in language assessment literacy training, while also addressing the practical needs of teachers, including ways to present technical information in a user-friendly manner. Finally, the paper identifies areas for new research.

Language assessment literacy: Communicating theories of language testing to users :: individual abstracts

Ofra Inbar-Lourie
Tel-Aviv University
& Beit Berl College

Language assessment literacy among prospective language teachers: From novice to proficient

Language assessment literacy is a relatively novel term which refers to the knowledge base required by different stakeholders for planning, conducting and analyzing assessment activities (Brindley, 2001; Inbar-Lourie, 2008; Taylor, 2009). Though the precise nature of this knowledge base is not clearly defined nor agreed upon, there seems to be overall consensus as to the fact that teachers need to become assessment literate due to their central role and involvement in the assessment process. This knowledge is gained via participation in assorted language assessment courses or professional workshops (Brown & Bailey, 2008; O'Loughlin, 2006). Though deemed vital, there is not much research available about the nature of the frameworks intended at developing language assessment literacy, their contents and points of emphasis, whether general or language specific, and the transformation the prospective teachers go through in acquiring language assessment skills, from novice to proficient.

This research aimed to gain insight into the language assessment literacy development process by following prospective EFL teachers through a year long assessment course comprised of two components: The first part included general assessment principles and procedures, while the contents of the second module focused on language assessment and was geared particularly to assessing English language proficiency in accordance with the national EFL curriculum. In addition, the EFL teacher candidates were compared to another group of pre-service language teachers (English and Arabic), who participated in a general assessment course together with teachers of other subject areas. The data collection tools included questionnaires, interviews and focus groups. Results will be discussed with reference to the format and contents, the "what" and the "how" of language assessment courses in teacher education programs, as well as with regard to recommendations for further research in this area.

Rating scales for writing/speaking: Issues of development and use

Friday, June 24 • 5:15–7:15 • Rackham Auditorium (1st Floor)

Session Chair
Sarah Briggs
Cambridge Michigan Language Assessments

Organizers
Evelina D. Galaczi
University of Cambridge ESOL Examinations

Gad S. Lim
University of Cambridge ESOL Examinations

Discussant
Sara Weigle

Paper Presenters
Anthony Green
University of Bedfordshire

Gad S. Lim
University of Cambridge ESOL Examinations

India C. Plough
Cambridge Michigan Language Assessments

Janna Fox
Carleton University

Natasha Artemeva
Carleton University

In the half century since Robert Lado's *Language Testing* (1961), much has changed with regard to testing and notions of test validity. Performance assessment is now the preferred way of testing the productive skills, and their validity depends in large part on the rating scales used, as rating scales are the de facto constructs on which these tests are based (McNamara, 2002; Turner, 2000; Weigle, 2002). Historically, however, such scales have been developed intuitively rather than empirically (Fulcher, 2003). Given the critical role of rating scales in the assessment process, it is important that they are demonstrably valid, both in their construction and in their different contexts of use (Fulcher & Davidson, 2009).

While rating scales are undeniably central to performance assessment, there is still relatively little publicly available research into issues salient to their development, interpretation and use, such as,

- considerations in the development of performance descriptors
- accounting for various validity requirements in scale design
- the role of different stakeholders in the development process
- the adequacy of traditional psychometric approaches to rating scales and their use.

The papers in this symposium explore these issues, drawing on current practice in different countries to present distinct yet complementary perspectives, addressing in turn, the integration of the CEFR into an operational rating scale; the need to balance and meet multiple validation requirements in scale development; stakeholder engagement in assessment scale development; issues of scale use in view of different stakeholders' criteria and scale interpretations.

Each paper will summarize its contribution to the theory and practice of rating scale development and use, addressing the challenges faced by scale developers and users. Collectively, the papers will contribute to the conference theme of validity in language testing and the provision of standards for the critical evaluation of rating scales.

Rating scales for writing/speaking: Issues of development and use :: individual abstracts

Anthony Green
University of Bedfordshire

From framework to scale: Integrating the CEFR into an operational rating scale

Situating a test in relation to a widely recognized framework such as the Common European Framework of Reference for Languages (CEFR) increases its transparency and relevance for users. Guidance on relating assessments to the CEFR suggests that "exploitation of the CEFR during the process of . . . rating scale development strengthens the content-related claim to linkage" (Council of Europe, 2009, p. 2). However, while the CEFR provides illustrative descriptions of levels, it does not offer substantial guidance to users on adapting this illustrative material to build locally relevant operational scales.

Addressing this gap, this paper describes the (re)development of a revised set of rating scales for use in the speaking components of a suite of English language tests targeting CEFR levels A2 to C2. The tests were not developed from the CEFR and reflect a distinct testing construct that predates its publication. The development process therefore involved a complex encounter between the CEFR, the existing rating scales, advances in the theoretical literature and specific operational requirements.

The paper will focus on:

- changes in the wording of the scales during the modification process
- the ways in which the CEFR was integrated into the updated scales, and
- factors influencing these modifications.

Consultants were asked to review the rating scales in light of the CEFR and of advances in the theoretical literature. Drawing on these sources, a team (involving the consultants and staff from the test provider) authored, reviewed and revised a set of scale descriptors, relating these to samples of test taker performance. Contributing to this process, wider questionnaire surveys invited examiners to comment on the existing scales and subsequent revisions. These were then piloted under operational conditions. Conclusions from this process will be related to the broader debate on the theory and practice of rating scale design.

Rating scales for writing/speaking: Issues of development and use :: individual abstracts

Gad S. Lim
University of Cambridge
ESOL Examinations

Meeting multiple validation requirements in rating scale development

Many requirements attend large-scale tests of writing ability. These tests need to meet the requirements of validity, reliability, and useability, they also need to be shown to relate to external benchmarks, and in addition meet operational demands and remain practically viable. As new requirements can emerge, test validation is necessarily a continual process, and suitability to new conditions needs to be demonstrated (Fulcher & Davidson, 2009).

In view of continual validation, this paper details the development of a new set of rating scales for a suite of writing exams, which needed to meet a number of requirements. In addition to reflecting the writing construct of the exams involved, the scales had to explicitly align outcomes to the CEFR, had to be useable across a number of exams at each CEFR level, and had to vertically link performance to adjacent CEFR levels. This paper focuses on the development process to meet those needs, a process that involved review of the writing construct; production of and repeated revisions of draft analytic scales involving expert input and empirical data; collection of qualitative feedback throughout the process; several iterations of scaling exercises to establish relationship to the CEFR; and a marking trial (n = 15,684 ratings) to provide converging a priori evidence for the validity, reliability, and useability of the scales with the above mentioned specifications.

Using this scale development project as an illustration, current views on test validation and recommendations for scale development, such as those proposed in the CEFR (Council of Europe, 2001), are confirmed, challenged, and extended.

India C. Plough
Cambridge Michigan
Language Assessments

Stakeholder engagement in test design and rubric development

In the last fifty years, notable improvements in the assessment of second language speaking performance have been made. There now exists a significant body of research into the interactions among examiner discourse, test taker performance, and raters' rating of said performances, which suggests that multiple stakeholders have important roles in speaking test design. Simultaneously, the value of data-driven methods for developing rating scales has been brought to the forefront (Council of Europe, 2001; Upshur & Turner, 2002). Deriving scales empirically from the performance of specific test takers, in the context of a specific test for a particular purpose, addresses the need for improved reliability and contributes to their validity argument. In the shift to empirical, data-driven scale development, the engagement, collaboration, and buy-in of stakeholders at each stage of the process is crucial.

This paper will present the development process of an empirically-derived set of scales for a high-stakes speaking test, explaining key stages in a multi-stage scale and test development process, as well as potential caveats. The theoretical foundation of the test design, based on a synthesis of task-based and assessment research, is briefly reviewed. The paper then focuses on the steps employed to involve four key stakeholders—the test takers, the examiners, the raters, and the test developers—in the development of the test instrument and then in the scoring rubric. The presentation concludes with a discussion of the age-old dilemmas of merging theory and practice, attending to multiple priorities, and of negotiating resolutions to differing stakeholder perspectives. Areas for potential compromise, as well as those scale and test development processes that should not be eliminated, are covered.

Rating scales for writing/speaking: Issues of development and use :: individual abstracts

Janna Fox
Carleton University

Natasha Artemeva
Carleton University

Raters as stakeholders: Uptake in the context of diagnostic assessment

Although the literature on rater training has stressed issues of reliability and consistency, some studies have provided evidence that variable rater interpretations, or uptake, can contribute useful information in the assessment of test performance (Brown, Iwashita, McNamara, & O'Hagan, 2005; Jacoby & McNamara, 1999). In the context of writing tests, descriptors and scales attempt to systematize raters' uptake so that they will interpret texts as similarly as possible; achieving high rater agreement is considered a fundamental requirement in most validity arguments. Following Moss (1992), however, this study suggests that traditional psychometric approaches to reliability may be insufficient in the context of diagnostic assessment.

The study investigated a diagnostic assessment of academic English (n = 300 first-year undergraduate engineering students), the results of which were to be linked to specific course support. One of the test tasks required students to provide a written interpretation of a graph. Eight experienced raters rated the writing task; analysis indicated that their rating was consistent and reliable (r = 0.93). In order to align the test results with specific instruction for the engineering students, twenty test scripts were remarked by stakeholders involved in providing classroom support, namely: engineering course designers (n = 3), an engineering and technical writing specialist, an engineering professor, and a professional engineer. The stakeholders were provided with scales and descriptors but had no rater training. The selected scripts represented a range of performance based on the original test scores awarded by the trained raters.

Findings from the study demonstrate how uptake of the written texts, scales and descriptors was informed by disciplinary perspectives. They suggest that in the context of diagnostic assessment, if valid inferences are to be drawn from test performance, traditional psychometric approaches to raters' use of scales may be inadequate, unless they are considered in relation to the uptake of stakeholders' varying indigenous criteria.

Workforce language assessment in the 21st-century knowledge economy: A quality management perspective

Saturday, June 25 • 3:25–5:25 • Rackham Auditorium (1st Floor)

Session Chair
Elana Shohamy
Tel Aviv University

Organizer
Kathleen M. Bailey
Monterey Institute of International Studies

Discussant
Jean Turner
Monterey Institute of International Studies

Paper Presenters
Kathleen M. Bailey
Monterey Institute of International Studies

Ryan Damerow
The International Research Foundation for English Language Education

Courtney Pahl
Monterey Institute of International Studies

Michael Milanovic
University of Cambridge ESOL Examinations

Nick Saville
University of Cambridge ESOL Examinations

Mary Ann Christison
University of Utah

In this symposium we will address assessment practices—those that are used and those that are needed—in the context of the 21st-century global knowledge economy. The presenters will discuss assessment in various businesses, industries and professions. Questions to be considered include the following:

1. How are language needs assessed in particular workforce contexts?
2. What procedures are used to test (potential) employees'language abilities?
3. What language abilities are the foci of assessment practices—language elements and/or integrated skills use?
4. What features of the workforce context and culture are considered in developing or choosing language assessment tools?
5. Who is responsible for the assessment of (potential) employees' language abilities?
6. How are the assessment procedures validated for their use in the particular context?
7. How are the assessment data used in decision making?

We will make the case for adopting a quality management approach to improve testing and to ensure that appropriate professional standards are met. In this respect, language test developers need to adopt the kind of managerial practices which enable successful organizations to implement error-free processes. In making this case, quality is defined in the context of language assessment. We will describe ways in which quality management can be employed to achieve the required quality goals. We will argue that there is convergence between the twin concepts of quality and validity; quality management provides an appropriate basis for guaranteeing the consistency of the processes which underpin a validity argument and provides the tools and techniques for linking theoretical concerns with the practical realities of establishing and administering assessment systems. By adopting a quality management approach, it is also possible to ensure that processes are continually improved and standards raised. This view is consistent with the concept of validation as a long-term, ongoing activity.

Workforce language assessment in the 21st-century knowledge economy :: individual abstracts

Kathleen M. Bailey
Monterey Institute
of International Studies

Ryan Damerow
The International Research
Foundation for English
Language Education

Courtney Pahl
Monterey Institute
of International Studies

Surveying English language assessment practices in international plurilingual organizations

This paper will discuss the initial findings of a survey about the roles and functions of English (and other languages) in international plurilingual organizations. These are businesses and nonprofit organizations that carry out their operations across national boundaries and conduct those operations in at least two languages. The entire study has six broad research questions:

1. What functions do English and other language serve in international plurilingual organizations in the 21st century knowledge-based economy?
2. What are the necessary language skills individuals must possess to be successful as competent employees and leaders in international plurilingual organizations?
3. What assessment procedures and instruments are used to determine whether such individuals possess the necessary skills to successfully accomplish those roles?
4. What economic impact do proficient (and less proficient) plurilingual employees and leaders have on the operations of international organizations?
5. How do international organizations determine their return on investment (ROI) when they provide language development opportunities for employees?
6. What is the ROI when organizations take steps to develop employees' language skills?

This paper will focus specifically on the first three questions above. This research follows from an investigation commissioned by TIRF—The International Research Foundation for English Language Education (www.tirfonline.org) about the impact of English and plurilingualism in global corporations.

Michael Milanovic
University of Cambridge
ESOL Examinations

Nick Saville
University of Cambridge
ESOL Examinations

Using quality management to improve language testing

This paper makes the case for adopting a quality management (QM) approach to improve testing and to ensure that appropriate professional standards are met. In this respect, language test developers need to adopt the kind of managerial practices which enable successful organisations to implement error-free processes.

In making this case, quality is defined in the context of language assessment and the ways is which quality management can be employed to achieve the required quality goals are described. It is argued that there is convergence between the twin concepts of quality and validity; quality management provides an appropriate basis for guaranteeing the consistency of the processes which underpin a validity argument and provides the tools and techniques for linking theoretical concerns with the practical realities of setting up and administering assessment systems. By adopting a QM approach, it is also possible to ensure that processes are continually improved and standards raised; this is in keeping with the concept of validation as a long-term, ongoing activity (Kane, 2006).

Workforce language assessment in the 21st-century knowledge economy :: individual abstracts

Mary Ann Christison
University of Utah

Issues in assessing workplace ESL instructional programs

This paper identifies five key issues related to workplace language instruction in terms of assessing both program outcomes and the language skills of the learners. These issues are related to understanding the amount and types of research available on workplace language instruction, moderating learner and employer expectations, communicating to lay persons about the complex nature of learning another language, measuring outcomes, connecting language training to worker on the job performance, and using assessment data in decision-making. Understanding these key issues can help language assessment professionals improve assessment practices in workplace environments.

Cho Bokyung
Korea Institute for
Curriculum and Evaluation

Developing the Internet-based testing system for national English ability test in Korea

This research introduces to the procedures of developing the Internet-based testing system for National English Ability Test (NEAT) in Korea. Since 2007, Korea Institute for Curriculum and Evaluation (KICE) has developed NEAT and already completed several field tests with a large size of examinees. The purpose of developing this test is to measure test-takers' four English skills (listening, reading, speaking, and writing) effectively and replace the English section, a paper-based test, on Korean University Entrance Examination which can assess only students' reading and listening abilities. Because over 600,000 Korean high school students take the University Entrance Examination every year, it is not possible to assess individual student's English speaking and writing skills by following a traditional system, such as face-to-face interview, audio tape recording and scoring. However, the Internet operating system can make it possible to evaluate a large number of students' speaking and writing abilities because over 50,000 examinees can take NEAT at the same time and raters can scores their scores effectively. Specifically, NEAT trained over 3,000 raters and designed the scoring system for these raters to grade these examinees' responses at the same time. Compared to the traditional operating system, this Internet-based system can save money, time, and energy to evaluate large size of students. To enhance the function and stability of this testing system, KICE has employed advanced networking technology, developed the rating system, adopted the clouding computing system, etc. NEAT can suggest an effective way to measure a large number of examinees' English reading and listening abilities as well as their speaking and writing abilities through the Internet.

Vanessa Borges-Almeida
University of Brasilia

Douglas Altamiro Consolo
State University of Sao Paulo

EPPLE–language proficiency examination for foreign language teachers

It has now been almost a decade since our first studies investigating language proficiency for teachers of foreign languages began in Brazil. In all these years, a group of researchers have been involved in collecting information to support the design and use of test tasks and scales for that purpose, leading to the development of the EPPLE examination. The EPPLE consists of a set of two integrated tests dealing with topics related to the field of language teaching and learning. The written test contains questions that require the candidate to read and understand a text in their area of expertise in order to be able to fulfill tasks considerably common for teachers, such as correcting a student's written essay. The oral test is taken in pairs and consists of two tasks. In the first one the candidate is involved in a discussion based on the content of a video extract dealing with the domain of language teaching, whereas in the second one they are presented with linguistic difficulties faced by students, and must offer explanations and solutions to them. The EPPLE examination is founded in a view of language that considers the language use domain for the language teacher to go beyond the general proficiency or the ability to provide simple class commands. More than that, this domain is deeply characterized by metalanguage. Several unfinished studies have been investigating the correlation between the type of language produced at the examination and that one produced in real-life settings in order to validate what has been claimed to be that language use domain. Thus, such moment represents a turning point for those studies and the EPPLE itself, and this poster intends to address some of their issues and results.

Nathan Carr
California State University, Fullerton

Training teachers to write good short-answer automated scoring keys

This study involves the automated scoring of limited-production reading comprehension questions. Clauser, Kane, and Swanson (2002) point out that previous studies of computer-automated scoring (CAS) have focused on the comparability of human- and computer-produced scores. Two additional studies (Clauser, Harik, & Clyman, 2000; Clauser, Swanson, & Clyman, 2000) further note important concerns over how representative the raters used in such comparability studies might be, and how representative the scoring criteria or algorithms used for CAS might be. To date, however, no studies appear to have been performed looking at this question in the context of language assessment, or in the area of limited-production tasks. This is surprising, as the generalizability of scoring keys is a crucial prerequisite for a test's construct validity. This poster describes the procedures used to train 7 pre- and in-service ESOL teachers—graduate students in a post-graduate TESOL program—to write scoring keys for a limited production computer-scored web-based test (WBT). It also addresses the lessons learned during this project, particularly what areas proved to be most problematic when the participants wrote the keys.

Frederick Cline
Educational Testing Service

Eleanor Sanford
MetaMetrics

Amber Aguirre
MetaMetrics

Linking TOEFL junior reading scores to the Lexile measure

The TOEFL® Junior test is a low-stakes assessment designed to assess the degree to which students aged 11–14 have attained proficiency in the academic and social English-language skills representative of English-medium instructional environments. The test includes three sections—Reading, Listening and Language Form and Meaning. The Reading section addresses the ability to read and comprehend both academic and nonacademic texts. TOEFL Junior examinees will be provided a Lexile® measure linked to their reading section results. Lexile measures use the same scale for both reader ability and text difficulty, so a single score provides an estimate of a person's reading level that can also be used to determine a range of reading material that is appropriate for the reader.The poster describes the results of the study used to link the Lexile scale to the TOEFL Junior reading scale. A sample of 1,159 examinees from the first TOEFL Junior administration in Asia, Europe, South America and the Middle East also completed a Lexile linking test. The correlation between the Lexile and TOEFL Junior scores was 0.80 and an equipercentile approach was selected for the linking process due to a slightly non-linear relationship.

Troy Cox
Brigham Young University

C. Ray Graham
Brigham Young University

Elicited imitation: A comparison of different scoring methods

Elicited Imitation (EI) has been used for years as a method to research oral language development in areas such as L1 progress in children, abnormal language development and second language acquisition. In recent years, more research has focused on how EI can be used to assess speaking ability. One of the reasons for this interest is the ability to use Automatic Speech Recognition (ASR) to rate the items, thus making the grading of oral language an economic possibility for situations in which spoken language is not tested at all. A further possibility of this research would be to use Item Response Theory (IRT) to create item banks that could be used in computer adaptive testing. With all of this research, however, very little has been done to examine the different scoring methods of EI. This presentation will compare different scoring methods to see how the effect of the method impacts the rating of student ability. The data to be analyzed will come from an EI test that has been administered over the last three years and has over 1,000 respondents. First, we'll use the entire dataset to compare human scoring and ASR scoring and see how the raw scores compare with the person ability estimates calculated from different IRT models including a binary scoring model, Andrich's rating scale model and Master's partial credit model. Finally, we'll compare these scoring methods with a subset of the data that has been scored holistically. The results of this comparison can be used to decide which scoring method provides the most information in assessing speaking. Furthermore, it will address the strengths and limitations of using ASR to rate EI so assessment professionals can make informed decisions in choosing how to score EI.

Francesca Di Silvio
Center for Applied Linguistics

Margaret E. Malone
Center for Applied Linguistics

Developments in language testing: Evolution of a computerized test of oral proficiency

With the movement towards communicative language teaching and its assessment in the United States (Hymes, 1971; Bachman and Palmer, 1996), there arose the need for instruments to assess the general proficiency of language learners. A direct, or face-to-face, interview was initially the norm. However, the associated time and costs for administering and rating such interviews can be prohibitive for some test users. In addition, the power structure of interviewer and interviewee has been investigated and critiqued extensively (Savignon, 1985; Lazaraton, 1992). In the intervening years, language testing professionals have investigated and applied the use of technology to make oral proficiency assessments that are more feasible and cost-effective for large-scale testing, including paired tests (UCLES, 2001) and semi-direct tests (ETS, 2001; Chalhoub-Deville, 1997; Stansfield and Kenyon, 1992). Such approaches led to the development of computerized oral proficiency tests. This poster describes the evolution of a computerized oral proficiency test in the context of growth within the language testing field towards varied approaches and technological innovations to facilitate testing of oral proficiency. It outlines the history of oral proficiency assessment with a particular focus on the development of different approaches for testing of speaking including semi-direct measures and tape- and computer-based deliveries. In examining the design of a computer-based test of oral proficiency the poster highlights differences in test formats and consequent benefits and limitations for test stakeholders in contrast with other approaches. Research findings on examinee performance and attitudes towards computer-based oral proficiency testing are also presented. This overview of the development of a particular oral proficiency assessment is intended to stimulate discussion among language testers of potential applications of available technology in the development of valid, reliable, and practical tests of speaking skills.

William Eilfort
Avant Assessment

A framework for applying target-language expertise in less commonly tested languages

This presentation discusses a framework for evaluating target language experts' contributions during test development. It arises from a context where language proficiency tests in listening and reading are developed in less commonly taught languages. Test development is often conducted in teams led by test developers who are not experts in the target language themselves. In such cases, the quality of the test depends heavily on the contributions and review comments of various target language experts. During the development process, multiple judgments are commonly gathered regarding the content of the test. However, it is common for these judgments to conflict with one another; in such cases, the test developer must reconcile these judgments and come to a final decision concerning any changes to test materials. This framework lays out details of two key dimensions on which judgments need to be assessed and shows examples of conflicting judgments, analysis, and decision resolution. The first key dimension relates to the target language itself. This is relevant when review comments involve the authenticity of test passages or whether passages adhere to test specifications for language variety and to any existing standard. The framework defines aspects of this dimension that the test developer needs to know to evaluate the merits of each comment, such as whether there is a recognized standard variety for the language at all, or whether there are more than one. The second key dimension relates to the target-language experts. It is important to know their educational background in the target language, as well as the amount and currency of their interaction with other members of the target-language community. In some of the languages in question, it can be difficult to assess target-language experts' qualifications. The framework provides an instrument to help determine the target-language experts' qualifications and perspectives on linguistic issues.

Ching-Ni Hsieh
Cambridge Michigan Language Assessments

ESL teachers' versus American undergraduates' judgments of oral proficiency, accentedness, and comprehensibility

Second language (L2) oral performance assessment always involves raters' subjective judgments, and is thus subject to rater variability. The variability due to rater characteristics has important consequential impacts on decision making (Bachman, Lynch, & Mason, 1995; Brown, 1995; Myford & Wolfe, 2000; Schaefer, 2008). This research project examines rater variability and rater rating orientations associated with the characteristics of two groups of raters, (1) English-as-a-Second-Language (ESL) teachers and (2) American undergraduate students, in a high-stakes testing situation—the screening of international teaching assistants (ITAs) at a large Midwestern university. Thirteen ESL teachers and 32 linguistically naïve undergraduate students were recruited to evaluate 28 potential ITAs' oral responses to an in-house oral proficiency test. Raters evaluated the examinees' oral proficiency, degree of foreign accent, and perceived comprehensibility using separate holistic rating scales. Raters also provided concurrent written comments regarding the factors that drew their attention to while they evaluated the examinees' oral proficiency. A FACETS analysis was employed to examine and compare ratings awarded by the two groups of raters. The written comments were analyzed using qualitative content analysis and six major categories tapping into raters' rating orientations were identified. Results of the study indicate that the ESL teachers and the undergraduate raters did not differ in severity they exercised when evaluating the examinees' oral proficiency. However, statistically significant results were found on the comparisons for the ratings of accentedness and comprehensibility. The undergraduate raters as a whole were harsher than the ESL teachers when they evaluated the examinees' foreign accent and the difficulty they experienced listening to the examinees' accented speech. The written comments suggest that most of the raters attended to the pronunciation and fluency aspects of the examinees' speech. Results of the study is hoped to shed light on the research in ITAs testing and training.

Rie Koizumi
Tokiwa University

Akiyo Hirai
University of Tsukuba

Comparing the Story Retelling Speaking Test with other types of speaking tests

Speaking test formats have been shown to differently affect test takers' performance and resulting test scores but to have some shared aspects (e.g., O'Loughlin, 2001). This study investigates such shared and varied aspects of three speaking tests. The Story Retelling Speaking Test (SRST), the focus of our study, requires test takers to read English texts silently and to retell them and add their opinions in English without looking at the original texts. Previous studies have shown some aspects of reliability, validity, and practicality of the SRST and its usefulness in eliciting extended monologues from low- and intermediate-level learners. The other speaking tests examined are the Standard Speaking Test (SST; ALC Press, 2010) and the Versant™ (Pearson Education, 2010). The SST is a modified version of the Oral Proficiency Interview, whereas the Versant is conducted on the telephone or computer and used to assess spoken language skills. We administered the three tests and a questionnaire to 64 undergraduate and graduate students in Japan and examined (a) to what extent the SRST scores are explained by the other two test scores, using multiple regression analysis, and (b) what test takers think each test measures, using results from the questionnaire. The results showed that a large proportion (46%) of the SRST scores was predicted by the other two test scores and that the SST by itself could explain the score variance almost as much as the Versant could (43% vs. 41%). This suggests that the SRST can measure speaking abilities similar to those measured by the SST and the Versant. Furthermore, test takers' responses from the questionnaire highlighted both perceived similarities and differences, indicating that while acknowledging that the three tests share speaking ability, test takers recognize differential aspects of ability involved. Pedagogical implications are presented.

Ana Lado
Marymount University

Margaret Lado-Schepis
Lado International College

Lucy Lado-Dzkowski
Central Florida University

Robert Lado's principles of testing applied to test revisions used at Lado International College

Teaching and testing are closely related and it is difficult to work in one area without impacting the other. Recent research in communicative competence has lead to beneficial changes in language programs by placing conscious attention to the authenticity of language context. However, while the merits of communicative teaching are supported by empirical evidence, subsequent changes in the area of language testing have not always been as empirically grounded. The enthusiasm behind a new teaching paradigm along with market forces, and accreditation mandates have worked in favor of the use of tests which might not adhere to the basic principles of testing. This paper examines recently emerging concepts in language teaching that when applied to testing have yielded flawed tests . For example, too much emphasis can be placed on authentic language context to the detriment of the use of basic principles of language testing. Specifically, tests administered at Lado International College will be used to illustrate the conflicts and potential pitfalls that arise from efforts by good teachers to include more authenticity. Examples will be shown from the English placement tests written decades ago by Dr. Robert Lado and more recently revised versions. Anecdotes will illustrate what was learned in the revision process. Memorable video footage will be shown of Dr. Lado teaching and training over the decades.

Zhi Li
Iowa State University

When video meets MCQ: A qualitative study of video listening comprehension test

In both listening teaching and testing, the discussion of video use has been nearly settled with a general agreement that videos are helpful in listening class and videos should be incorporated into listening tests as visual input to mirror the common practice in listening teaching. However, for any language tests, using visuals or videos in listening test is not an easy decision mainly because of the concerns of construct validity. Different from quantitative studies which focus on statistical significance in comparing different input modes, qualitative study on test-takers' comprehension can help better understand on the effect of visuals in listening comprehension test. Following an interactionist perspective, this study treats listening comprehension as a process of learner's interaction with both input and listening task. It aims to explore the learners' interaction with video input and multiple-choice questions in an English listening placement test at Midwestern university in the U.S. Both quantitative and qualitative measures are used to explore the effect of video in a multiple choice question listening comprehension test. An IRT item analysis of the existing data is conducted to identify item features, along with a content analysis of video and the questions. Based on the information, a coding system is established for the follow-up qualitative analysis. 6 ESL learners with different proficiency levels are invited to take a computer-delivered version of listening comprehension test and all the test-taking process and computer operations will be recorded by a web-camera and Camtasia respectively as stimuli. Then a stimulated recall experiment is conducted to explore learner's listening comprehension process. The qualitative analysis results indicate that many test-takers ignore the visual input and still rely on the audio channel and item reading skills. On the other hand, only few test-takers can take advantage of visual information and successfully use it in answering questions.

Chien-Yu Lin
University of Maryland, College Park

Validating a theory-based strategic process model for L2 reading comprehension via Bayes Nets

This study applied Bayes Theorem, a unique approach rarely used in assessing strategic competence, to validate a theory-based strategic model for L2 reading comprehension. First, a strategic process model for L2 reading was constructed by drawing upon a text comprehension model, Integrated and Constructed Model (Kintsch, 1998; van Dijk & Kintsch, 1983) and metacognitive theories (Baker & Brown, 1984; Mokhtari & Reichard, 2004).Then, the component of language proficiency was integrated into two Bayes nets by specifying different conditional probabilities of reading comprehension levels for proficient L2 readers and less proficient L2 learners in terms of readers' sufficient or superficial use of textbase strategies, situation model strategies and metacognitive strategies. A data set obtained from 3 proficient EFL readers and 3 less proficient EFL readers using think-aloud protocols was fed into the two Bayes nets. The study examined how these readers' strategy use patterns led into different reading comprehension levels as predicted by the Bayes networks. Further implications about how to use Bayes nets to assess integrated language skills in a task-based context were also discussed.

John Michael Linacre
Winsteps.com

Current developments in the "Facets" software for Rasch Analysis

The "Facets" software for Many-Facets Rasch Measurement (MFRM) was first implemented in 1986. In the ensuing 25 years, Facets has been considerably enhanced. Current developments are illustrated, including improvements to data handling, analysis and the user interface. The capabilities of Facets are also compared with those of two-facet (person-item) Rasch software, such as Winsteps.

Megan Montee
Georgia State University

Sara Cushing Weigle
Georgia State University

Raters' perceptions of textual borrowing in integrated writing tasks

Integrated assessment tasks are intended to more closely reflect language use in real-world academic settings than tasks that measure only one skill. In the case of integrated reading and writing tasks, test takers must produce source-based writing. While integrated tasks offer important benefits in terms of task authenticity, their use of these tasks also raises issues about how test takers incorporate source texts into their writing. Textual borrowing refers to the direct use of language from the source text. While previous research (Shi, 2004; Weigle, n.d.) has looked at patterns of textual borrowing in writing assessment, no research to date has examined how test raters perceive source-based writing and how textual borrowing may affect their ratings. Textual borrowing may lead to inaccurate ratings by masking students' writing proficiency (Weigle, 2002). In addition, issues of what raters perceive as appropriate and inappropriate source-based writing, and the extent to which these perceptions reflect the expectations of real-world academic writing, are essential to task authenticity. This poster presents the results of a study to explore how test raters identify, perceive and make scoring decisions about textual borrowing. The context of the study is a locally developed writing exam used for placement in ESL courses. The study included three stages. First, exploratory focus groups (n=2) were conducted to gather data about raters' general perceptions of textual borrowing in student essays. Next, stimulated recalls (n=12) were conducted in which participants scored and discussed essays with textual borrowing. Finally, a follow-up focus group was conducted to revise the test rubric to address textual borrowing. This poster presents selected results from the study focusing on implications for rubric design and rater training.

Margaret van Naerssen
Immaculata University

Faking a lower language proficiency: How easy is it?

An experimental protocol was developed for determining the ease/difficulty of faking a lower than truthful language proficiency in legal cases involving nonnative speakers (NNSs) of English. The focus of this presentation is to report on a replication of this protocol with 18 subjects, including two "fakers." In many cases claims by NNSs about their problems in understanding/speaking are truthful. However, some may try faking low second language (L2) proficiency levels for legal advantages. Detecting possible faking is challenging, especially when there is limited or no audio evidence. To address this, an experimental protocol was developed and applied in a federal case using an alternating language story retell task along with an oral proficiency assessment. This was also replicated with two other subjects with different language proficiency levels. Initial results suggested the degree of subconscious "leakage" of content across languages might be possible evidence of a general listening proficiency level. However, as this data only came from a single criminal case, a research study was designed to replicate the protocol. The protocol was replicated with 18 subjects outside of the legal system, with a few changes to improve the procedures. The alternating language retell task (English and Chinese-Taiwanese) was paired with an adult basic English language proficiency test, BEST-Plus. The general goal was to determine if the protocol could be an appropriate and practical tool for wider use. The subjects responded using their truthful language proficiency. However, also of interest was what their performance would look like if they were asked to intentionally attempt to fake a lower proficiency? Thus, two subjects were asked to role-play an arrested drug dealer, trying to fake a lower language proficiency. Afterwards their truthful language proficiency was also tested, and their language performance across the two samples were compared for consistencies and inconsistencies.

Nina Padden
Concordia University

Heike Neumann
Concordia University

Developing a comprehensive academic English placement test

This poster describes the development process for a new postadmission placement test for a university English as a second language (ESL) program. Students will take this placement test after they have obtained the required score on a standardized English proficiency test and have already been admitted for university study. The revision of the ESL curriculum at the institution called for a new placement test that was closely tied to the new curriculum and the course objectives. This circumstance also provided an opportunity to replace an outdated placement and proficiency test. The purpose of the new test is to place incoming students into one of two sequential academic ESL courses (with an emphasis on reading, writing, grammar, and vocabulary development) and one of two oral communication courses. This purpose determined that the placement test had to assess test takers' reading, writing, listening, and speaking ability as well as their level of grammar and vocabulary knowledge. The poster will outline the development process of this test from its conceptualization in relation to the curriculum and the selection of task types to item writing and the piloting of test items. In keeping with the LTRC 2011 conference theme, this poster will provide an opportunity to compare Lado's (1961) view on testing "language elements" and "integrated language skills" with current assessment practice. By contrasting the old and new placement tests, one can observe, on the one hand, the move from decontextualized to contextualized assessment of "language elements". On the other hand, integration is now not only evident in the assessment of "integrated language skills" but also in the links between different parts of the test, such as reading-to-write tasks. In this sense, the poster provides a window into "developments in language testing in the last fifty years" (LTRC call for papers).

Elizabeth Park
Educational Testing Service

Timothy Pelletreau
Educational Testing Service

Fairness and the role of culture in assessing a global test-taking population

This poster presents issues in the design and delivery of language assessments for large-scale, international populations. The validity of a language test that is administered globally hinges upon its fairness and its appropriateness for test takers in the countries in which the test will be utilized. As what is considered appropriate varies from country to country, it is important to consider the culture of the test takers throughout the process of test development, from the early design stage to the reporting of scores to stakeholders. Included in this poster are strategies for developing fairness guidelines, considerations for minimizing cognitive, affective, and physical sources of construct-irrelevant variance, and approaches for standardizing and implementing fairness reviews.

Ève Ryan
Avant Assessment

It takes two to tango: Examining the collaboration between test developers and language informants

Globalization, increased immigration, and humanitarian and military intervention abroad have all led to an ever-increasing demand for language tests, especially in less commonly taught languages. However, this situation does not automatically ensure that a workforce of test developers is locally accessible or adequately trained, and in many instances, the test will be developed by a team who does not speak the language. As such, test development professionals rely on native speakers of the target language for their linguistic and sociocultural expertise. The purpose of this poster is to describe the process of developing reading tests when the tests developers do not speak the target language, with a particular focus on the partnership between test development experts (TDEs) and target language experts (TLEs). Four TDEs and three TLEs participated in this study. Study participants were interviewed using a standardized protocol drawn from a comprehensive review of the literature. Interview transcripts were analyzed using computer-assisted methods. Results of this study shed light on some of the linguistic and cultural challenges encountered during this collaboration and provide details on the "checks and balances" of this process. It was found that cross-cultural communication is a central focus of the test development process, with language informants stressing the importance of cultural competency in their test development colleagues. TDEs reciprocally valued TLEs with strong intra- and inter-cultural knowledge, metalinguistic awareness, pragmatic knowledge and an ability to explain these differences clearly in English. This research provides a new perspective on this largely unexplored area of language testing and has implications not only for researchers but also for practitioners who build or use tests in languages they do not speak, especially in lesser-known languages. Given the high stakes of such tests, an understanding of the process of test development through collaboration between test developers and language informants is critical.

David Ryan
Universidad Veracruzana

Pushing the envelope, moving towards democratic assessment in a non-democratic context: A case-study in Mexico

In 2004, Cumming stated that more research is needed on the role of stakeholders in language testing contexts that have traditionally been overlooked. One such area is Latin America. The present poster helps fill this gap by examining the context of a Mexican university where initial efforts are underway to advance a more democratic assessment agenda within a society that is, in many ways, still struggling to come to terms with certain democratic ideals. A questionnaire focusing on candidate perceptions of an English language proficiency test was sent to 964 students, of which 245 responded, and of which a further 99 agreed to participate in follow-up interviews. Extreme case sampling was then used to select four candidates with opposing views on an attitudinal scale ranging from "overwhelmingly positive feelings about the test" to "overwhelmingly negative feelings." The interview protocol, drawn from a literature review, focused on: (1) a further, and deeper, exploration of test takers' perceptions and feelings about the test; (2) test takers' beliefs and opinions about education and assessment in Mexico. Participants' responses were then analyzed based on Hofstede, Hofstede and Minkov's (2010) cultural framework of power distance. The findings of this study suggest that Mexican test takers have mixed feelings about democratic notions of assessment, such as those espoused by Lynch (2001) and Shohamy (2001a; 2001b). Suggestions are made to reach a more balanced distribution of power between the testing institution and test takers, while also considering the role of Mexican societal and cultural values in assessment. This research provides a new perspective on a relatively unexplored area of language testing and has implications for assessment practitioners and stakeholders who welcome more democratic forms of assessment. Given the undeniable influence of culture and values in language tests, efforts to examine the "assessment culture" of under-represented geographical areas are vital.

Sara Smith
University of Oxford

Developing receptive and productive tests of verb + noun collocational knowledge for use with young learners

Assessment of vocabulary in reading comprehension has tended to emphasize counting the number of words in a text an individual knows, or vocabulary breadth (Pearson, Hiebert, & Kamil, 2007). However, in order to process these lexical items effectively the learner must also have knowledge of grammatical functions, register, appropriate usage, idioms and collocations, or vocabulary depth (Milton, 2009). One area of vocabulary depth of increasing interest is the role multi-word phrases, collocations, play in language acquisition and processing. For the purposes of the current study, a 'collocation' will be defined as a composite unit of words that expresses meaning as a whole and whose components co-occur more often than would be expected by chance (e.g. take vs take place). While it has been long established through corpus linguistic evidence that such multi-word items are ubiquitous in naturally-occurring discourse (Wray, 2002; Sinclair, 1991), and that such items can negatively affect the reading comprehension of adult L2 learners of English (Martinez & Murphy, in press), at present the nature of collocational knowledge in young learners and its possible correlates with literacy remain under-explored, largely because there are no available measures appropriate for use with children. The current study details the creation and validation of a receptive and a productive measure of verb+noun collocation knowledge for children between 7 and 10. This poster reports the findings from the initial test administrations with 80 British EAL and L1 English children. Results show the tests yield reliable scores and show evidence of internal consistency and concurrent validity. The tests discriminate well between learners at different ages and between native and non-native English speakers. Scores correlate highly with other measures of vocabulary knowledge, indicating a relationship between vocabulary size and collocation knowledge. This paper discusses the study's limitations and offers an outlook on possible future research.

Alan Urmston
Hong Kong Polytechnic University

Michelle Raquel
Hong Kong University of Science and Technology

Jerry Gray
Lingnan University of Hong Kong

Carrie Tsang
Hong Kong Polytechnic University

Felicia Fang
Hong Kong Polytechnic University

Designing a valid and reliable diagnostic assessment and tracking system for tertiary level ESL learners

The eight universities in Hong Kong all have language centres which are charged with enhancing the English proficiency levels and academic English skills of students as they proceed with their programmes of study, which are in the main through the medium of English. Many of these language centres have self-access facilities which enable students to take the initiative to improve their language skills. However, what is lacking is a means whereby students can reliably diagnose their weaknesses in English and monitor their progress in learning it, thereby enabling them to spend their limited time productively. The Diagnostic English Language Tracking Assessment (DELTA) is a collaborative development project involving university language centres in Hong Kong. The DELTA system utilizes Item Response Theory to produce reliable and useful results and targeted assessments, as well as computer technology to facilitate testing operationalisation and student progress tracking. Students take tests of reading, listening, grammar and vocabulary under supervision in a language laboratory and their results are analyzed using Winsteps (Linacre, 2010). Winsteps analysis enables test items to be calibrated at distinct difficulty levels and from this proficiency levels (DELTA Measure), component skills profiles and item response reports are generated. The reports provide invaluable information for students, teachers and course planners. In addition, the DELTA system provides tests targeted to test takers' proficiency levels and tracks their progress towards a desired or (in some cases) mandated exit level of English proficiency. This poster reports on the design and development of the DELTA, highlighting the use of Winsteps for item calibration, item banking, test design and diagnostic reporting. The paper also looks at the design of the reporting and tracking system and explains how valuable a tool the DELTA will be for English learners and teachers in universities in Hong Kong and elsewhere.

Zunmin Wu
Beijing Normal University

Luxia Qi Guangdong
University of Foreign Trade and Foreign Languages

Developing an educational assessment system for China's basic education

With the curriculum reform in the basic education undergoing for about ten years, it is high time that China constructed a formal assessment of education quality that can provide information necessary both for important educational policy making and for curriculum development. The proposed presentation intends to report on an ongoing Chinese project for the development of an assessment system for the country's basic education. The National Assessment of Education Quality (NEAQ) project was officially launched in 2009. It involves the major school subjects including English language education. NAEQ has set Grade 4 (10-year-olds) and Grade 8 (14-year-olds) as target assessment groups, drawing on experiences of some of the world's educational assessment programs, such as NAEP, PISA and TIMMS. While the assessment system has encompassed both a language test and an investigation of factors affecting learning, the present presentation focuses on the construction of a pen-and-paper test in relation to the test result reporting issues. The most important issue has been the decision on the test construct. It has been generally agreed that the assessment should measure students' actual ability to do things with English rather than let them demonstrate simply what they remember. In that vein, integrative test formats that require the combined use of language skills are much favoured although research needs to be done in relation to the necessity of control of candidate responses and the degree of it.

Jing Xu
Iowa State University

Elena Cotos
Iowa State University

L2 speech prosodic features in a human-annotated ITA spoken corpus

One of the challenges that automated speech evaluation (ASE) faces is to identify quantifiable yet construct-relevant features in spontaneous second language (L2) speech (Xi, et al., 2008). Prosody, a construct addressed by numerous L2 pronunciation textbooks, has been bypassed by the developers of ASS systems due to the limitations of current speech recognition and processing technologies (Xi, 2010; Chapelle & Chung, 2010). It also remains an open question as to what specific prosodic features affect human raters' judgments of L2 oral proficiency. This poster reports on an L2 spoken corpus currently being developed at a U.S. university with a primary goal to identify and analyze problematic prosodic features in L2 speech in order to inform assessment and pedagogical practices. At present, the corpus contains 109 five-minute monologue speech samples produced by international teaching assistants (ITAs) in a simulated teaching task, which is part of an institutional oral proficiency test. These samples were randomly selected from three language groups (Chinese, Hindi, and Korean) and four proficiency levels representative of all score bands of the test. The corpus is transcribed by three trained research assistants and annotated for prosodic features based on a coding scheme adapted from Du Bois (1993, 1991). The coded prosodic features are: intonation unit, disfluency, terminal pitch direction, word stress, prominence, and non-meaningful pauses. Statistical procedures including correlation and ANOVA will be used to investigate the relationships between these prosodic features and the ITAs' holistic test scores given by human raters. The findings from this corpus study will provide helpful insights for the development of automated analysis of L2 speech prosody by identifying (1) prosodic patterns typical of Chinese, Hindi, and Korean speakers of English and (2) construct-relevant prosodic features to be included in the scoring model of ASE systems.

Martyn Clark
Center for Applied Second Language Studies

Open Collaborative Assessment Platform (Open CAP)

Open CAP (Open Collaborative Assessment Platform) is an attempt to develop an open source online tool which will allow the "crowd sourcing" of test item content. The goal is to eventually enable end users (teachers) to create and deliver custom-built assessments from a centralized bank of community developed items in a variety of foreign languages, especially less commonly taught languages. The primary educational mission of the Open CAP project is to increase assessment literacy and improve learning outcomes by allowing foreign language professionals to collaborate in the development of proficiency-based items for low stakes, classroom use. A key challenge for this project is determining how to incorporate accepted test development practices into a tool geared towards nonspecialists without the resources available to professional test development teams. Initial technical discussions for Open CAP include concepts of test and item specifications, item review, exemplar items and responses, and standard setting. It is hoped that discussion of Open CAP as a work in progress will provide useful suggestions and identify areas of concern while the project is in its formative stages.

Jee Wha Dakin

Examining gain in grammatical knowledge among adult learners with low English proficiency

In the context of adult learners with low English proficiency enrolled in an organization offering instruction in language and civics content, the purpose of the study is to examine in which ways examinees' responses have improved over the course of twelve weeks of instruction. The Grammar Test was operationalized using Purpura's (2004) model of grammatical knowledge. It was designed to measure grammatical form and grammatical meaning in three types of tasks: selected-response; limited-production; and extended production. Data from a sample of fifteen adult ESL learners with low English proficiency were collected using a nonexperimental, pretest-posttest design with intact classes. A post hoc analysis is being conducted on the extended-production items, examining linguistic patterns from examinees' responses. The results of the study could help identify where examinees may have problems in their ability to use grammatical forms to express their knowledge of civics content. Such results can provide important achievement details to the teachers, students, and the schools of the target-culture. It could also help inform formative or diagnostic decisions about students' next steps in closing certain learning gaps related to the simultaneous learning of language and content. The present study might also have important pedagogical implications for other schools with similar populations.

Larry Davis
University of Hawai'i, Mānoa

Rater cognition and behavior in speaking assessment

Performance tests typically employ raters to produce scores; accordingly, an understanding of how and why raters make particular scoring decisions is necessary to fully capture what test scores mean. In writing assessment, studies of rater decision-making employing verbal report techniques have been used to illuminate rater variability (e.g., Sakyi, 2000), examine the effects of training and experience (e.g., Lim, 2009; Weigle, 1994), and clarify rater scoring processes and criteria (e.g., Lumley, 2005; Milanovic, Saville, & Shen, 1996). In contrast, few studies have examined rater decision-making within speaking tests. Moreover, the relevance of findings from writing assessment to speaking is unclear given that the two modalities may place different cognitive demands on the rater. This study aims to describe the cognition and behavior of raters within the context of a speaking performance test, and to illuminate how decision-making changes with training and experience. Experienced teachers of English will score recorded TOEFL iBT speaking test responses prior to training and in four sessions following training (90 responses for each iteration). For a subset of judgments, raters will verbally report (via stimulated recall) what they were thinking as they listened to the response and made the scoring decision. Analysis of stimulated recall data will be used to describe decision-making processes and scoring criteria across time as raters develop increasing familiarity with the scoring context and procedures. Scores will be analyzed using Rasch analysis, with scoring behaviors including consistency, severity, and use of the rating scale compared across dates. Findings from the qualitative analyses will also be compared with scores to determine whether specific types of decision-making are associated with particular scoring patterns. Finally, the project will consider the cognitive and behavioral nature of rater expertise, and the contribution of decision-making behaviors to scoring phenomena such as reliability and severity.

Candace Farris
McGill University

Immanuel Barshi
NASA Ames Research Center

The assessment of communicative effectiveness in non-routine aviation communications: A socio-cognitive perspective

This work in progress is phase one of a research project investigating the nature of effective communications between air traffic controllers and pilots. The International Civil Aviation Organization (ICAO) has introduced language proficiency requirements (LPRs), due to take effect in March 2011. The goal of the ICAO LPRs is to ensure that pilots and air traffic controllers can communicate effectively in nonroutine or emergency aviation situations. These requirements were drafted after a number of aviation incident/accident reports revealed miscommunications to be a factor in the mishap. In some situations, one or both interlocutors were nonnative speakers of English, and it appeared that English language proficiency was a factor in the miscommunication. ICAO therefore introduced LPRs, which stipulate that pilots and air traffic controllers must demonstrate general proficiency in the language of radiotelephonic communications—often English. According to the criteria of the rating scale provided by ICAO, most native speakers of English will achieve an Expert Level of proficiency. We question the assumption that high proficiency in English necessarily implies communicative effectiveness in nonroutine (emergency) aviation situations. Furthermore, we question ICAO's definition of proficiency in relation to the goal of communicative effectiveness for pilots and air traffic controllers. From a sociocognitive perspective that emphasizes the co-constructed nature of communication, we apply conversation analysis to incident/accident transcripts to identify elements key to effective communications. We illustrate the complexity of successful and unsuccessful nonroutine aviation communications. We demonstrate that while language proficiency is a factor in successful communications, factors that extend beyond mere proficiency are also important. These factors apply to native and nonnative speakers alike. Finally, in accordance with the conference theme, we suggest assessment criteria that are aligned with the elements of successful communications identified in our analysis. We are seeking feedback on our methodology and critique of our inferences.

Eun Hee Jeon
The University of North Carolina at Pembroke

Tran Phuong
University of Melbourne

Assessment of L2 reading-related metacognition

Assessment of metacognitive awareness and strategy use in L2 reading research is a challenging and underinvestigated area. Currently, assessment options of metacognition are largely limited to questionnaires and self-report surveys inquiring about the respondent's awareness or knowledge of reading strategies, strategy use, and text characteristics (e.g., Jeon, forthcoming; Schoonen et al., 1998; Shiotsu, 2010; Van Gelderen et al., 2007). However, as noted by Grabe (2009), the problem with such assessment practice is that while questionnaires and self-surveys do capture reader's metacognitive awareness of reading, they fall short of measuring the actual use of metacognitive knowledge and strategies in reading (i.e., how successfully the reader applies his metacognitive knowledge to aid reading comprehension). In other words, the current assessment tools of metacognition may have potential problems with construct validity. There is therefore, a clear need for developing valid measurement tools of metacognition.Acknowledging these limitations, the proposed study adopts two different types of assessment tools of metacognition: a test of metacognitive knowledge and reading strategies, and a discourse-level comprehension monitoring and predictive ability test, the latter of which was newly developed for this study. It is hypothesized that each of these two tests will tap into different subconstructs of reading-related metacognition. It order to investigate this hypothesis, a large-scale data collection involving approximately 90 university level L1 Vietnamese-L2 English learners was conducted in January, 2011. The study also aims to examine the comprehensive and separate contributions of the two subconstructs of metacognition to overall L2 reading outcomes. To this end, study participants were also assessed on 8 additional key reading variables (i.e., pseudoword decoding, word recognition, morphological knowledge, vocabulary size, online syntactic processing efficiency, grammar knowledge, listening comprehension, motivation for reading) as well as L2 reading comprehension. Study data will be analyzed using Factor Analysis and Multiple Regression Analysis.

Sahbi Hidri

Testing listening comprehension: Towards an integration of assessment for learning and assessment of learning

Tests have a powerful effect on the life of people and institutions and they influence pedagogy in fundamental ways. Therefore, test designers should ethically adopt a scrutinizing attitude to design tests in a careful and informed way. This paper addressed both types of assessment: assessment of learning (AoL) and assessment for learning (AfL) of listening comprehension (LC) for learners of English in an EFL context at the university level. The study gathered quantitative and qualitative data: a progress dynamic LC test for 130 test takers where two test takers worked on the test in pairs and with one test interviewer. Also, the same test takers were administered an achievement static LC test at the end of the course to be worked on individually. Two qualitative instruments were administered, each with one test: a think-aloud protocol while working on the progress dynamic LC test and a retrospective interview with the static achievement LC test. Four raters, who were also administered a retrospective interview, scored the two tests. The scores were analyzed using the FACETS program. Results of the study showed that AoL and AfL of LC were both germane to having a comprehensive view about the test takers' language ability. In addition, dynamic assessment proved to be an effective method of teaching and testing. Recommendations were made to consider both types of assessment to develop the learners' autonomy in a classroom assessment environment.

Becky Huang
Educational Testing Service

The influence of language experience and professional background on raters' judgment of oral language proficiency

Human judgment is widely used in the assessment of second language (L2) speech. Given that rater variability directly relates to the validity of the assessment, researchers have examined potential sources of rater variation, such as raters' native language background (Kim, 2009) and effects of training (Lumley & McNamara, 1995; Weigle, 1994; Xi & Mollaun, 2009). The current study aims to expand on this line of research by investigating two rater variables that have been relatively underexplored: familiarity with the speakers' first language (L1), and professional experience in teaching English as a second/foreign language (ESL/EFL). The study will utilize an operational dataset from the TOEFL iBT speaking test, which includes 38 Chinese L1 test takers with varying English oral proficiency levels. The study will utilize a quasi-experimental design with four groups of native English-speaking raters (n = 80 raters) varying in their experience with the speakers' L1 (with/without familiarity) and their experience in ESL/EFL teaching (with/without teaching experience). All raters will be students or staff affiliated with universities, and have similar age range and educational level. A quarter of the raters in each of the four groups (n= 20 total) will also provide retrospective verbal reports of their decision-making process. The L2 dimensions to be examined include a holistic rating of general oral proficiency as well as analytic ratings of accentedness, grammar and vocabulary use, and content. Both quantitative and qualitative analyses will be employed to answer research questions about the influence of the two rater background variables on raters' rating behavior. The findings will have implications for the selection of raters for oral language assessments, and the natural biases revealed by naïve raters in this study can also inform the rater training procedure.

Glyn Jones
Pearson

Ying Zheng
Pearson

Accumulating evidence of test validity: An ongoing study

The work in progress to be presented is an ongoing study in which IRT is being used to align two high stakes tests. The tests in question were developed for different but overlapping target groups: one—the Pearson Test of English Academic (PTE Academic)—is a test of English for Special Purposes targeting a restricted range of proficiency; the other—the Pearson Test of English General—(PTE General) is a General English test offered at a range of levels. While the immediate aim of the study is to link the tests to each other (so that both can draw on the same item bank and both report to the same scale), this aim serves in turn as a means to broader objectives to do with building arguments for validity, in particular:

- To establish a concordance table between these tests and other high stakes tests (concurrent validity)
- To align both test to the Common European Framework (CEF) (criterion related validity)

The methodology adopted for the study is fundamentally that of field testing with linking items. However, due to practical constraints and the need to link to external criteria, the data gathering proceeds by a complex series separate trials in which

- Candidates taking a live administration of PTE General take a trial test consisting of items written for PTE Academic
- Candidates with recent scores on other high stakes tests take a trial test containing items written for PTE Academic or PTE General, or both
- Candidate responses obtained in the above trials are rated independently according to CEF scales.

The presenters will outline the practical challenges involved in managing this project from the point of view of the logistics of data gathering and the complexity of analyzing data from different sources, and will invite comments on theoretical issues concerned with aligning different tests to common standards.

Yen-Fen Liao
National Taiwan University

Investigating the relationships between grammatical knowledge and L2 comprehension across proficiency levels

Grammatical knowledge has been widely identified as playing a critical role in comprehending second language (L2) input. A recent review of the language testing literature reveals an urgent need for empirical studies on the role of grammatical knowledge in L2 reading and listening comprehension across different proficiency levels. It seems intuitive to hypothesize that high- and low-ability L2 learners differ in their mental processing (Rost, 2002). Low-ability learners might tend to rely more on bottom-up processing than on top-down processing (Alderson, 2000). Much is still unknown as to whether the relations of grammatical knowledge to L2 reading and listening abilities may vary across ability levels. The current study thus attempts to investigate how and to what degree the dimensions of grammatical knowledge are related to L2 reading and listening comprehension across different proficiency levels in the context of the General English Proficiency Test (GEPT), a high-stakes test commonly used in Taiwan for selecting prospective students or screening job applicants. One form of the high-intermediate GEPT reading and listening tests was administered to 260 college students in Taiwan. A multi-group structural equation modeling (SEM) approach was adopted to address the issue of interest. The discussion first centers on the factorial structures of the GEPT reading and listening tests, and then turns to the role of grammatical knowledge in the prediction of L2 reading and listening test performance across high- and low-ability test takers.

Yu-Chen Tina Lin
Indiana University, Bloomington

Exploring integrated reading-writing constructs through SEM: Relationships among L2 proficiency, strategy use, and summary performance

Integrated writing is a new focus of EFL standardized tests because using integrated skills is similar to authentic tasks of the target language use. However, the nature of reading-writing activities and the complexities of cognition, metacognition, schemata, L2 acquisition, and other aspects of summary writing become the overwhelming cognitive load on ESL/EFL students (Kirkland & Saunders, 1991). Since standardized tests usually report holistic scores of integrated tasks, it is hard for students/teachers to identify critical aspects requiring further improvement when reading and writing in EFL countries are usually regarded as two separate skills and instructed by different teachers. This study aims to explore underlying constructs of the summarizing task and relationships among students' L2 reading and writing proficiency, cognitive operations, and performance on summarizing in L2. Research questions are: 1) what are the relationships among students' general reading comprehension, writing ability, and summary performance? 2) Based on the same text, how do students perform on two tasks: summary and reading comprehension? 3) How do students' cognitive operations relate to their summary writing? 4) To what degree do L2 reading and writing proficiency and cognitive operations contribute to or interact with the quality of summary writing? Participants are 200 EFL students taking reading/writing courses at three universities in Taiwan. Every class meets two times in a computer lab for taking tasks online. The first time focuses on the TOEFL independent writing test, the TOEFL reading test, and surveying participants' background information. The second time includes text summarizing tasks, text reading comprehension tests, and a questionnaire inquiring participants' perceived difficulty about cognitive operations for summarizing tasks. Structural Equation Modeling (SEM) is used to estimate relationships among variables mentioned in the research questions. The expected results will attempt to provide implications for second language writing assessment and instruction, teaching material development, and theory in second language academic writing.

Liandi Liu
Macquarie University

Interactional features of Chinese EFL learners' discourse in a paired speaking test

As conversational strategies such as the ability to initiate, respond and negotiate meaning in situated performances have become a focus of L2 oral proficiency assessment during the last decade, there has been a growing interest among language testing researchers in analyzing candidates' spoken production in paired and group orals. The peer-peer testing formats possess a number of advantages over the individual-centered model, some of which Lado (1961:247) even drew attention to a half-century ago. Although there has been a sizable body of research studies exploring the nature of test talk elicited by these two formats (e.g., Davies, 2009; Galaczi, 2004, 2010; Gan, 2008, 2010; Gan, Davidson & Hamp-Lyons, 2009; He & Dai; 2006; Lazaraton & Davies, 2008; May, 2000, 2007, 2009; Luk, 2010; Nakatsuhara, 2006, 2009, 2010, Van Moere, 2007, 2010, yet published studies on oral test-taker interactive discourse by Chinese EFL learners are rare. The present study, drawing on the notion of interactional competence (Ducasse & Brown, 2009; Hall, 1999; He & Young, 1998; Kramsch, 1986; McNamara, 1996, 1997; Young, 2000) and using conversation analysis techniques, investigates interactional features of dyadic conversation by Chinese EFL learners in the Public English Testing System (Level 5) Spoken English Test (PETS-5-SET) through comparative analysis. Data were collected by a mock paired speaking test conducted by both Chinese (N=60) and Australian (N=30) university students. Fifteen transcripts of audio-recorded discourse performances from each group are analyzed and compared, focusing on the generic structure, topic development, and turn-taking behaviors. The comparative analysis shows both marked similarities and substantive differences between the two groups in the interaction. The findings will have implications for a better understanding of the construct of interactional competence underlying candidate-candidate paired test format in the PET-5-SET, for empirically based rating scale construction for the PETS-5-SET, and for the PET-5-SET examiner training, among others.

Jia Ma
Queen's University,
Faculty of Education

Liying Cheng
Queen's University,
Faculty of Education

Test preparation effects on Chinese test-takers' TOEFL iBT test performance and English language proficiency

By 2008, one-third of nearly 1.4 million overseas Chinese students studying in Untied States and Canada ("Overseas Chinese students", 2009) for better education and future career (Maslen, 2007). TOEFL is the gatekeeper for these students' university entrance. In 2007, China had the second largest number of students in the world taking the TOEFL (Qian, 2009). Because of the importance of TOEFL score in offering admissions, Chinese students willingly spend considerable amount of money on test preparation for TOEFL (Hamp-Lyons, 1998). Although many have achieved high TOEFL scores and gained university acceptance, English language proficiency remains a major challenge these students encounter in English-speaking academic environment (Campbell & Li, 2008). The limited research conducted on test preparation for SAT, TOEIC and IELTS produced mixed conclusions on test takers' performance (Gan, 2009; Powers, 1993; Robb & Ercanbrack, 1999). The studies specifically addressing TOEFL preparation (Alderson & Hamp-Lyons, 1996; Hamp-Lyons, 1998; Wall & Horák, 2008) had not investigated the effects of TOEFL preparation on test takers' test scores and their English proficiency. Therefore, this study empirically examines a potential causal relationship between test preparation effects on Chinese test takers' TOEFL iBT test performance and their English language proficiency. A nonequivalent groups pretest and posttest control/comparison design will be adopted. Students (n=200) from one Chinese university who are taking a TOEFL iBT preparation course will be the experimental group and those (n=200) from the same university who are not will be the control group. A reading test will be used as a control measure. TOEFL iBT and College English Test (indicator of English language proficiency) will be administrated before and after the 16-week test preparation to test statistical group differences on test preparation effect. The findings will provide empirical evidence on the effect of test preparation on Chinese test takers' TOEFL iBT performance and their English language proficiency.

Philippa Kim
Borough of Manhattan Community College, City University of New York

Tom Means
Borough of Manhattan Community College, City University of New York

Online reading comprehension tests for elementary Spanish at an urban community college

As part of a larger college-wide focus on assessment, we created an online reading comprehension test for all 50 sections of Spanish 101 in our department, Modern Languages. This paper will focus on the design and field testing of the test (content of the reading passage: should it be literary, functional or informational?) and types of comprehension questions to be asked: (a mix of explicit, interpretive and lexical/idiomatic), and the 15 pilot tests (approx 300 students) that ran one semester prior to the actual assessment. Mention will also be made of the online survey services we tested to find a good fit for our institution. All 50 sections (approx 1000 students) took the test in Fall 2010 and all students knew that the test would count toward their final exam grade (approx 5%). We are currently in the process of sorting and analyzing the data. Initial interpretations and logistical lessons learned will be shared.

Rebecca Present-Thomas
Vrije Universiteit Amsterdam

Academic English writing assessment and the Common European Framework

An increasingly popular model for describing levels of language proficiency, the Common European Framework of Reference for Languages (CEF; Council of Europe, 2001) classifies learners into one of six levels (A1 to C2) and describes their skills in terms of can-do statements. However, for writing skills in particular, the CEF descriptors are far from comprehensive. Though many widely used tests of language proficiency currently make claims about how scores on their tests are linked to particular CEF levels, the integrative tasks and scoring methods favored by these tests do little to clarify the linguistic reality underlying the abstract levels. In this project, written skills of advanced learners of English will be analyzed in terms of linguistic characteristics. Advanced English student responses to items assessing writing skills will form the basic material for the research. Item responses will vary according to length and timing restrictions and each response text will be classified according to the CEF levels using a combination of independent scoring methods. Results will shed light on the specific elements that distinguish an advanced academic (CEF C1/C2) English writer's linguistic repertoire. Additionally, the appropriateness of limited-time and limited-scope tasks in measuring advanced writing proficiency will be evaluated. This paper will discuss the findings from the first round of data collection, completed in 2010. The dataset includes mainly responses representative of CEF levels B1 and B2 and will serve as a baseline for comparison with C1/C2 data to be collected in subsequent periods of data collection between 2011 and 2013.

Steven Ross
Center for Advanced Study of Language

Megan Masters
Center for Advanced Study of Language

Margaret E. Malone
Center for Applied Linguistics

Katherine Riestenberg
Center for Applied Linguistics

Assessing learning outcomes in short-term language programs: Issues and challenges

As the language testing community has developed definitions and means of measuring language proficiency over the last half-century, researchers have investigated the validity of student and teacher reports of student proficiency (Glisan & Folz, 1998; Malabonga, Kenyon, & Carpenter, 2005). This presentation describes effective uses of teacher- and student-based instruments for reporting proficiency as compared with results from a computerized proficiency test. The presentation is contextualized in the challenge of assessing student progress in short-term, less commonly taught language programs. To identify appropriate assessments for such programs, a study (N=396) investigated the efficacy of three assessment instruments administered to high school students in short-term Arabic and Chinese programs: a learner self-assessment, a teacher retrospective assessment, and a computerized proficiency assessment. Additional program observations were conducted for qualitative analysis. Data from the three instruments were triangulated to examine evidence of convergent validity and comparative reliability. Overall, the teacher and student-based instruments correlated with the external proficiency test; however, 8 of the 15 programs showed misfitting patterns. Characteristics of misfitting programs are being explored to inform future implementation of these assessment instruments in short-term language programs. Researchers have suggested that training for teachers and students on the assessment tools, including purposes of self-assessment and the nature of "can-do" proficiency statements, would increase the validity of these instruments. Discussion of this work in progress will focus on reasons for alignment and misalignment of pilot study results, and potential use, including training, of teacher- and student-based proficiency reporting instruments in short-term language programs. Glisan, E. W., & Foltz, D. A. (1998). Assessing students' oral proficiency in an outcome-based curriculum: student performance and teacher intuitions. The Modern Language Journal, 82, 1–18. Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. Language Testing, 22(1), 59–92.

Nehal Sadek
Indiana University of Pennsylvania (IUP)

A qualitative study exploring the impact of a hybrid dynamic assessment model on essay writing by English language learners (ELL)

The present study explores the impact of a proposed Hybrid Dynamic Assessment (HDA) model on the essay writing of English Language Learners (ELL). The proposed model combines characteristics of both currently used interventionist and interactionist DA models rooted in Vygotsky's Sociocultural Theory (SCT) concept of Zone of Proximal Development (ZPD). The participants of the study include five ELL students and two writing teachers within the Intensive English Program (IEP) at the American University in Cairo (AUC), Egypt. Using pre- and post-tests, observations, and interviews, the study explores the impact of the proposed assessment model on students' writing. The study also investigates students' and teachers' evaluation of the proposed HDA model.

Jamie Schissel
University of Pennsylvania

Aubrey Logan-Terry
Georgetown University

Test accommodations for immigrant students: Wolves in sheep's clothing?

Immigrant, emergent bi/multilingual students in U.S. public schools are required to take annual standardized assessments with federal and state mandated accommodations. However, the effectiveness of test accommodations in reducing construct-irrelevant variance due to language proficiency is, at best, uncertain and, at worst, unsuccessful (e.g., Kieffer et al., 2009). Despite these issues, test accommodations continue to be used, but to what effect? Do they help or hurt immigrant students? To investigate these questions, we analyze accommodation use with emergent bi/multilingual students (n = 60) from six content classrooms across grades three through eight in three Northeastern U.S. schools. Qualitative and quantitative analyses of the test data, classroom ethnographies, students' workbooks, and cognitive laboratory interviews indicate that popular test accommodations (e.g., test translation and linguistic simplification) are incongruent with students' multilingual processing of content knowledge. Whereas accommodated tests frame content and language learning as static, our findings suggest a continuum of dynamic multilingualism for immigrant students. For example, in-depth case studies of individual students illustrate their fluid navigation of content through multiple languages and multiple modalities. This multilingual processing is in contrast to the monolingual standard inherent in the design of accommodated tests. We therefore argue that test accommodations are potentially harmful for immigrant students because they neither address the dynamic multilingualism of these students, nor the larger issues surrounding problematic monolingual biases in the design of high-stakes, standardized assessments. To account for these limitations, we conclude that there is a need to re-examine the current paradigm of research on test accommodations that perpetuates these harmful practices. Additionally, language testers need to discuss what ideological shifts are necessary in order to develop instruments that address the skills that immigrant students employ in processing content and language in school contexts. We advocate for the development of multilingual, multimodal assessments.

Toshihiko Shiotsu
Kurume University

Test-taker and task characteristics as sources of variability in reading comprehension and speed

Developments in language testing in the last fifty years have identified test tasks as important sources of intrapersonal variability in language performances, while research on the roles of various language components or elements in more integrated skills such as reading comprehension has attempted to shed light on possible sources of individual performance differences. Empirical research on reading comprehension tests has generated data pointing to the effects of text characteristics and response formats (e.g., Kobayashi, 2009), but further research is needed on their relationships to both reading comprehension and speed, the latter of which has been notably under-researched compared to the former. With the increased awareness of the need for fluency and automaticity development in language performance (e.g., Grabe, 2010), systematic investigations on the relationship between test tasks and reading comprehension speed can lead to significant implications. This research project in progress aims to explore test-taker characteristics that explain the individual differences in reading comprehension performance and in reading comprehension speed. It also simultaneously addresses task characteristics accounting for the intrapersonal differences in both comprehension and speed. The study involves a sample of 120 test takers whose reading comprehension and speed are assessed on computer-based reading tests and whose individual characteristics are recorded through a questionnaire. Systematic variations in the text characteristics and response formats enable examinations of their effects on performance differences. The presentation will be an opportunity for exchange of ideas related to the topics of test-taker characteristics, test-task characteristics, and reading comprehension and speed.

Liping Su
Ohio State University

Preschool teachers' attitudes towards assessment on young English language learners

In recent years, there has been a steady increase in young English Language Learners in U.S. In Ohio, more than 35,000 ELLs were enrolled in the state's elementary and secondary public schools during the 2006–2007 school year. With the No Child Left Behind Act (NCLB) in 2001, each state is required to assess the English Language Learner's proficiency on an annual basis. Also the federal Head Start program has been set up to ensure equal early childhood education (ECE). Following NCLB and ECE, in the past decade, there has been an increasing effort in the test development and research on young language learners, including children from preschool to the earliest elementary years (Bailey, 2008). Studies has focused on different types and purposes of assessment, for example normalized assessment and formative assessment (Tsagari, 2004), high-stakes assessment and diagnostic assessment (Genesee and Upshur, 1996). However, seldom is attention paid to the role of language assessment in helping young language learners' transition from preschool to kindergarten, which according to Ohio Department of Education (ODE) is a weak link in the schooling system. The proposed study is to explore a university affiliated preschool teachers' attitudes towards three tests used for assessing young language learners—Ohio KRA-L test (Kindergarten Readiness Assessment), OTELA (Ohio Test of English Language Acquisition) and IPT (IDEA Proficiency Test). Data will be collected via survey, classroom observation, and interviews, and will be triangulated for analysis. The major research questions are: (1) how do they think about the tests in terms of their format and content? (2) how do the tests address children's cognitive development level and cultural background? (3) how would the tests influence their teaching practices with ELLs in their class? The presentation will conclude with a discussion of the study's findings with respect to test reliability and validity issues.

Youyi Sun
Queen's University

Construct validity: Exploring the dimensionality of the college English proficiency

This study aims to investigate construct validity of the College English Test (CET) in China focusing on the dimensionality of the construct measured by the test. The following research question is to be addressed: What are the features of the college English proficiency as measured by the CET? The CET is designed to measure the English language proficiency of students in tertiary institutions in China. It adopts a componential view of the language construct and operationalizes it in terms of listening; reading; integrated skills (cloze or error identification and correction); as well as writing and translation. Both the overall score and the profile score for each component are reported (Jin, 2010). The CET has gone through some major and important changes in content, format and score reporting in the past decade; however, there have been few studies on its validation. These studies (Yang and Weir, 1998; Jin & Wu, 1998) were mostly conducted 10 years ago and have focused mainly on the reading section (Zheng & Cheng, 2008). In this study, data are collected from two sources: quantitative CET scores and retrospective verbal report of test takers' strategies used to respond the test items. Factor analysis is used to define factors that account for consistency and variation in the test scores. Convergence among scores on different sections of the test suggests unidimensionality of the construct measured by the CET whereas divergence indicates that different components of college English proficiency can be separately assessed. Data from retrospective interviews are analyzed to provide in-depth understanding of the CET construct validity. The study illuminates the nature of the construct measured by the CET and provides evidence for the test validation. The results of the study will also be used for a further study on the CET washback within Messick's (1989; 1996) validity theory.

Marta Tecedor Cabrero
University of Iowa

Communicative and interactional competence in computer-based oral tests for beginning learners

The use of new technologies is becoming a crucial part of the foreign language curriculum, thus it should be an important component of the assessment process as well. Current development of technology facilitates both the integration of multimedia interactive materials and the recording and storage of learners' speech samples, which allows us to start focusing on productive skills such as speaking. This study explores the ways in which two types of computer-based oral assessment provide information about beginning learners' oral production. Two computer-based exam formats were designed to elicit participants' communicative competence in a picture description task and in a role-play situation. Exam 1 uses videoconferencing techniques to present the input in the picture description task as well as to facilitate interaction between the participants in the role-play. Exam 2 uses written and static visual input for the picture description task and written and prerecorded video input during the dialogic task to simulate a conversation. Students' language samples are recorded using a voice recorder. The main guiding questions of the study are: Is there an interaction between degree of task structure and accuracy and fluency of students' performance? Are the two tests comparable in terms of lexical density, speech functions and communicative and interactional strategies they elicit? Discourse analysis techniques are used to explore the discourse produced in the two tests. Preliminary results show that even though both tests are similar in terms of the language functions they elicit, the direct version of the test, consistently with an Interactionist approach (Hall, 1993: Young, 2000), promotes greater use of interactional features (i.e. turn-taking, means for signaling boundaries, topic management strategies, compensatory strategies). However, the results also indicate that the guided structure of the indirect version helped beginning learners to stay on task, and contribute to greater fluency.

Aylin Unaldi
Bogaziçi University

Theories and assessment of reading for academic purposes: Sentence, text and multiple-text levels

Although reading for academic purposes is a widely researched area, there is still need to improve our understanding of it. This study draws on the reading theories that explain reading comprehension with emphasis on different levels of careful reading such as sentence, text and multiple texts in order to explicate that increasingly more complex cognitive processes explain higher levels of reading comprehension. It is suggested that reading tests of English for Academic Purposes (EAP) should involve not only local level comprehension questions but also reading tasks at text and multiple texts levels. For this aim, cognitive processes extracted from the theories defining each level of reading, and contextual features extracted through textual analysis of university course books were combined to form the test specifications for each level of careful reading and sample tests assessing careful reading at sentence, text and intertextual levels were designed. Statistical findings through ANOVA and PCA confirmed the differential nature of the three levels of careful reading; however, the expected difficulty continuum could not be observed among the tests. Possible reasons underlying this are discussed, suggestions on reading tasks that might operationalise text level reading more efficiently and intertextual level reading more extensively are made, and additional components of intertextual reading are offered for the existing frameworks that explain reading at textual level.

Xuechun Zhou
Michigan State University

Liyang Mao
Michigan State University

Xiaoqing Chen
Michigan State University

Comparing proficiency classification in EFL standardized exam: Unidimensional and multidimensional IRT approaches

This study examines the effect on proficiency classification in reading comprehension when unidimensional and multidimensional item response theory (UIRT and MIRT) models are applied to an EFL large scale standardized exam. The study investigates: 1) how classification of examinees' lexical and syntactic knowledge as estimated from MIRT model differs from that of UIRT models; 2) how the differences in classification relate to the test information. Previous research has shown that the UIRT model works well in estimating examinees' reading comprehension ability and the proficiency classifications from various UIRT models are comparable (Wainer, 1995; Zhang, 2010). However, since vocabulary and syntactic knowledge/skills are known to be relevant to foreign language reading comprehension ability (Barnett, 1986; Bernhardt, 1999; Shiotsu & Weir, 2007), it is worthwhile to explore how classification differs when MIRT models are used. Also, test information from the two approaches will be used to explain the classification accuracy across the five proficiency levels. The exam has 100 multiple choice items measuring four components: grammar, vocabulary, cloze, and reading comprehension. The sample size is about 35,000 examinees. The study is conducted by: 1) estimating examinees' ability using UIRT three parameter logistic (3PL) model and the multidimensional extension of the 3PL (M3PL) model; 2) determining the corresponding cut points on the ability continuum from the empirical scale score distribution; 3) classifying the examinees according to their ability estimates; 4) comparing the classifications from the 3PL model to the classification of lexical and syntactic ability as estimated from the M3PL model; 5) linking the test information and corresponding standard error of measurement (SEM) on the corresponding cut points to the classification results. The implications of the study include using diagnostic information obtained from MIRT model for more accurate classification, and constructing tests with multimodal test information to achieve this goal.

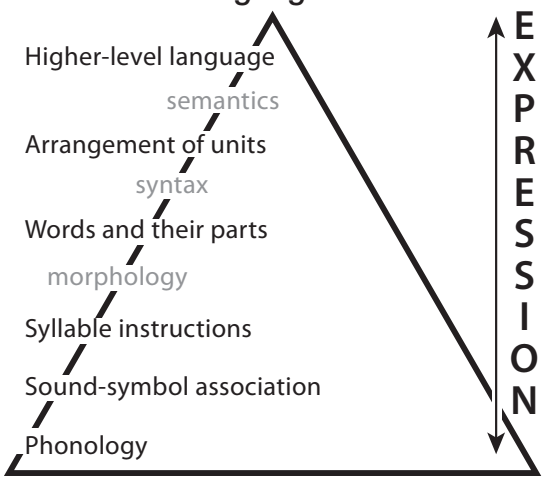
The Learning Pyramid

Assess your student's language skills. Use Decoding to Composition Tests

These tests are based on the theoretical concept that language is a puzzle with many pieces that needs to be put together to form the whole.

The components of the test assess the student's ability to put together the units of language. The components of the test evaluate the student's phonological, morphological and syntactical knowledge as well as how all these levels generate meaning. A lesson plan can then be devised to strengthen the gaps in the student's capacity to master the skills needed for reading and writing proficiency.

Structure of Language



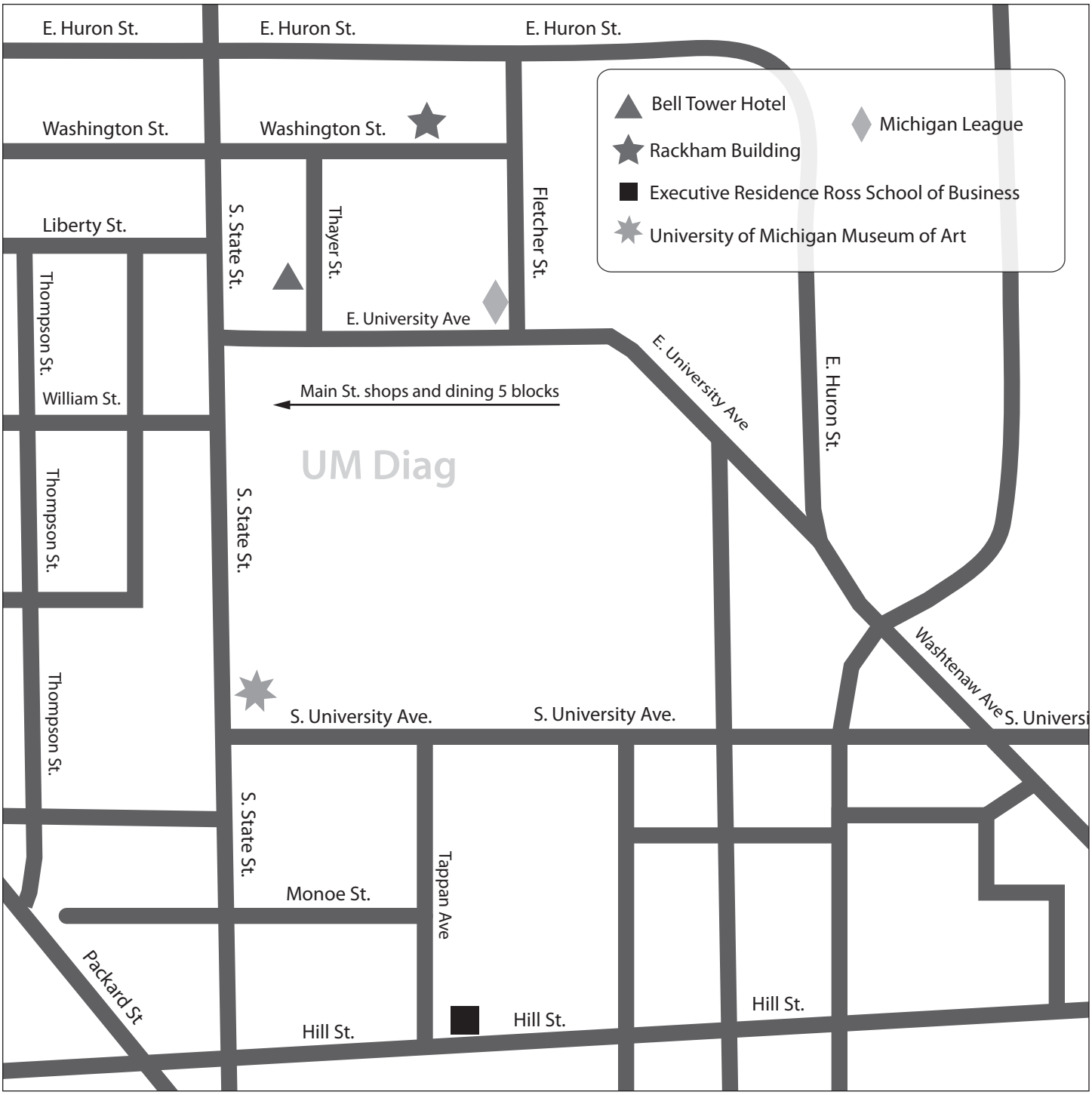
This assessment finds the broken links in the chain and shows how to mend them!

These tests have been developed by:

The Learning Pyramid
Director Stephanie Miller MA
2330 Shelter Bay Avenue
Mill Valley, CA 94941

Tel: 415 389 1094
Email: clack04@sbcglobal.net
Web: http://web.me.com/clack04/TheLearningPuzzle/Welcome.html

Map


E. Huron St.
E. Huron St.
E. Huron St.
Bell Tower Hotel
Michigan League
Rackham Building
Executive Residence Ross School of Business
University of Michigan Museum of Art
Washington St.
Washington St.
Liberty St.
S. State St.
Thayer St.
Fletcher St.
Thompson St.
E. University Ave
E. University Ave
E. Huron St.
William St.
Main St. shops and dining 5 blocks
UM Diag
Thompson St.
S. State St.
Thompson St.
S. University Ave.
S. University Ave.
Washtenaw Ave
S. State St.
Tappan Ave
Monoe St.
Packard St
Hill St.
Hill St.
Hill St.

Index

Index

Getting a message across is not always this easy.

Our mission is to provide test development, translation/adaptation, and other related services in the areas of second language proficiency testing and the testing of nonnative English-speakers.

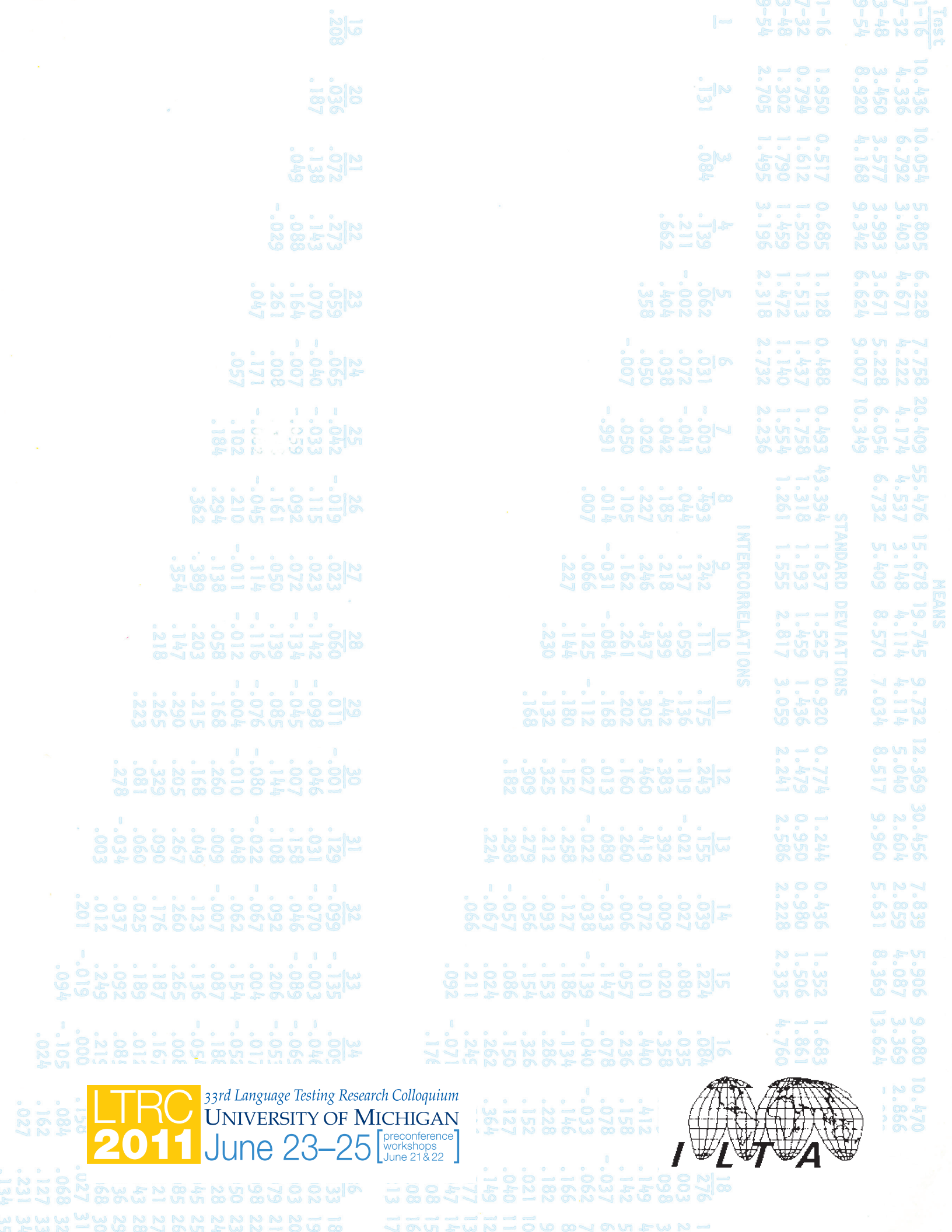
2LT SECOND LANGUAGE TESTING INC.

A Berlitz® Company

For further information:

Please be sure to visit our table in the exhibits area or contact Charles Stansfield at cstansfield@2LTI.com or at:

6135 Executive Boulevard
Rockville, MD 20852-3901
United States of America
Ph:301 231 6046 • F: 301 231 9536
www.2LTI.com


LTRC
2011
33rd Language Testing Research Colloquium
UNIVERSITY OF MICHIGAN
June 23–25
preconference workshops June 21 & 22
ILTA