23rd Annual
Language Testing Research Colloquium

LTRC 2001

Theme:
Language Testing and Technology

Marriott Pavilion Hotel
St. Louis, Missouri, USA

February 20-24, 2001
### Friday, February 23

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00-11:00</td>
<td>Continental Breakfast</td>
<td>East Foyer</td>
</tr>
<tr>
<td>8:00-9:30</td>
<td>Registration</td>
<td>Outside of the Pavilion Ballroom</td>
</tr>
<tr>
<td>8:30-12:00</td>
<td>Book Exhibit</td>
<td>Pavilion Suite</td>
</tr>
<tr>
<td>8:30-10:00</td>
<td><strong>Paper Session 5</strong></td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>Chair: Barry O’Sullivan, <em>University of Reading</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Construct Validation of an Integrated Communicative Language Test</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nathan Carr, <em>University of California at Los Angeles</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Component Dependency and Adult ESL Students' Performance in a Language Test</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hameed Esmaeili, <em>University of Toronto</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Mirror, Mirror on the Wall: Seeking Identities in Classroom Based Assessment</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pauline Rea-Dickins, <em>University of Bristol</em></td>
<td></td>
</tr>
<tr>
<td>10:00-10:20</td>
<td>Break</td>
<td></td>
</tr>
</tbody>
</table>
Friday, February 23

10:20-11:50   Paper Session 6
   Chair: Beryl Meiron, Cambridge Examinations and IELTS International

   Examining a Process-oriented Writing Assessment for Large Scale Assessment
   Yeonsuk Cho, University of Illinois at Urbana-Champaign

   Access to Technology: An Analysis of the Language Learning Background of Examinees Taking the Test of English as a Foreign Language (TOEFL)
   April Ginther, Purdue University; Carol Taylor, Biola University; and Krishna Prasad, Purdue University

   Developing a Computerized System for Investigating Non-linguistic Factors in L2 Learning and Test Performance
   Nick Saville, University of Cambridge Local Examinations Syndicate, and Jim Purpura, Columbia University

11:50-1:15     Lunch

1:15-3:15   Symposium 3
   Organizer: Charles Alderson, Lancaster University

   Learning-centred Assessment Using Information Technology
   Charles Alderson, Lancaster University; Ari Huhta, University of Jyvaskyla; Neus Figueras, Catalonia Department of Education; Sari Luoma, University of Jyvaskyla; Mirja Tarnanen, University of Jyvaskyla; and Steve Fligelstone, Lancaster University

3:15-3:30     Break
Friday, February 23

**3:30-5:15 Poster Session**
Chair: Felicity Douglas, *Iowa State University*

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:30-4:30</td>
<td>Introduction to Posters</td>
</tr>
<tr>
<td>4:30-5:15</td>
<td>Viewing</td>
</tr>
<tr>
<td></td>
<td>Ana Borderia-Garcia and Ana Oskoz, <em>The University of Iowa</em></td>
</tr>
<tr>
<td></td>
<td>2. Test-taker Attitudes Towards the Computer-based TOEFL</td>
</tr>
<tr>
<td></td>
<td>Lindsay Brooks, <em>University of Toronto</em></td>
</tr>
<tr>
<td></td>
<td>3. Development and Initial Validation of a Web-based ESL Placement Test</td>
</tr>
<tr>
<td></td>
<td>Nathan Carr, Brent Green, Viphavee Vongpumivitch, and Xiaoming Xi, <em>University of California at Los Angeles</em></td>
</tr>
<tr>
<td></td>
<td>4. The Framework for the 2003 Foreign Language National Assessment of</td>
</tr>
<tr>
<td></td>
<td>Educational Progress (NAEP)</td>
</tr>
<tr>
<td></td>
<td>Dorry Kenyon, <em>Center for Applied Linguistics</em></td>
</tr>
<tr>
<td></td>
<td>5. Effect of Topics and Cognitive Skills in Assessing Foreign Language Listening Ability</td>
</tr>
<tr>
<td></td>
<td>Yi Lin, <em>The University of Iowa</em></td>
</tr>
<tr>
<td></td>
<td>6. Technology is Like an Egg: Metaphors and Analogies in Performance Testing</td>
</tr>
<tr>
<td></td>
<td>Carol Lynn Moder and Gene Hallek, <em>Oklahoma State University</em></td>
</tr>
<tr>
<td></td>
<td>7. Spanish Speaking Test for Elementary Students: SOPI</td>
</tr>
<tr>
<td></td>
<td>Lucia Osa-Melero and Rebeca Bataller, <em>The University of Iowa</em></td>
</tr>
<tr>
<td></td>
<td>8. The Effects of Contrastive Test Equating Strategies on the Decision</td>
</tr>
<tr>
<td></td>
<td>Consistency of Criterion-referenced Foreign Language Tests</td>
</tr>
<tr>
<td></td>
<td>Steven Ross, <em>Kwansei Gakuin University</em></td>
</tr>
<tr>
<td></td>
<td>9. A Multivariate Study of L2 Reading: A Preliminary Analysis</td>
</tr>
<tr>
<td></td>
<td>Toshihiko Shitsu, <em>Kurume University</em></td>
</tr>
<tr>
<td></td>
<td>10. Accommodating Students with Disabilities</td>
</tr>
<tr>
<td></td>
<td>Mary Spaan, <em>University of Michigan</em></td>
</tr>
<tr>
<td></td>
<td>Stephanie Stauffer and Dorry Kenyon, <em>Center for Applied Linguistics</em></td>
</tr>
</tbody>
</table>
### Friday, February 23

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:15-5:30</td>
<td><strong>Concluding Remarks</strong></td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>Micheline Chalhoub-Deville, <em>The University of Iowa</em></td>
<td></td>
</tr>
<tr>
<td>6:00-7:00</td>
<td><strong>Cash bar</strong></td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td>7:00-10:00</td>
<td><strong>Banquet, Awards and Jazz Music</strong></td>
<td>Pavilion Ballroom</td>
</tr>
</tbody>
</table>

Awards sponsored by IELTS partners--UCLES, The British Council, and IDP Education Australia

- The MA Dissertation Award
  - **Lindsay Brooks**, *Ontario Institute for Studies in Education at the University of Toronto*
  - **Sally O'Hagan**, *University of Melbourne*

Awards sponsored by the International Language Testing Association

- **1998 ILTA Best Article Award**
  - **Carol Chapelle**, *Iowa State University*

- **1999 ILTA Best Article Award**
  - **John A. Upshur**, *Concordia University &
    **Carolyn E. Turner**, *McGill University*

- Robert Lado Award for Best Student Paper at LTRC
  - To be announced

Awards sponsored by the TOEFL Board, Educational Testing Service

- The Samuel Messick Memorial Lecture Award
  - **Richard Luecht**, *University of North Carolina at Greensboro*

- The Jacqueline A. Ross Dissertation Award
  - **Dianne Wall**, *Lancaster University*

- The Outstanding Young Scholar Award
  - **Micheline Chalhoub-Deville**, *The University of Iowa*
Saturday, February 24

8:30-12:00  Joint LTRC/AAAL Symposium  Grand Ballroom
Organizer: Dan Douglas, Iowa State University  Hyatt Regency
                   St. Louis

The Brahmin and the Elephant: Defining and Assessing Speaking Ability

Discussants: Elaine Tarone, University of Minnesota, and Alan Davies, University of Edinburgh

The Speaking Construct in Historical Perspective
Bernard Spolsky, Bar-Ilan University

Speaking as a Psycholinguistic Process: The Machine Within
Kees de Bot, University of Nijmegen

Speaking as Register
Susan Conrad, Iowa State University

Speaking as Performance Within a Discourse Domain
Larry Selinker, Birkbeck College

Speaking as a Cognitive Tool
Merrill Swain, Ontario Institute for Studies in Education

Speaking as a Realization of Communicative Competence
Lyle Bachman, University of California at Los Angeles

Speaking as Proficiency
Judith Liskin-Gasparro, The University of Iowa

The Role of Speaking in Discursive Practice
Richard Young, University of Wisconsin-Madison
ABSTRACTS

Abstracts are listed in the following order: Workshops, Plenary, Symposia, Papers, Works in Progress, and Posters.

WORKSHOPS

An Introduction to Multivariate Generalizability Theory
Robert Brennan, The University of Iowa

Generalizability theory liberalizes and extends classical test theory through marrying analysis of variance procedures with a rich conceptual framework for measurement. In doing so, generalizability theory enables an investigator to disentangle multiple sources of error that influence the dependability of measurements. This workshop will begin with an overview of the basics of univariate generalizability theory for balanced designs. Then attention will be directed to unbalanced designs, followed by an introduction to multivariate generalizability theory. Mathematical details will be treated only minimally, with major emphasis focused on conceptual understanding. Computer programs will be demonstrated that perform computations for virtually all of the procedures discussed in the workshop. It is assumed that participants in this workshop have some familiarity with univariate generalizability theory.

Robert L. Brennan received his doctorate from Harvard Graduate School of Education in 1970. He then served on the faculty of SUNY Stony Brook until 1976 when he became Senior Research Psychologist at American College Testing (ACT). Subsequently, he held several other positions at ACT including Assistant Vice President for Measurement Research and Distinguished Research Scientist. In 1994 he became the E.F. Lindquist Professor of Educational Measurement and Director of the Iowa Testing Programs in the College of Education of The University of Iowa.

Dr. Brennan is the author or co-author of two books: Elements of Generalizability Theory and Test Equating Methods and Practices, and is completing a new book on generalizability theory (G-theory). In addition, he has published numerous articles in professional journals on various topics in educational measurement, including generalizability theory, equating, scaling, performance assessment, standard setting, and domain-referenced testing. He has also developed a number of computer programs used for G-theory and equating. Dr. Brennan has served as President of the Mid-Western Educational Research Association, President of the Iowa Academy of Education, Vice President of Division D (Measurement) of AERA, and President of NCME. He is the co-recipient of the 1980 AERA Division D Award in Measurement Theory and Practice, the 1997 NCME Award for Outstanding Technical Contribution to the Field of Educational Measurement, and the 2000 NCME Award for Career Contributions to Educational Measurement.
PLENARY

New Directions in Computerized Testing Research
Richard Luecht, University of North Carolina at Greensboro

The potential uses of computers in testing go far beyond the implementation of adaptive-testing algorithms, the integration multimedia and sounds into tests, and the use of new, creative item formats. Computers make it possible to actually change the very nature and type of assessment information gathered and the method of collecting it. This paper reviews some of the state-of-the-art developments in computerized testing and presents some of the challenges and possibilities for future research.

Richard M. Luecht (Ph.D., University of Wisconsin-Milwaukee, 1989) is currently Professor and Chair of the Department of Educational Research Methodology at the University of North Carolina Greensboro (UNCG) and Director of the Center for Educational Research and Evaluation at that institution. Prior to joining the UNCG faculty, Dr. Luecht was involved with several large scale testing programs: he was the Project Director for Computerized Adaptive Testing Research and Senior Psychometrician at the National Board of Medical Examiners (NBME) where he also helped develop large-scale automated test assembly technologies for the NBME; he was a Psychometrician with ACT, Inc. where he conducted research on automated test assembly, computer-adaptive testing and IRT estimation issues; he served as chief psychometrician for the National Assessment of Educational Progress (NAEP) standard setting project; he has served as a consultant/advisor for the computerization of numerous exams such as the Medical College Admissions Test, the Law School Admissions Test, the United States Medical Licensing Examination, and even Microsoft certification exams with responsibilities ranging from test construction to calibration, equating and scoring.

The majority of his research has focused on extending theoretical developments in psychometrics, statistics, and operations research to practical assessment applications. Examples include: extending applications of multidimensional item response theory to computer-adaptive testing; developing new item selection algorithms and heuristics for automated test assembly including the recent introduction of a new and comprehensive test development framework known as computer-adaptive sequential testing (CAST); designing robust calibration methods for estimating polytomous IRT model parameters used with complex performance assessments; exploring Bayesian estimation issues in IRT; applications of structural equation modeling to construct validation; exploring the use of latent class models and developing scoring and analysis systems based on these models for computer-based testing; designing effective computerized testing interfaces; evaluating relationships between item features and test speededness; and research into statistical methods for detecting aberrant test-taking practices.
SYMPOSIUM 1

Computer-Based TOEFL: A Discussion of Some Technological and Psychometric Issues.
Xiang Bo Wang, Daniel Eignor, Marna Golub-Smith, Yong-Wong Lee, and Patricia Carey,
Educational Testing Service

The Test of English as a Foreign Language (TOEFL) has been a computer-based testing (CBT) program since July 1998. During the past two years, a substantial amount of practical knowledge and experience has been obtained with regard to the technology that drives a technically sound large-scale CBT program. The purpose of this symposium is to summarize some key technological components of the TOEFL CBT implementation and to discuss some of the test’s psychometric properties.

This symposium consists of four presentations. The first presentation, entitled “Item Selection Algorithms Used in TOEFL CBT,” will summarize both the computer adaptive and linear-on-the-fly (LOFT) item selection mechanisms used with TOEFL CBT, as well as the details of how and why these two mechanisms were chosen for use. Furthermore, it will discuss the pros and cons by comparing these procedures to other available procedures that have been developed.

The diverse cultures and linguistic backgrounds of the TOEFL examinee population pose unique challenges for the calibration and scaling of TOEFL items. This is especially true in CBT due to the unidimensional assumption of item response theory (IRT). In the past, pretest items were only administered to examinees who take TOEFL in the United States and Canada. Due to the increasing need for large numbers of new items to replenish item pools, the TOEFL program has expanded its item pretesting internationally. The second paper, entitled “Challenges in Maintaining Online Calibration and Scaling with Multilingual Examinee Populations” will explore the effects of such cultural and linguistic diversity on TOEFL online calibration, and describe methods that can be used to monitor and maintain the stability of the TOEFL scale.

One of the keys to a cost-efficient CBT testing program, especially if it is administered continuously, lies in equivalent use of as many items as possible in the item pool, subject to content constraints, test specifications and item exposure control. The third presentation, entitled “Summary of TOEFL CBT Item Pool Creation, Maintenance, and Usage” will summarize the algorithm, process and challenges of creating viable item pools for the continuous testing, and maintaining a sound collection of items. Furthermore, it will report on the overall and conditional item pool usage, given the test specifications and the continuous administration nature of the TOEFL CBT. New approaches on how to improve item pool integrity and pool usage will be offered for further research.

To make the direct measure of writing ability as part of the TOEFL score, TOEFL CBT has also incorporated into the Structure and Written Expression section an essay component that its paper-and-pencil (P&P) predecessor did not. These constructive-response essay items are scored by trained readers via an on-line scoring network system (OSN). The final presentation, entitled “The Essay Scoring and Scorer Reliability in TOEFL CBT,” will first describe the OSN system, discuss psychometric issues related to the scoring process, such as interrater reliability and rating stability monitoring via OSN, and outline future research avenues for CBT essay scoring including multifaceted analyses of ratings and e-raters.

It is the sincere hope of the presenters that the four presentations in this symposium will provide a concise summary of the theory and technology that support TOEFL CBT to the field of language testing at large. More importantly, we wish to use this symposium as a platform to solicit valuable feedback and suggestions from language testing experts in the audience.
SYMPOSIUM 2

Using Language Corpora in Language Testing. Sarah Briggs, University of Michigan; Dan Douglas, Iowa State University; Susan Nissan, Educational Testing Service; John Read, Victoria University of Wellington; Rita Simpson, University of Michigan; and Charles Alderson, Lancaster University

Technological advancements in the last few decades have enabled the collection of large computerized language corpora. Written corpora have been available for the past few decades, and spoken corpora are just now becoming available. These resources have proven useful for linguistic research as well as lexicographic projects, pedagogical grammars, and language teaching materials, but as Alderson (1996) observes, "in the assessment of language learning and proficiency-language testing-corpora have yet to find an application" (p. 249). This symposium composed of three presentations focuses on recent initiatives that attempt to use language corpora as resources for test development.

1. Developing Listening Prototypes, Using a Corpus of Spoken Academic English

As part of a major revision of the Test of English as a Foreign Language (TOEFL), known as TOEFL 2000, the Research and Development Oversight Committee commissioned the collection of a million-word corpus of spoken academic English. The corpus was compiled by students at five tertiary-level institutions across North America, who took tape recorders with them to record interactions in lectures, in-class discussions, office hour visits, service encounters, and informal study groups. The resulting tapes were transcribed for analysis and use by the TOEFL 2000 Listening Team.

In this presentation we will describe how the text types included in the corpus were defined, how the corpus was developed, how excerpts were selected for listening task prototyping, and what problems were encountered when using the excerpts in developing test tasks. We will also describe tools being developed for corpus analysis, using a framework proposed by Biber (1995): Involved/Informational Production, Narrative/Non-narrative Discourse, Situation-depended/Elaborated Reference, Overt Expression of Argumentation, and Abstract/Non-abstract Style.

We will particularly focus on problems that test developers encountered during the creation of prototype tasks, including mistakes in content, inappropriate content, content requiring background knowledge from previous classes, and a paucity of testable content. We will conclude by discussing how the corpus may be used by test development in the future.
2. Contributions of Corpus Analysis to Vocabulary Assessment

Since the frequency and range of word forms are so easily counted by computer, corpus analysis has obvious potential for the production of modern vocabulary lists to replace such venerable sources as the General Service List and the Teacher's Word Book. However, contemporary vocabulary assessment requires not only general inventories of word forms but also information about the lexical dimension of language use in particular social and educational contexts. This presentation will explore how analysis of corpora may contribute to an expanded view of vocabulary testing which incorporates multiword lexical items, the vocabulary of informal speech and the lexical features of different genres or registers.

3. Using an Academic Corpus to Evaluate the Lexis of EAP tests

Just as those responsible for EAP courses are concerned that the materials they use are relevant to the goals of their students, test takers and developers have concerns about whether EAP tests are relevant and appropriate. This presentation reports the results of a study using corpus data to evaluate the lexical content of certain EAP tests. We explored the tests with regard to content relevance, which Fulcher (1999) points out is more related to construct validity than content validity.

Word frequency information from two newly compiled corpora of academic English, one written corpus and one spoken corpus, is used to study the content of various components of two tests. One test is a test designed and used to evaluate incoming non-native speakers at an American university and the other test is an advanced level general English proficiency exam that, among other things, can be used as evidence of English proficiency by applicants to English-medium universities. Three sub-components of the tests were selected for analysis: the vocabulary sections, the reading sections, and the listening sections.

The first part of the research project involved comparing frequency information of words in the spoken academic corpus (MICASE, 2000) with three existing word lists-two lists of 1,000 words each from A General Service List of English Words (West, 1953) that have been updated to include up to Level 6 words of Bauer and Nation (1993); and the Academic Word List (Coxhead, 1998, 2000). Since all of these word lists were derived from written sources, this comparison is useful for determining lexical overlap and lexical uniqueness of spoken and written genres. The second part of the project involved analyzing the lexical content of the three sub-components of the two tests to determine content relevance with respect to naturally occurring lexis in the spoken corpus and the word lists. Finally, item facility information for the vocabulary test items was compared with relative frequency of occurrence in the corpora.
Acknowledgements

**LTRC 2001 Organizing Committee**

Micheline Chalhoub-Deville, *The University of Iowa*, LTRC Co-Chair

Craig Deville, *The University of Iowa*, LTRC Co-Chair

Dan Douglas, *Iowa State University*, LTRC Co-Chair

Felicity Douglas, *Iowa State University*, LTRC Co-Chair

We would like to thank the following for their financial support:

- English Language Institute, University of Michigan
- TOEFL Board, Educational Testing Service
- University of Cambridge Local Examinations Syndicate
- Those who contributed to the jazz band fund

We also wish to thank the following organizations for participating in the book exhibit:

- Cambridge Examinations and IELTS International
- Cambridge University Press
- Educational Testing Service
- Yale University Press
SYMPOSIA

SYMPOSIUM 3

Learning-centered Assessment Using Information Technology. Charles Alderson, Lancaster University; Ari Huhta, University of Jyvaskyla; Neus Figueras, Catalonia Department of Education; Sari Luoma, University of Jyvaskyla; Mirja Tarnanen, University of Jyvaskyla; and Steve Fligelstone, Lancaster University

This symposium will provide a forum for a discussion of the potential benefit to language learning of the use of information technology (IT) in language assessment. We will illustrate this by reference to the new DIALANG assessment system.

DIALANG is a European multi-national project which is creating a diagnostic language assessment system on the Internet. The purpose of DIALANG is to give its users diagnostic feedback on their language proficiency and their self-assessment abilities. One of the aims of DIALANG is to contribute to increased learner autonomy by encouraging reflection on learners language use, and on their ability to assess their learning. It hopes to make language learners more aware of the nature of language proficiency in general and to develop their ability to judge their own proficiency. The system will be operational in fourteen European languages, and will be launched on the Internet during the European Year of Languages (2001).

Four complementary papers will be presented. The first one describes the DIALANG philosophy and its implementation on the Web, in particular how self-assessment is integrated into the assessment system and how feedback is provided to the users.

The second paper presents issues of implementing a large-scale assessment system in the Internet environment and describes and discusses the solutions that DIALANG is working towards.

The third paper presents the rationale behind the feedback system that has been developed, describes the system as implemented, and presents the results of ongoing qualitative research (using observation and interviews) into how learners use, understand and appreciate the feedback presented.

The fourth paper explores in detail the benefits and effects of integrating self assessment and computer delivery through the development of, and experimentation with, a benchmarking system, which involves eliciting learner productions in writing and speech, and getting them to compare their performance with benchmarked performances, which have already been commented on in detail by expert judges. The results of a study of the use of this system with learners of Finnish will be presented and discussed.

The symposium will conclude with an open discussion of the use of IT and in particular the Internet, for the delivery of language tests in general, and of learning-related, low-stakes, diagnostic tests in particular, focussing on the potential for innovation and the added value to users of this mode of delivery.
 symposia

LTRC/AAAL SYMPOSIUM: THE BRAHMIN AND THE ELEPHANT: DEFINING AND ASSESSING SPEAKING ABILITY

The Speaking Construct in Historical Perspective. Bernard Spolsky, Bar-Ilan University

While those who start their histories of language testing in the 1960s assume that the FSI Oral Interview was the first venture into testing speaking, the previous half-century was filled with research in how to test speaking. I will talk in particular about the 1913 committee, the Yale and Columbia efforts, and the British and American experience in the Second World War.

Speaking as a Psycholinguistic Process: The Machine Within. Kees de Bot, University of Nijmegen

In this contribution I want to argue that speaking involves many subprocesses that are mechanistic in nature and that cannot be controlled. In language testing very little attention has been paid to the mechanistic part of speaking as opposed to the intentional or conceptual part.

Speaking as Register. Susan Conrad, Iowa State University

Is it misleading to discuss "speaking" as a single variety of language? Are spoken registers just as different from each other as from written registers? I will address these questions with reference to the linguistic characteristics of several spoken registers at universities: lectures, discussion groups, advising sessions, and service encounters.

Speaking as Performance Within a Discourse Domain. Larry Selinker, Birkbeck College

The 'discourse domains view of interlanguage' (Douglas, 2000) is reviewed, noting deficiencies. Here, DDs are viewed NOT atomistic but dynamic, shading into one another and overlapping. We revive the hypothesis (Douglas & Selinker, 1985) that there exists a 'test-taking domain' that test-takers unavoidably engage and try to tease out this factor by comparing 'test' speaking performance from 'non-test' speaking performance.

Speaking as a Cognitive Tool. Merrill Swain, Ontario Institute for Studies in Education

Speaking is a cognitive tool which mediates our problem-solving activities. Thus, examining speaking as process (rather than product) can reveal cognitive processes and strategies which underlie problem-solving performance. This will be illustrated with examples of the dialogue of pairs of students using their L1 while solving L2 problems.
SYMPOSIA

Speaking as a Realization of Communicative Competence. Lyle Bachman, University of California at Los Angeles

To view speaking as a "skill" that can be acquired, used and assessed confounds what language users have with what they do, and leads to irresolvable ambiguities in the way we interpret the results of research into the nature of speaking. Speaking can be more fruitfully viewed as an activity in which language users realize their communicative competence.

Speaking as Proficiency. Judith Liskin-Gasparro, The University of Iowa

A substantial body of research has documented the effect of task and method on the construct of L2 speaking proficiency. We must conclude, therefore, that the L2 oral proficiency is not constant. How it is operationalized for a particular language-use context will determine which dimensions are most salient for that context.

The Role of Speaking in Discursive Practice. Richard Young, University of Wisconsin-Madison

Many scholars of language accept a view of speaking as a self-contained system of internally coherent structures. I propose a very different way of approaching and understanding speaking. I focus on the role of speaking in discursive practice rather than on speaking in general. And I take speaking as part of a broader range of systems that underlie the organization of social life.

As language technologies have developed, standard methods of performance measurement have evolved along with them. The technologies include text-to-speech synthesis, speech-to-text conversion (speech recognition), text and speech translation, and, recently, systems for extracting and "understanding" the content of extended texts and spoken dialogs. The systems perform tasks that are similar to human tasks, such as reading or listening, and speaking. Each technology is conventionally measured in its own functional protocol, some of which are similar to those used in human language testing, e.g., read aloud, dictation, and even problem solving via open ended spoken interaction.

The paper reviews and analyzes language technology performance metrics from an historical perspective, and relates them to second language testing. Metrics have become central elements in government-sponsored R&D programs, where testing is an annual ritual followed by large meetings to discuss results (and metrics). Some cross-over is now emerging. Lippmann (1996) produced a meta-analysis comparing machine and human transcription performance. Hirschman (1999) built an automated system that takes 3rd grade reading tests; she recently proposed ETS-like reading comprehension formats to measure machine comprehension of text. We analyze an example of negative washback in synthesis engineering resulting from inappropriate performance evaluations of text-to-speech synthesis systems.

The paper describes:
1. example functional protocols and performance metrics,
2. comparative human and machine performance ranges, where applicable,
3. washback effects (good and bad) from specific evaluation processes, and
4. resources and organizations involved in language technology evaluation.

This information implies that essential differences between human and near-term machine performance will raise issues about the use of common performance measures for human and machine performance with language and about how such measures should be interpreted, but in the long run, evaluation of language automata, as such, and the testing of human language skills may converge.
Investigating Raters' Orientation in Specific-purpose Task-based Oral Assessment. Annie Brown, Noriko Iwashita, Tim McNamara, and Sally O'Hagan, University of Melbourne

This paper reports on a study concerned with the development of rating scales for the assessment of oral language skills in an English for Academic Purposes (EAP) context, as part of the TOEFL 2000 project. The study investigates the perceptions of field experts as a possible basis for the construction of these scales.

Performance tasks used in the assessment of speaking in EAP tests increasingly seek to replicate the roles and demands of students in academic contexts. Such tasks will typically require the student to integrate information from various sources (written texts, lectures, etc.), which is then used as the basis for speaking. Such spoken performances are more complex and more demanding than the kinds of stand-alone tasks where little help is provided on the content of what is to be spoken about. The absence of input materials means that stand-alone tasks are often restricted to fairly bland topics drawing on the candidate's general knowledge, and arguably under-represent the construct of speaking within academic contexts.

The question of the criteria for the assessment of performance on complex integrated tasks emerges as an urgent question. Performance on such tasks involves the integration of cognitive skills and more narrowly defined language proficiency. To what extent do raters attend to each of these different dimensions of performance?

A verbal protocol methodology was used to investigate the natural or unguided judgements of university-based oral communication skills specialists on a range of stand-alone and integrated speaking tasks of the type under consideration for the new test. Stand-alone tasks included argument and the presentation of graphic information. Integrated tasks included summarizing oral and written academic text. Each of twelve oral communication skills experts listened to a series of taped performances on the different tasks by learners at different proficiency levels. Immediately after each listening they reported their perceptions of the quality of the performances.

The protocol data were analyzed qualitatively using NUD*IST software to consider the following:
1. the balance between an orientation to general cognitive skills and language proficiency, narrowly defined
2. possible differences in raters' orientation to criteria in stand alone and integrated tasks
3. possible differences in raters' orientation to criteria according to proficiency level.

The results are useful in the establishment of criteria to be used in rating scale construction, in understanding constraints on the type of scale to be used (holistic or multiple), and in informing rater training in the use of such scales.
Construct Validation of an Integrated Communicative Language Test. Nathan Carr, University of California at Los Angeles

Test designs incorporating the principles of thematic unity and task interdependence have been discussed in the language assessment literature for a number of years, and several examples have been developed. These features, which are intended to improve the authenticity, interactivity, and construct validity of the test, are particularly attractive with computer- and web-based language testing, which have the potential for presenting authentic material and engaging test takers in tasks that may be more interactive than those possible in P&P formats. However, while the authenticity and interactivity may be enhanced by thematic unity and task interdependence, especially when combined with the benefits offered by greater technological sophistication, these integrative qualities may at the same time lead to test bias, as well as to violations of the measurement assumptions of unidimensionality and local independence. Such questions regarding dimensionality and construct validity in so-called communicative language tests have been discussed extensively in the language testing literature. This study attempts to investigate the factor structure of one such test to determine the extent to which evidence can be found in support of reporting separate scores for its four interdependent, thematically unified component parts.

The test in question, a test of academic Spanish language ability used by the University of California Education Abroad Program in order to assess the readiness of 214 students for study overseas, was designed to test academic reading, writing, speaking, and listening. What makes this test of even greater potential interest to test developers, however, is that its use of broadcast quality videotaped lectures for the listening passages probably represents the pinnacle of appropriate technology use in a P&P test.

Initial exploratory factor analyses identified six distinct factors: speaking, written rhetorical ability (content and organization), written linguistic accuracy (vocabulary, cohesion, and grammar), ability to apply concepts read in a text, reading for specific details, and academic listening. Subsequent analyses using structural equation modeling confirmed this factor structure, and indicated that the six first-order factors in turn load on a second-order factor identified as academic Spanish language ability. Results further indicated that the increased authenticity gained by using videotaped listening passages did not necessarily translate into improvements in the factor structure—and by inference the construct validity—of the items in that section of the test. These findings will be discussed in terms of their implications for test design, the dimensionality of thematically unified tests, score reporting, and the need to investigate the construct validity of score-based inferences regardless of the increased authenticity and interactiveness made possible by more sophisticated testing technologies.
Exposing a Process-oriented Writing Assessment for Large Scale Assessment. Yeonsuk Cho, University of Illinois at Urbana-Champaign

Research shows that writing ability is a competence that requires the coordination of various complex cognitive skills, and revision is an essential step that determines the quality of text. Consistent with the cognitive account of writing as a recursive process, many composition teachers denounce the use of the timed single draft test in their classrooms. They note that a single writing sample written with severe time pressure reflects less than what an individual can do as a writer. For this reason, portfolio assessment is rapidly replacing the timed single draft essay in the classrooms. Contrary to findings in writing research and the movement toward the process-oriented writing assessment in classrooms, the timed single draft essay test is dominant in large-scale EAP testing contexts for the sake of logistics but at the expense of validity.

The present study introduces a new process-oriented writing assessment that can be realized in a severely timed large-scale EAP testing context. The new writing assessment, called a workshop essay test, was administered to incoming international students as an experimental ESL Placement Test (EPT) at the University of Illinois. The workshop essay test, based on current writing theories, provided several facilitative activities such as a peer-feedback session to help examinees show their best work. The preliminary findings of the study show that the majority of examinees performed significantly better on the workshop essay test than on the current EPT.

Further analysis will be made to answer the following questions: (1) Which writing test measures examinees' true writing ability more accurately with reference to a criterion writing measure? Content-field faculty members' ratings on examinees' content papers in terms of writing ability will be used as the criterion measure. (2) Are there textual differences between the essays from the workshop and the EPT? Furthermore, examinees' responses to a questionnaire on the general writing process and test-taking experiences will be presented in order to help interpret the findings of the study. The author predicts that the workshop essay test will help enhance the validity and the accuracy of the EAP writing assessment in the EPT context. Although the present study is quite contextually bound, the author believes that the findings of the study can be discussed in other large-scale EAP writing assessment contexts.
A Tree-based Modeling Approach to Construct Validation of a Computer-delivered Speaking Test. Diane Strong-Krause, Brigham Young University

During the last decade of the twentieth century, a call for tests useful to teaching and learning has been issued. However, in order for language tests to be informative, test developers need a clear theoretical understanding of how the underlying constructs develop as proficiency increases. Proficiency scaling, or behavioral scaling as termed by Carroll (1993), is one way to connect underlying abilities with performance. This approach involves interpreting examinees' observed test scores in terms of probabilistic statements about mastery of particular skills associated with increasing levels of expertise.

A promising statistical approach to this type of scaling is tree-based modeling. The basic idea behind a tree-based approach is to uncover the structure in a set of data by organizing the test tasks into clusters that have comparable difficulty values and that require similar skills to complete. The results provide a tree-like structure detailing which skills are needed to complete similar type tasks as well as how required skills vary as proficiency increases. Further analysis provides the probabilities of completing a test task correctly with the required skills for particular levels of proficiency.

Data for this study was obtained from 168 English language students who responded to 40 speaking tasks delivered via computer. Responses were rated and analyzed using FACETS to determine the estimated b-parameter for each task. Additionally, characteristics of each task were coded based on Bachman and Palmer's (1996) task characteristics framework. A tree-based regression using a recursive partitioning algorithm (Breiman, Friedman, Olshen, & Stone, 1984) was then completed. Based on this information, probability curves were developed for each cluster of tasks.

The findings have implications for:

a) providing insights into how underlying language abilities affect speaking performance;

b) providing useful feedback to students, teachers, and administrators;

c) developing test tasks aimed at particular skill levels.

References


Revising Instruments for Rating Speaking: Combining Qualitative and Quantitative Insights. Lynda Taylor and Simon Beeston, University of Cambridge Local Examinations Syndicate

Traditionally the design and construction of rating scales for direct tests of writing and speaking ability have resorted to an a priori approach in which assessment criteria and rating scale descriptors are developed by 'experts' (i.e. teachers, applied linguists and language testers) using their own intuitive judgement. In recent years writers in these fields have advocated a more empirically-based approach to rating scale construction (Fulcher, 1987; Shohamy, 1990; Upshur and Turner, 1995; Milanovic, Saville, Pollitt and Cook, 1996). An empirically-based approach involves analysing samples of language performance to construct new assessment criteria and rating scale descriptors; it also involves investigating the way in which these are likely to be interpreted and applied by human raters.

This paper describes the approach recently used to revise the assessment criteria and rating scales for two well-established, direct tests of spoken language ability in English (IELTS and CPE). The presenter will describe how conversational and discourse analytic techniques were exploited to investigate samples of language performance from the existing speaking tests at different proficiency levels; the results of these studies helped identify salient features of developing proficiency and to inform the construction of analytic and holistic rating scales. In addition, quantitative analyses (using Rasch partial credit and multi-faceted Rasch) were used to explore issues of scale functioning, and the extent to which different raters might be interpreting and applying the scales differentially. By combining the use of qualitative and quantitative methodologies, it has proved possible to redevelop valid, reliable and practical assessment criteria and rating scale descriptors; these studies have also proved valuable in informing plans to retrain and standardise raters in preparation for the introduction of the two revised speaking tests in 2001 and 2002.
An Investigation of Judges' Point of View. Dianne Wall, Lancaster University

Standard-setting has become increasingly important to language testing researchers and test developers as demand has risen for portability of language certificates, accountability of entry and exit levels for professional training programmes, international recognition for new regional and national examinations, and comparability of assessment across skills and languages. This demand is particularly strong in the European context where an already considerable mobile labor force is expected to grow with the expansion of the European Union.

Typical standard-setting procedures involve collecting judgements about item and test difficulty and levels from expert judges, aggregating these and then performing statistical manipulations in order to arrive at a consensus view of the standards or levels of the test being developed.

However, the research that has been carried out to date on standard-setting in language assessment (Alderson 2000; Kaftandjieva, Verhelst and Takala 2000) has tended to concentrate on the large picture, and has not looked critically at the process of standard-setting from the point of view of response validity. How are the judgements arrived at, and what constructs can they be related to? Just as the process of test-taking is now recognized to affect test validity, and is seen as an important component of construct validity in the Messickian sense, so too must the process of making judgements be worthy of systematic investigation.

This paper will report on the results of an analysis of introspections and discussions amongst judges involved in a standard-setting exercise for diagnostic tests of English. The standard-setting procedures involved several expert judges familiarizing themselves with scales derived from the Council of Europe Common European Framework, undergoing a degree of harmonization of judgements, and then making decisions about the ability of candidates at given levels to answer specified items correctly. As cut scores were required to enable distinctions amongst the six different levels of the Framework, five iterations of the judgements were made for each skill tested. These procedures were repeated for tests of reading, writing, listening, vocabulary and structures.

The results throw considerable and at times controversial light on how judges arrive at their decisions, and challenge the bases on which standards are currently set in language assessment programmes.
Acknowledgements

We are also grateful to the following people who served as proposal reviewers:

Lyle Bachman, *University of California at Los Angeles, USA*

Deena Boraie, *American University of Cairo, Egypt*

Carol Chapelle, *Iowa State University, USA*

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

John de Jong, *Language Testing Services, The Netherlands*

Barbara Dodson, *University of Michigan, USA*

Cathie Elder, *University of Auckland, New Zealand*

Janna Fox, *Carleton University, Canada*

Glenn Fulcher, *Higher Colleges of Technology, United Arab Emirates*

April Ginther, *Purdue University, USA*

Leslie Grant, *Central Michigan University, USA*

Liz Hamp-Lyons, *Hong Kong Polytechnic University, Hong Kong*

Sari Luoma, *University of Jyväskyla, Finland*

Tim McNamara, *University of Melbourne, Australia*

Adrian Palmer, *University of Utah, USA*

Jim Purpura, *Columbia University, USA*

Elana Shohamy, *Tel Aviv University, Israel*

Bernard Spolsky, *Bar-Ilan University, Israel*

Gillian Wigglesworth, *Macquarie University, Australia*

We also wish to acknowledge the following people for their help in various capacities:

Glenn Fulcher, *Higher Colleges of Technology, United Arab Emirates*

Jee Wha Kim, *Columbia University*

Carrie Kiser-Wacker, *Conference Coordinator, The University of Iowa*

Ana Oskoz, *The University of Iowa*

Elvis Wagner, *Columbia University*
# LTRC 2001 Program

**Tuesday, February 20**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-12:00</td>
<td>Workshop (by pre-registration only)</td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>An Introduction to Multivariate Generalizability Theory</td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>Robert Brennan, <em>The University of Iowa</em></td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td>12:00-1:30</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>1:30-5:00</td>
<td>Workshop continued</td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>An Introduction to Multivariate Generalizability Theory</td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>Robert Brennan, <em>The University of Iowa</em></td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td>7:00-9:00</td>
<td>Registration</td>
<td>Outside of the Pavilion Ballroom</td>
</tr>
<tr>
<td>7:00-10:00</td>
<td>Welcoming Reception</td>
<td>East Foyer</td>
</tr>
<tr>
<td></td>
<td>Co-hosted by the English Language Institute, University of Michigan, and University of Cambridge Local Examinations Syndicate</td>
<td>East Foyer</td>
</tr>
</tbody>
</table>
### Wednesday, February 21

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00-11:00</td>
<td>Continental Breakfast</td>
<td>East Foyer</td>
</tr>
<tr>
<td>8:00-2:00</td>
<td>Registration</td>
<td>Outside of the Pavilion Ballroom</td>
</tr>
<tr>
<td>8:30-10:00</td>
<td>Opening session</td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>Welcome: Micheline Chalhoub-Deville, The University of Iowa</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dan Douglas, Iowa State University</td>
</tr>
<tr>
<td></td>
<td>Introduction of Keynote Speaker:</td>
<td>Craig Deville, The University of Iowa</td>
</tr>
<tr>
<td></td>
<td>Keynote Address:</td>
<td>New Directions in Computerized Testing Research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Richard Luecht, University of North Carolina at Greensboro</td>
</tr>
<tr>
<td></td>
<td><em>This is the Samuel Messick Memorial Lecture, Sponsored by the TOEFL Board, Educational Testing Service</em></td>
<td></td>
</tr>
<tr>
<td>10:00-10:20</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>10:20-11:50</td>
<td>Paper Session 1</td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>Chair: Sari Luoma, University of Jyvaskyla</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evolution of Performance Measures for Language Technologies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jared Bernstein, Ordinate Corporation, and Lynette Hirschman, MITRE Information Technology Center</td>
<td></td>
</tr>
<tr>
<td></td>
<td>What Do Listeners Do in a Computer-based Listening Test?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mikyung Kim, Soojeong Kim, and Sangkeun Shin, University of California at Los Angeles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Tree-based Modeling Approach to Construct Validation of a Computer-delivered Speaking Test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diane Strong-Krause, Brigham Young University</td>
<td></td>
</tr>
</tbody>
</table>
Wednesday, February 21

11:50-1:15 Lunch

1:15-3:15 Symposium 1
Organizer: Xiang Bo Wang, Educational Testing Service

Pavilion Ballroom

Computer-Based TOEFL: A Discussion of Some Technological and Psychometric Issues
Xiang Bo Wang, Daniel Eignor, Marna Golub-Smith, Yong-Wong Lee, and Patricia Carey, Educational Testing Service

3:15-3:30 Break

3:30-5:00 Paper Session 2
Chair: Carolyn Turner, McGill University

Pavilion Ballroom

Effects of Examinee Control on Examinee Attitudes and Performance on a Computerized Oral Proficiency Test
Dorry Kenyon, Valerie Malabonga, and Helen Carpenter, Center for Applied Linguistics

Uncovering the Nature of the Systematic Rating Discrepancies in the Test of Spoken English (TSE): Application of G-theory and the Many-faceted Rasch Model
Yong-Won Lee and Patricia Carey, Educational Testing Service

An Investigation of Judges’ Point of View
Dianne Wall, Lancaster University

5:00-5:15 Break

5:15 ILTA Board Meeting
Pavilion Ballroom
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00-11:00</td>
<td>Continental Breakfast</td>
<td>East Foyer</td>
</tr>
<tr>
<td>8:00-12:00</td>
<td>Registration</td>
<td>Outside of the Pavilion Ballroom</td>
</tr>
<tr>
<td>8:30-6:00</td>
<td>Book Exhibit</td>
<td>Pavilion Suite</td>
</tr>
<tr>
<td>8:30-10:00</td>
<td>Paper Session 3</td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td></td>
<td>Chair: Jim Purpura, Columbia University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technical Means to Predict Functional Language Proficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>John de Jong, Language Testing Services; Jared Bernstein, Ordinate Corporation; Brian North, Eurocentres; and Brent Townshend, Ordinate Corporation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Validating a Computer Adaptive Test of Vocabulary Size and Strength</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Batia Laufer, University of Haifa; Cathie Elder, University of Auckland; and Kathryn Hill, University of Melbourne</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Task Difficulty in Testing Spoken Language: A Socio-cognitive Perspective</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cyril Weir, University of Surrey Roehampton; Barry O'Sullivan, University of Reading; and Angela French, University of Cambridge Local Examinations Syndicate</td>
<td></td>
</tr>
<tr>
<td>10:00-10:20</td>
<td>Break</td>
<td></td>
</tr>
</tbody>
</table>
Thursday, February 22

10:20-11:50 **Paper Session 4**
Chair: Lijing Cheng, *Queen's University*

*Investigating Raters' Orientation in Specific-purpose Task-based Oral Assessment*
Annie Brown, Noriko Iwashita, Tim McNamara, and Sally O'Hagan, *University of Melbourne*

*Using Can-do Statements to Equate Computer-based Tests Across Languages*
Neil Jones, *University of Cambridge Local Examinations Syndicate*

*Revising Instruments for Rating Speaking: Combining Qualitative and Quantitative Insights*
Lynda Taylor and Simon Beeston, *University of Cambridge Local Examinations Syndicate*

---

11:50-3:00 **Lunch and visit to the St. Louis Arch and Museum**

**Language Testing Editorial Advisory Board Meeting**
Pavilion Ballroom

---

3:00-5:00 **Symposium 2**
Organizer: Sarah Briggs, *University of Michigan*

*Using Language Corpora in Language Testing*

---

5:00-5:20 **Break**

**Group photograph**
East Foyer
### Thursday, February 22

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:20-6:50</td>
<td><strong>Works in Progress</strong>&lt;br&gt;Chair: Liz Hamp-Lyons, <em>Hong Kong Polytechnic University</em></td>
<td>Pavilion Ballroom</td>
</tr>
<tr>
<td>5:20-6:05</td>
<td>Introduction to Works in Progress</td>
<td></td>
</tr>
<tr>
<td>6:05-6:50</td>
<td>Viewing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. The Effect of Time-compressed Speech Training on Listening Comprehension&lt;br&gt;Martyn Clark, <em>University of Hawaii</em>, and John Clark, <em>Defense Language Institute</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Towards a Common Performance Description Scale for Non-native Speaker Writing&lt;br&gt;Roger Hawkey, Nick Saville, and Chris Banks, <em>University of Cambridge Local Examinations Syndicate</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Is Speaking Performance Assessment Based Mainly on Grammar?&lt;br&gt;Tom Lumley and David Qian, <em>Hong Kong Polytechnic University</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Development of a Web-based Listening and Reading Test for Less Commonly Taught Languages&lt;br&gt;Margaret Malone, Helen Carpenter, Paula Winke and Dorry Kenyon, <em>Center for Applied Linguistics</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Differential Item Functioning in Language Tests and its Implication to Second/Foreign Language Learning and Teaching&lt;br&gt;Tai-Ill Pae, Susan Maller, David O'Brien, and Beverly Cox, <em>Purdue University</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6. Is There a Place for Vocabulary Knowledge in TOEFL Reading Assessment?&lt;br&gt;David Qian, <em>Hong Kong Polytechnic University</em>, and Mary Schedl, <em>Educational Testing Service</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7. How Examinees Take Conventional Versus Web-based Japanese Reading Tests&lt;br&gt;Yasuyo Sawaki, <em>University of California at Los Angeles</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8. Designing Validation Research for an ESL Placement Exam Outline Completion Task&lt;br&gt;Viphavee Vongpumivitch, <em>University of California at Los Angeles</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9. Establishing Comparability and Stability of Cut Scores Across Forms&lt;br&gt;Amy Yamashiro, Barbara Dcbson, and Mary Spaan, <em>University of Michigan</em></td>
<td></td>
</tr>
</tbody>
</table>

| 7:00-8:30 | **LTRC/ILTA Business Meeting**                                                            | Pavilion Ballroom  |