20th Annual
Language Testing Research
Colloquium

Theme: Interrelationships Between
Language Pedagogy and Language Testing

Program and Abstracts

Monterey Holiday Inn Resort
Monterey, California, USA
March 9-12, 1998
Acknowledgments

LTRC 98 Steering Committee

Dariush Hooshmand, LTRC '98 Program Chair
Antony John Kunnan, LTRC '97 Program Chair
Mary C. Spaan, LTRC '97 Program Chair

Organizations

California State University, Los Angeles
Defense Language Institute, Foreign Language Center
Educational Testing Service
English Language Institute, University of Michigan, Ann Arbor
International Language Testing Association
Teachers College, Columbia University
Lancaster University
University of Surrey

Individuals

Lyle Bachman, University of California, Los Angeles
Caroline Clapham, Lancaster University
Glenn Fulcher, University of Surrey
Adrian Palmer, University of Utah
James E. Purpura, Teachers College, Columbia University

Volunteers

Christine Campbell, Defense Language Institute, Foreign Language Center
Barbara Evans, Defense Language Institute, Foreign Language Center
Ba-Nhon Le, Defense Language Institute, Foreign Language Center
Delia Lugo, Defense Language Institute, Foreign Language Center
Margaret Van Daalen, Defense Language Institute, Foreign Language Center
Carrie Wolf, Teachers College, Columbia University
Acknowledgments

Abstract Evaluators

Lyle Bachman, University of California, Los Angeles, USA
Kathleen M. Bailey, Monterey Institute of International Studies, USA
Carol Chappelle, Iowa State University, USA
Caroline Clapham, Lancaster University, UK
John L. D. Clark, Defense Language Institute, Foreign Language Center, USA
Fred Davidson, University of Illinois, Urbana - Champaign, USA
John H. A. L. de Jong, Dutch National Institute for Educational Measurement, the Netherlands
Dariush Hooshmand, Defense Language Institute, Foreign Language Center, USA
Christine Jensen, University of Kansas, USA
Dorry Kenyon, Center for Applied Linguistics, USA
Antony John Kunnan, California State University, Los Angeles, USA
Adrian Palmer, University of Utah, USA
Kyle Perkins, University of Illinois, Carbondale, USA
James E. Purpura, Teachers College, Columbia University, USA
Mary C. Spaan, University of Michigan, Ann Arbor, USA
Jean Turner, Monterey Institute of International Studies, USA
PROGRAM
MONDAY, MARCH 9

9:00 - 5:00  On-Site Registration
Second Floor Foyer outside Presidio Room

1:00 - 5:00  Pre-Colloquium Workshop
Defense Language Institute, Foreign Language Center

An Introduction to Structural Equation Modeling with EQS

Presenter:

James E. Purpura, Teachers College, Columbia University
With assistance from: Anthony Kunnan, California State University, Los Angeles

6:00 - 8:00  Welcoming Reception
Big Sur Room

Co-hosted by the Defense Language Institute, Foreign Language Center (DLIFLC) and the International Language Testing Association (ILTA)
PROGRAM
Tuesday, March 10

8:00 - 5:00  On-Site Registration
Second Floor Foyer outside Presidio Room

8:30 - 10:00  Opening Plenary
Presidio Room

Welcome

COL Daniel D. Devlin, Commandant
Defense Language Institute, Foreign Language Center (DLIFLC)

Dariush Hooshmand, LTRC '98 Program Chair, DLIFLC

Introduction of Plenary Speaker, Martha Herzog, DLIFLC

9:00 -10:00  Responding to Demands for “Better, Faster, and Cheaper” Testing
Ray T. Clifford, Provost, DLIFLC

10:00 - 10:15  Break

10:15 - 12:00  Session 1 - Papers
Presidio Room

Chair: Caroline Clapham, Lancaster University

10:15 - 11:00  Detecting and Evaluating the Impact of
Multidimensionality in Language Test Data
R. J. Adams, Australian Council for Educational Research
Tim F. McNamara, University of Melbourne
Susan Zamit, Australian Council for Educational Research

11:00 - 11:45  Does item discrimination make a difference?
Randy Thrasher, International Christian University
Tuesday, March 10

11:45 - 12:30  The Development of Task-Dependent and Task-Independent Scales in Performance Assessment
Thom Hudson, James Dean Brown, & John M. Morris, *University of Hawaii at Manoa*

12:30 - 1:30  Lunch Break

1:30 - 4:45  Session 2 - Papers
Presidio Room

**Chair:** Jean Turner, *Monterey Institute of International Studies*

1:30 - 2:15  Using Multimedia to Create Interactive Assessments
Louis Mang, *Educational Testing Service*
April Ginther, *Purdue University*

2:15 - 3:00  Validation of an Automated Spoken Language Test
Jared Bernstein, *Ordinate Corporation*
Eduardo Cascallo, *Regents College*

3:00 - 3:15  Break

3:15 - 4:00  TOEFL CBT Listening Comprehension: An Analysis of LC Minilink Stimulus and Item Characteristics
Mario Yepes-Baraya & Jean V. Yepes, *Educational Testing Service*

4:00 - 4:45  Adapting the Recall Protocol Method for Classroom Testing
Christine E. B. Moritz, *University of South Carolina*

5:00 - 7:00  Meeting: Editorial Advisory Board for the Journal of Language Testing
Peninsula Room

5:15 - 8:30  Tour of Cannery Row and Carmel by the Sea
PROGRAM
Wednesday, 11 March

Chair: John H. A. L. De Jong, CITO, Dutch National Institute for Educational Measurement

8:30 - 10:30 Session 3 - Colloquium/Papers
Presidio Room

Language Assessment in Education in Hong Kong

Organizer:
Lyle F. Bachman, University of California, Los Angeles

Presenters:
Lyle F. Bachman, University of California, Los Angeles
Liz Hamp Lyons, Hong Kong Polytechnic University
Tom Lumley, University of Melbourne
Lynda Chapple, The Chinese University of Hong Kong
Kathie Bailey, CSLEL, Monterey Institute of International Studies
Sonya Saunders, The Chinese University of Hong Kong
Lifying Cheng, Open University of Hong Kong
Howard Sou, TOC Assessment Unit
Peter Falvey, University of Hong Kong
David Coniam, The Chinese University of Hong Kong

10:30 - 10:45 Break

10:45 - 5:00 Book Exhibits
Big Sur Room

10:45 - 11:30 Validating a Topic-based Test of Language Proficiency
Janna Fox, Carleton University
Martha Jennings, Carleton University
Barbara Graves, Carleton University
Elana Shohamy, Tel Aviv University

11:30 - 12:15 Mapping Rates of Progress in Proficiency
Tony Lee, Elaine Wylie & David Ingram
Griffith University
Wednesday, 11 March

12:15 Group Photograph

12:15 - 1:30 Lunch Break

1:30 - 4:14 Session 4 - Papers/Posters
Presidio Room

Chair: Eduardo Cascallo, Regents College

1:30 - 2:15 Analytic vs. Holistic Rating of an ESL Composition Placement Test: A Validation Study
Cynthia Taskessen, University of California, Los Angeles

2:15 - 3:00 Task Difficulty in ESL Writing Assessment
Geoff Brindley & Gillian Wigginsworth, Macquarie University

3:00 - 3:30 Break

3:30 - 4:15 Investigating test impact: public perceptions of basic skills tests and their ethical implications
Cathie Elder & Brian Lynch, University of Melbourne

4:15 - 5:15 Introduction to Works in Progress

1. The Development of Second Language Graduation Proficiency Tests in Reading and Writing
Cheryl Alcaya, Melody Jacobs-Cassuto, Marcos Holzner, Ursula Lentz, Gabriala Sweet, & Dan Reed, University of Minnesota

2. Problems Associated with Setting Language Admission Standards for Teachers Wanting to Teach English or French as a Second Language
Doreen Bayliss, University of Ottawa

3. Towards a Multi-lingual Framework: the Can Do Statements Project
Sabille Bolton & Neil Jones, University of Cambridge

4. Evaluation of the Dimensionality of a Free Writing Task and Multiple-Choice Questions for ESL Writing Assessment
Yeonsuk Cho, University of Illinois, Urbana-Champaign

5. Curriculum/Testing Revision-A Synergic Methodology
Christa Hansen & Christine Jensen, University of Kansas
Wednesday, 11 March

6. Assessing Speaking Ability: Two Aspects (Linguistic and Pragmatic) from Three Dimensions (Monologue, Dialogue, and Multilogue)  
Uji Nakamura, Tokyo Keizai University

7. EAP: What is Important in Testing Reading  
Liliana Fortuny, Martha Botto de Pocovi, Silvia Sastre, & Susana Briones  
Universidad Nacional de Salta, Argentina

8. Nationality Group Differences in the ESL Placement Test and Students’ Satisfaction over Placement. Kyong Hyon Pyo, University of Illinois, Urbana-Champaign

9. Creating an Assessment instrument for ESL Conversation  
Diane Steele, University of Illinois, Urbana-Champaign

10. A Cognitive Perspective: Validation of Test Performance as a Measure of Communicative Competence  
Minako Yamada, The Pennsylvania State University

5:15 - 6:00  
Session 4 - Papers/Posters (Continued)  
Poster Review, Vista Room

6:00 - 7:00  
Business Meeting and no-host Bar  
Peninsula Room
**PROGRAM**  
**Thursday, 12 March**

**Chair:** Elana Shohamy, Tel Aviv University

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 - 10:00</td>
<td>Session 5 - Invited Speaker</td>
<td>Presidio Room</td>
</tr>
</tbody>
</table>

**Language Factors in Educational and Mental Testing**

John Oller, *University of Southern Louisiana*

**10:00 - 10:15 Break**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:15 - 12:30</td>
<td>Session 5 - Papers</td>
<td>Presidio Room</td>
</tr>
</tbody>
</table>

10:15 - 11:00  Setting up a Dynamic Language Testing System in National Language Test Reform: The Public English Test System in China  
Liang Yuming & Liu Quingsi, *National Education Examinations Authority*  
Michael Milanovic & Lynda Taylor, *University of Cambridge Local Examinations Syndicate*

11:00 - 11:45  The Comparability of Second and Foreign Language Testing Programs  

11:45 - 12:30  Co-construction in Test Discourse: Interactional Authenticity and Pedagogical Implications  
Carol Lynn Moder & Gene B. Halleck, *Oklahoma State University*

**12:30 - 1:30 Lunch Break**
Thursday, 12 March

1:30 - 6:15  
Session 6 - Papers/Posters
Presidio Room

Chair: Dorry Kenyon, Center for Applied Linguistics

1:30 - 2:15  
The Effect of Candidate Acquaintanceship on OPI Performance
Barry O’Sullivan, Okoyama University
Don Porter, The University of Reading

2:15 - 3:00  
Interview Style and Candidate Performance in the IELTS Oral Interview
Annie Brown, University of Melbourne

3:00 - 3:15  
Break

3:15 - 4:00  
Rating Scales and Rater Agreement in Testing Second Language Oral Proficiency
Helen A. Van Berlo & John H. A. L. de Jong
CITO, Dutch National Institute for Educational Measurement

4:00 - 4:45  
Describing the Discourse of Scale Developers: A Closer Look at the Process of Empirically Derived Scales
Carolyn Turner, McGill University

4:45 - 5:15  
Session 6- Papers/Posters
Introduction to Posters, Presidio Room

1. The Local Item Banking System (LIBS)-a PC-based Item Bank for Developing Language Tests. Simon Beeston, University of Cambridge Local Examinations Syndicate


3. Assessment as a Critical Component of a Successful Articulation Project
Micheline Chalhoub-Deville, University of Iowa

Neil Jones, University of Cambridge Local Examinations Syndicate
Thursday, 12 March

   Beryl E. Meiron, California Institute of Technology (Caltech) & California State University, Los Angeles

6. Investigating the Impact of IELTS
   Michael Millanovic, Nick Saville, University of Cambridge Local Examinations Syndicate

7. The Development of English Language Tests for Young Learners (age 7-12)
   Nick Saville, University of Cambridge Local Examinations Syndicate

8. A Study on the Role of the TOEIC IP
   Ishikawa Shoichi, National Defense Academy, Japan

9. Providing Diagnostic Information for Understanding Students’ Learning Processes
   Naoki Taira, The National Center for University Entrance Examinations, Japan
   Hisami Saito-Scott, University of Illinois, Urbana-Champaign
   Masahiro Kasai, State of Florida, Department of Business and Professional Regulations, Bureau of Testing

5:30 - 6:00     Session 6 - Paper/Posters (Continued)
                  Poster Review, Vista Room

6:00 - 6:30     Colloquium Summation
                  Presidio Room

   Caroline Clapham, Lancaster University
   Fred Davidson, University of Illinois, Urbana - Champaign
   Alan Davies, University of Melbourne
   Mary C. Spaan, University of Michigan, Ann Arbor

7:00 - 10:00    Banquet / ILTA & ETS Awards
                  Big Sur Room

Coordinator: Antony John Kunnan, California State University
            Award Recipients: To be announced
ABSTRACTS

Abstracts are listed in the following order: Plenary, Invited speaker, Pre-Colloquium Workshop, Colloquium, Papers, Works in Progress, and Posters.

PLENARY

Responding to Demands for “Better, Faster, and Cheaper” Testing

Ray T. Clifford, Provost, Defense Language Institute Foreign Language Center

The field of language testing is subject to the same market and budgetary forces as the fields of language teaching and general education. Some current manifestations of these forces are:
1. Recognition of the need to validate the skills of individuals with high levels of language competence.
2. Reductions in testing time and a push for “just-in-time” assessment.
3. Shrinking budgets for both development and testing operations.

How we respond to these pressures will have a lasting impact on the credibility and survivability of the language testing profession.

Ray T. Clifford received B.A. and M.A. degrees in German from Brigham Young University and a Ph.D. in second language education from the University of Minnesota. He has taught German at elementary, high school, and college levels, as well as preservice and in-service courses for foreign language teachers. Since 1978 he has been involved in foreign language program administration and is currently Provost of the Defense Language Institute in Monterey, California.
INVITED SPEAKER

Language Factors in Educational and Mental Testing

John Oller: University of Southwestern Louisiana

Recent advances in the general theory of signs, especially in the theory of true narrative representations (TNR-theory) and in the theory of abstraction (A-theory) have shown that many mental tests are essentially dependent on conventional signs in a way that was only partially understood before. In particular, nonverbal IQ tests and their derivatives (the performance inventories of various sorts) are dependent on conventional signs in the defined way. As a result, it is now possible to offer a coherent explanation for the discrepancies between VIQ and NVIQ scores in both monolingual and bilingual cases, and the meaning of intelligence test scores across the board, unsurprisingly, is revised. Experimental demonstrations of the necessity for this adjustment, predicted by the general theory of signs (especially the two components of it mentioned), are also discussed. The upshot is that language factors play a more central role in mental tests than might have been expected prior to the theoretical developments in question.

An elected member of the New York Academy of Sciences, John Oller (a native New Mexican and bilingual Hispanic, "Oller" is a Catalan name meaning "potter") is an experimentalist and measurement specialist in language, literacy, communication, and intelligence. He is author or co-author of 12 books and 192 articles on these and related issues. His most recent work concerns the hierarchy of sign systems used by human beings—how it develops and how it can stall or break down, for instance, in autism.

PRE-COLLOQUIUM WORKSHOP

An Introduction to Structural Modeling Using EQS

James E. Purpura: Teachers College, Columbia University

Structural equation modeling (SEM) is a multivariate, analytic procedure used for specifying, estimating and testing hypothesized inter-relationships among a set of substantively meaningful variables. More specifically, this methodology allows researchers to posit and test hypothesized models of inter-relationships (1) between observed variables and constructs and (2) among constructs based on substantive theory or previous empirical research.

The use of SEM in second language (SL) research is not a new phenomenon. In the early 1980’s, Purcell (1983) used SEM to investigate the effects of predictor variables on SL pronunciation accuracy, while Gardner (1983, 1985) used it to examine the effects of attitudes and motivation on SL proficiency. However, in recent years, SEM has become increasingly popular among SLA and SL assessment researchers for investigating models of SL acquisition or performance (e.g., Brindley & Ross, 1997; Clement & Knudsen, 1985; Gardner et. al. 1987; Kunneman, 1991; Nelson, Lomax & Perlman, 1984; Purpura, 1996, 1997; Sasaki, 1991). This is perhaps due to the availability of a number of user-friendly programs, which provide a relatively simple, yet technically sophisticated means of performing a wide range of multivariate analyses. One such program is EQS, a leading SEM program that has earned a reputation for its ease of use and practical features.
The purpose of the current workshop is to provide participants with a hands-on introduction to the use of EQS in performing a number of statistical procedures from computing basic, descriptive statistics to doing SEM. In the first part of the workshop, I will demonstrate how to create data sets in EQS, how to import data, and how to generate plots. I will then demonstrate how to compute basics statistics (descriptives and frequencies) and how to perform exploratory factor analysis. Participants will then be given a computer task to become familiar with these basic procedures.

In the second part of the workshop, I will present a non-mathematical introduction to the basic concepts of SEM and will demonstrate how to perform a confirmatory factor analysis, a special case of SEM, on EQS. Participants will then perform a similar analysis and we will discuss the output.

Finally, I will demonstrate how to generate a full latent variable model on EQS and we will again discuss the output. Participants will then perform these analyses on the computer. For purposes of illustration and discussion, I will provide a data set to work from; however, all participants are encouraged to bring their own data to the workshop and to work with their own data. Both MACs and IBMs will be available for use.

James E. Purpura is Assistant Professor of linguistics and education in the TESOL and Applied Linguistics programs at Teachers College, Columbia University. He teaches courses in second language assessment, second language acquisition, pedagogical grammar and discourse analysis. He serves on the Test of Spoken English Committee at ETS and has worked on a number of projects with the English as a foreign language division of UCLES. Prior to his current position, he served as Coordinator of the ESL Placement Exam at UCLA. Jim's most recent book is entitled Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeling Approach, to appear in the Studies in Language Testing Series with Cambridge University Press.
Relationships Among Course Objectives, Self-assessment, and Achievement in a Learning Strategies Course

Kathi Bailey, Monterey Institute of International Studies  
Sonya Saunders, Chinese University of Hong Kong

At the Chinese University of Hong Kong, a twenty-item self-assessment questionnaire (using a five-point Lickert scale format) was developed on the basis of existing course objectives for an intermediate EFL learning strategies course with an emphasis on speaking and listening. Students completed this questionnaire and also took a commercially developed listening test as pre- and post-course measures. The group means showed statistically significant improvement on the commercially developed test and each item on the self-assessment questionnaire. We will discuss the relationship between the test results and the self-assessment results, and discuss issues related to trying to introduce a self-assessment mechanism in the Hong Kong context.

Developing a Video-based Test of Listening Strategies

Lynda Chapple, Chinese University of Hong Kong  
Kathi Bailey, Monterey Institute of International Studies

At the Chinese University of Hong Kong we offer an elective course in English listening and speaking strategy training. A major objective of the course is that students will develop a range of listening strategies for understanding native speaker speech, including the ability to deal with speed, accent, and tone.

As part of a larger assessment instrument, we have been developing a video-based listening subtest to evaluate students' acquisition of such listening strategies. The test specifically targets students' ability to identify key points, to extrapolate meaning from tone, and to infer meaning of unfamiliar or idiomatic vocabulary from context. Video clips seemed to provide a useful medium as they contain many of the contextual and visual clues that exist in ‘real life’ situations. Video also provided an opportunity for us to experiment with test items which required a more holistic understanding of natural spoken English, thus requiring students to use a broader range of strategies than discrete-point items.

In this presentation, we will discuss the rationale for the subtest and the test development process, including item analyses and reliability indices. We will also briefly discuss students' responses to this approach to testing, and the positive washback for teaching and learning.

The Washback Effect of Public Examination Change on Classroom Teaching

Liying Cheng, Open University of Hong Kong

Public examinations have always been used as instruments and targets of control in the school system. A belief that assessment can leverage educational change and bring positive washback on teaching has often led to top-down educational reforms. This is the exact case in Hong Kong, where major changes
were introduced into the existing Hong Kong Certificate of Education Examination in the English subject with the intention of bringing about a positive washback effect on classroom teaching so that what happens in the examination room can be relevant to what happens in the real world. An investigation into the washback effect on classroom teaching has demonstrated the impact of the examination change on classroom behaviours such as teacher talk, time allocation of classroom activities, and student practice opportunities. However, the extent of washback on these behaviours varies within teachers as well as between teachers. Washback on classroom teaching behaviours is seen to occur slowly, and is limited. Only a combined effort of teacher education and material development, together with support from all parties within the educational system, can realize the belief in washback and achieve the task of a genuine change in classroom teaching.

Setting Language Benchmarks for English Language Teachers in Hong Kong Secondary Schools

Peter Falvey, The University of Hong Kong
David Coniam, The Chinese University of Hong Kong

The work described in this presentation arises from the 1995 Hong Kong Government Education Commission Report Number 6 (ECR6), which recommended that minimum levels of language and professional competence be established for all language teachers. The presenters were commissioned by the Hong Kong Government in order to make recommendations for the establishment of language benchmarks.

The presentation describes the process involved in establishing English language benchmarks for teachers of English in Hong Kong. The presenters will describe the background to the project, the survey instrument that was used to gauge teacher reaction to the issue, existing and new test types that were developed in order to initially assess teacher competence, the prototype specifications that were created for the formal establishment of assessment instruments, and current developments which include piloting the instruments in the summer of 1998.

Expectations of Exit Language Proficiency of University Graduates in Hong Kong

Liz Hamp-Lyons, Hong Kong Polytechnic University
Tom Lumley, Hong Kong Polytechnic University

Hong Kong has been thought of as an ESL environment, but in fact it has always effectively been an EFL environment for all but the privileged minority (Pennington, 1992). As the university system has expanded in recent years, it has become more difficult to admit only students with the stated levels of English on the key high school exit test, the British-based HKALE ("A" level) Use of English, and there to maintain the universities as truly English-medium institutions. Claims have become common among university staff and local employers (see, for example, Leung, 1997; Choi, 1997 and other papers in the Building Hong Kong on Education Symposium held at Lingnan College in June 1997, as well as many editorials and 'feature' articles in the Hong Kong press) that university students' and graduates' standards of English are declining.
The Graduating Students’ Language Proficiency Assessment (English strand) project was set up with University Grants Council funding to explore the feasibility of a Hong Kong-wide assessment of the English language proficiency of first degree level students at graduation. It has now been adopted, in an operational form, as the exit English proficiency measure by two of the seven degree-awarding institutions in Hong Kong, the Polytechnic University, and Lingnan College. This paper reports the third and final phase of the project, in which the assessment developed for the funding body was adapted to meet the needs of an operational assessment.

Hong Kong is an extremely language-aware social context, and a survey of potential score consumers showed clear views about what the test should be like and what it should achieve. The paper discusses the ways in which the original test design was adapted to provide more of what the users want; it also considers the processes and problems of responding to the sometimes-conflicting expectations of university faculty and of potential employers of the graduates. The process of accommodating such influences has led to some interesting test developments.
Validation of an Automated Spoken Language Test.
Jared Bernstein, Ordinate Corporation
Eduardo Cascallar, Regents College
John de Jong, CITO, Dutch Institute for Educational Measurement

Background. Oral language assessment traditionally requires significant examiner time per candidate. Preliminary reports have suggested the feasibility of automatic systems for evaluating and scoring some aspects of spoken language.

Research Goal. This research seeks validation for a fully automatic prototype test of oral communication skills. The study also aims to establish criterion referenced score transformation and to estimate reliability and precision of scores with reference to traditional measures.

Design and Method. Data are gathered from many samples (typical size 50-150) of native and non-native speakers of English from North America, Europe, and Asia. The overall total will comprise approximately 4000 subjects with 10% native speakers. Tests are administered by telephone and yield both machine-and human-generated scores. For some samples, data are gathered concurrently from regular tests of speaking and/or listening such as OPI and standardized examinations. Data analyses aim at estimating concurrent validity for machine vs. regular measures for subjects sampled from within a number of particular institutional settings. Concurrent measures of machine-human correspondence over the whole set of 4000 subjects on test-internal material will be reported. In addition, IRT analysis will be used to evaluate the assumption of essential unidimensionality across measures.

Results. Preliminary results show that concurrent validity varies across samples, but refinement of scoring algorithms (during August-November 1997) should yield higher and more consistent validity estimates.

Implications. Convenient testing with on-line scoring may increase focus on oral skills and partly relieve teachers and assessors of a time-intensive task.
Task Difficulty in ESL Writing Assessment. Geoff Brindley and Gillian Wigglesworth, Macquarie University

Educational authorities around the world are increasingly adopting assessment and reporting systems which use teacher-conducted assessments rather than standardized tests as the basis for reporting outcomes of language programs. However, since the tasks used to assess the same learning outcome may vary across teachers and sites, there is the potential for considerable variability in task difficulty, creating a situation which may disadvantage some learners. It is therefore important for teachers to be aware of those factors that may contribute to task difficulty so that these can be considered when assessment tasks are constructed.

The paper reports on a research study into the question of task difficulty in ESL writing. First, ratings of writing performances by forty students on each of three tasks aimed at assessing two different competencies were analyzed using generalizability theory and many-facet Rasch analysis in order to investigate task comparability and to examine the effects on dependability of using different numbers of tasks and raters. The second phase of the study investigated the effects on task difficulty of manipulating a range of variables, such as the amount of structure provided and the type of pre-task preparation allowed.

Analysis of results revealed a number of specific task characteristics contributing to task difficulty. These will be described and some of the major implications for teaching and assessment practice will be outlined. Practical suggestions will be made as to how these factors might be taken into account in assessment task design.
Interviewer Style and Candidate Performance in the IELTS Oral Interview. Annie Brown, The University of Melbourne

Aims. Research involving oral language tests has begun to investigate the discourse produced in these interviews. However, most research focuses on what is common across interviews rather than the extent to which they differ. While some studies have indicated that, despite training, there may be substantial variation in interviewer behaviour, little is known about the relationship between interviewer behaviour and candidate performance.

This paper reports on a study which investigated:
(a) the extent to which differential behaviour by IELTS interviewers affected the scores awarded to candidates, and
(b) the interviewing styles of interviewers identified as being "difficult" or "easy" interviewers.

Methodology. Thirty-two students from IELTS preparation courses and six accredited interviewers participated in the study. Each of the thirty-two candidates was interviewed twice by two different interviewers and each tape was later rated by four accredited IELTS raters.

The test data were analyzed using the multi-faceted Rasch analysis program FACETS (Linacre, 1989) in order to identify cases where candidates perform differentially in the two interviews, as well as to identify interviewers who consistently elicit poorer or better performance.

In stage two of the study, four pairs of interviews where the same candidate performed at different levels were transcribed. An analysis was then undertaken in order to identify whether there are particular interviewing styles which are representative of "easy" or "difficult" interviewers and which may contribute to better or worse performance by candidates. This paper presents some findings from the analysis.

Implications. The findings of this study will contribute towards ensuring fairness in assessment for all candidates through their relevance to the training of interviewers, in increasing interviewers' understanding of the effect of their performance on that of the candidate and in ensuring comparability of the challenge presented to different candidates.

Reference

Investigating Test Impact: Public Perceptions of Basic Skills Tests and Their Ethical Implications.
Cathie Elder and Brian Lynch, University of Melbourne

Recent work on language testing (Messick 1993, Moss 1992, Gipps 1994) has stressed the political nature of the testing enterprise and the need for an expanded view of test validity including evidence of a test’s impact on those involved, directly or indirectly, in the testing process.

This paper extends the scope of test validation into the social consequences of test use as instanced by public views of the controversial Learning Assessment Project (LAP) introduced recently in the Australian State of Victoria. The LAP tests are measures of school achievement in basic literacy, numeracy, and other key learning areas administered annually to Victorian school children in Years 3 and 5. The impact of this testing enterprise was assessed through:

1. Interviews with bureaucrats & other key players in the tests’ implementation
2. questionnaires administered to a representative sample of primary teachers
3. in-depth case studies involving teachers, students, and parents at two different schools - one favourable to the LAP and the other which boycotted the test administration
4. a survey of media reportage of the LAP.

Data were coded thematically using event, time-ordered, and case dynamics matrices. Findings revealed considerable diversity in public perceptions of the LAP project including some quite nuanced reactions (e.g. some respondents objected not to testing per se but to the potential misuse of LAP test results to create “league tables” of schools). The paper concludes with a discussion of the implications of this study for a) conceptualizations of test validity b) the formulation and implementation of educational policies.
Validating a Topic-based Test of Language Proficiency.
Janna Fox, Martha Jennings & Barbara Graves, Carleton University
Elana Shohamy, Tel Aviv University

Many topic-based tests of language proficiency used in university admission procedures (CAEL, IELTS, OTESL) claim pedagogical validity by reflecting language use in academic domains. One possible threat to this validity is test bias in that familiarity with the test topic and background knowledge may influence test-taker performance. The effect of topic on language test performance has been researched with reference to background knowledge operationalized as domain of study, declared subject major, and preteaching (Alderson & Urquhart, 1988; Clapham, 1995; and Fox et al, in press). While background knowledge, thus defined, is one factor influencing test-takers’ performance, the present study investigates additional factors in the interaction between the test-taker and the topic of the test. A sample of 250 ESL university applicants taking a topic-based test of language proficiency were randomly assigned to one of two conditions: no choice of topic vs. choice among five topics representing various fields (science, engineering, social sciences). Data from their test scores and results of interviews and questionnaires are analyzed to investigate topic-effect on performance under the two conditions. Variables which influence the relationship between topic and test-taker performance including level of language proficiency and test version are examined. Verbal protocol analysis is used to examine test-taker perceptions of the factors influencing their choice of topic. Implications of the study including the involvement of the test-taker as a stakeholder and pragmatic considerations for test administration and development will be discussed.
Using Multimedia to Create Interactive Assessments.
Louis Mang, Educational Testing Service
April Ginther, Purdue University

This presentation is a demonstration of an interactive multimedia prototype designed to test listening and speaking in an integrated manner. Our intention was to create an assessment in which interactivity and context are central rather than secondary aspects of the task. The project was funded by Educational Testing Service as part of its effort to explore the potential of multimedia in computer-based testing.

The prototype is based on a simulation which requires subjects to “interact” with a number of different people in order to accomplish a series of specific objectives. In order to successfully complete the simulation, subjects are required to understand spoken English and respond to spoken statements and questions. Subjects must assume an active role, produce requests appropriate to the presented contexts, and use language to solve immediate problems. The statements and questions within the simulation are designed to elicit spoken responses to basic conversational routines, and the exchanges between the subjects and the characters take place in real time. Subject responses are recorded for later analysis or scoring.

The prototype will be presented, followed by a series of recorded subject responses. Results of a pilot study based on the responses of 20 subjects will be summarized. Typical response patterns, the development of scoring rubrics, and program design will be discussed. The prototype is intended as a point of departure for research and development of future item types.

Background. The council of Europe has proposed a common framework for language learning, teaching, and assessment (Second draft: 1997) which offers a criterion referenced approach to language testing. Migrants should therefore have equal access to rights related to particular levels of mastery of languages of host countries, whether acquired as foreign or as second language learners.

Purpose of the research. This research investigated the comparability of two testing programs, one measuring foreign, the other second language ability. Both programs provide separate tests for the four skills, each tested at a number of levels. The research aimed to assess whether in both programs (1) the same trait underlies the corresponding tests and (2) the same criterion underlies the pass/fail decision on those tests.

Research design and method. For the eight comparisons (four skills, two levels) data were gathered from candidates in both testing programs on tests containing items from both programs. The data were fitted to two unidimensional IRT models, the one-parameter Rasch model and a model allowing varying discrimination parameters. Whenever IRT-modeling was not feasible, linear regression models were applied.

Results. The unidimensionality assumption was rejected applying the Rasch model, but could not be rejected for five comparisons when inputting varying discrimination parameters. Regression models were rejected for one test pair only. For the seven pairs where unidimensionality was not rejected, pass/fail decisions did not differ statistically.

Implications. The results have lead to mutual recognition of certificates from both programs. In addition, both testing programs have initiated revisions of scoring procedures and one program has adopted the equating procedures already operational in the other program.
Mapping Rates of Progress in Proficiency. Tony Lee, Elaine Wylie, and David Ingram, Griffith University

For language students and those who need their skills, and for agencies which fund language programs and institutions which provide those programs, it is important to know how long it takes to reach desired levels of proficiency. This paper reports on a project funded by the Australian Government. The project aims to provide data on how long it takes particular types of learners of English to progress through particular proficiency levels in particular types of learning situations. Data on over 32,000 adult learners were obtained from five different sources, mainly government-funded programs and ELICOS (English Language Intensive Courses for Overseas Students) programs. All proficiency scores were provided in terms of the International Second Language Proficiency Ratings (ISLPR), formerly the ASLPR. To estimate rates of progress a two-wave path model within the LISREL family of models was used. Entry and exit latent proficiency levels were estimated from observed ISLPR ratings for the sample of learners. The two proficiency factors were then entered into a path model with other available critical data (e.g., L1 background, age, and type and length of courses attended). The two-wave path model estimated the increase (positive or negative) between course entry and exit after adjusting for the other (situational) variables in the model, including possible auto-regressive processes. Investigation of the possible influence of variables not included in the model (including affective factors such as learner motivation) by means of case studies will constitute a second phase of the project.
Detecting and Evaluating the Impact of Multidimensionality in Language Test Data.
R. J. Adams & Susan Zammit, Australian Council for Educational Research
T.F. McNamara, University of Melbourne

The question of dimensionality in language test data has been the subject of debate for many years, and has usually involved discussion of the clash between construct and measurement conceptions of dimensionality, particularly in the context of IRT-based approaches to the analysis of test data. In this paper, a number of recent empirical techniques, Rasch-based and non-Rasch based, for the detection of multidimensionality in language test data are compared. These include the factor analysis of matrices of tetrachoric correlations, the use of so-called full information factor analysis and multidimensional item response models. The investigation considers the issue of whether different detection methods produce different answers to the question of whether the data display multidimensionality, and explores the assumptions underlying each of these techniques.

From a practical point of view, we have little understanding of the impact of multidimensionality on the estimates of candidate ability resulting when any such multidimensionality is not modeled, which is the case with most standard analytic techniques, particularly IRT ones. We investigate this question by comparing ability estimates from a pair of data analyses in which multidimensionality is and is not modeled in the analysis. The data used are performances by secondary school students on national tests of listening and reading comprehension in two foreign/second languages.
Co-construction in Test Disclosure: International Authenticity and Pedagogical Implications.
Carol Lynn Moder & Gene B. Halleck, Oklahoma State University

In discourse research, investigators have increasingly focused on the co-construction of discourse and the linguistic features which it engages (Schiffrin, 1987, 1994; Sacks, Schegloff, & Jefferson, 1974; Ochs, Schegloff & Thompson, 1996). The findings of such studies are increasingly being incorporated into communicative classroom settings but have been incorporated more slowly into testing practice. For example, McNamara (1996) observes that although current theories of language knowledge which are widely used in testing, such as that of Bachman (1990, 1991) and Bachman & Palmer (1996), incorporate concepts relevant to co-constructed discourse, they focus too heavily on the individual candidate rather than on the “candidate in interaction.”

The purpose of the current study is to focus on the way in which an interview, the ACTFL OPI, and a performance test, the ITA Test (Smith, Meyers, & Burkhalter, 1992), engage the candidate-in-interaction by sampling textual and sociolinguistic abilities crucial to interaction. The analysis focuses on the interactional functions of two kinds of discourse markers: propositional coherence makers like advanced organizers, lexicalized markers, and hypotaxis and more interactional markers like well, and, but, and so.

The results indicate that there were important differences in the contribution of co-construction strategies in the two testing situations. Contrary to what test designers and candidates might expect, sensitivity to the interlocutors was more critical in the ITA Test than in the OPI. This finding has significant implications both for test designers and for the training of ITAs, who tend to view lectures as a nonreciprocal task.
Adapting the Recall Protocol Method for Classroom Testing. Christine E. B. Moritz, The University of South Carolina

In recent years, traditional techniques of measuring second language reading comprehension have been broadly criticized. Many claim that both selected-response tasks such as multiple choice questions and constructed-response tasks such as open-ended questions and cloze tests encourage only surface-level reading, or matching of vocabulary items, and result in reduced awareness of text cohesion and ability to synthesize material from a reading passage. Recall and summary protocols, on the other hand, have gained recognition in the field of reading research as truer and more effective measures of second language learners’ reading comprehension. In the classroom, however, conventional testing techniques commonly reign, in part because they are familiar, and in part because they are easy to administer and score (especially compared with the apparently complex scoring procedures used for recall protocols). This paper describes the pilot stage of an experiment to determine the feasibility of adapting the recall/summary protocol method for ongoing use as an assessment tool in a proficiency-based classroom.

Whereas a single text and recall protocol may be sufficient for a given research project, adapting this instrument for classroom testing and year-end exams entails using many different texts, of varying difficulty, and length. Procedures were developed to include text choice, identification of idea units (taking into account hierarchy of and relationships between propositions), and a reliable, efficient scoring method. Interrater reliability was calculated for both idea unit classification and protocol scoring. Ways in which the use of recall protocols for classroom testing influenced instruction and materials will also be discussed.
The Development of Task-Dependent and Task-Independent Scales in Performance Assessment.
Thom Hudson, James Dean Brown, and John M. Norris, University of Hawaii at Manoa

Curriculum-related tasks were developed systematically so that they differed in terms of 64 possible combinations of performance difficulty (based on plus or minus values for linguistic code, cognitive complexity, and communicative demand, as discussed in Norris, Brown, Hudson, & Yoshioka, in press, after Skehan, 1996). In addition, a set of rating scales was created based on task-dependent and task-independent categories. The criteria for the task-dependent categories were created in consultation with advanced language learners, language teachers, and domain experts for each individual task. These criteria for success were allowed to differ from task to task depending on the input of our consultants. The task-independent categories were created on the basis of three theoretically motivated components of inherent task difficulty as follows: code (linguistic) accuracy, cognitive adequacy, and communicative appropriacy. Data were gathered from ESL students at a wide range of proficiency levels. Their performances were scored by trained raters using the task-dependent and task-independent criteria. Analyses included descriptive statistics, one-parameter IRT item difficulty estimates, reliability estimates (intrarater, Cronbach alpha, etc.), and dependability estimates (using generalizability theory). The results are interpreted and discussed in terms of: (a) the adequacy of our original task difficulty estimates, (b) similarities and differences between task-dependent and task-independent ratings, (c) test reliability and ways to improve the consistency of measurement, (d) test validity and the relationship of our task-based test to the curriculum, and (e) the effectiveness of performance measurement for pedagogical decision making.
PAPERS

The Effect of Candidate Acquaintanceship on OPI Performance.
Barry O’Sullivan, Okayama University, Japan
Don Porter, University of Reading, UK

Renewed interest in direct performance-oriented language assessment has brought with it a growing body of research into the oral proficiency interview and its variations. While many of these studies have concerned themselves with establishing the validity of the procedure through an exploration of the language generated therein, a number of others have focused on the effect of characteristics of an interlocutor on the linguistic performance of a test candidate (Porter 1991; O’Sullivan 1995; O’Sullivan & Porter 1975/6/7; Berry 1997; Iwashita 1997).

It has been suggested in the psychology literature that the spontaneous support offered by a friend positively affects anxiety and task performance under experimental conditions (Lindner, Sarason, and Sarason 1988, Matsuzaki, Tanaka, and Kojo 1990, Sarason and Sarason 1986, and Matsuzaki, Kojo, and Tanaka 1993). The importance to language testers is clear in that where a candidate is paired with a person considered to be a friend, he or she will potentially produce a superior performance than if the partner is a stranger.

This study examines the performance of a group of Japanese female university students (N = 30) on a series of tasks performed under two conditions, once with a friend and once with a stranger. The results indicate that there is a significant and systematic difference in performance.

The paper concludes by suggesting a model of factors which affect linguistic performance. This model reflects the results reported here and in the earlier studies referred to above.
Analytic vs. Holistic Rating of an ESL Composition Placement Test: A Validation Study. Cynthia Taskessen, University of California, Los Angeles

To provide the most direct method of placement into levels of ESL writing courses, many language programs now administer direct writing tests. While there remain questions on the authenticity of the limited time and topics associated with such tests, practical considerations necessitate the one-shot, timed writing test.

Given resource constraints in the administration of direct writing tests, the validity of different rating methods, namely holistic vs. analytic, can be examined. This study reports on the comparative validity of the use and interpretation of the holistic vs. analytic methods of composition scoring. Compositions from the English as a Second Language Placement Exam at UCLA, which were rated on the former three-subscale (content, organization and language) analytic scale, were rated again using a newly-instituted holistic scale. The presentation will address technical questions of the interrater agreement and accuracy of level prediction. In addition, the study will explore whether test-taker background characteristics interact with the rating method.

But more importantly, this study will consider what Messick (1996) refers to as the “structural” and “consequential” aspects of validity. How well do the rating methods represent the construct of writing? How does the placement test contribute to the system of examinations in the writing courses?

These questions will be considered in light of Bachman and Palmer’s (1996) concept of test usefulness, investigating how the practicality, authenticity, impact and construct validity considerations contribute to the usefulness of this ESL placement test when using holistic vs. analytic scoring.
Setting up a Dynamic Language Testing System in National Language Test Reform: the Public English Test System in China.
Liang Yumin & Liu Quingsi, National Education Examinations Authority, China
Michael Milanovic & Lynda Taylor, University of Cambridge Local Examinations Syndicate, UK

The Public English Test System (PETS) development was established in response to growing concern within China over the inadequacy of standards of communication in English. The need to improve communication has been identified as a priority for the success of China’s continuing Open Door economic policy. The PETS project aims to rationalize publicly available English language tests within a 5-level framework ranging from the level of English required at Junior High School to the level required for Chinese graduates to study/work overseas. This presentation will focus on the first phase of the project and will present work related to the development of level criteria, test specifications and sample materials.

The project is of particular interest because it represents a major attempt on the part of a previously traditionally-oriented national testing system to adopt a more criterion-referenced approach which also takes into account the impact that the tests are likely to have not only in the immediate pedagogical context but also in wider Chinese society. In addition, the project provides a valuable opportunity to apply and validate a cyclical and interactive model of test development which has been proposed in recent years in which full account is taken of the many different features of the test development context.
Does Item Discrimination Make a Difference? Randy Thrasher, *International Christian University, Tokyo*

With the most recent publication of McNamara’s *Measuring Second Language Performance* and Otomo’s *Komoku oto riron nyumon* (Introduction to Item Response Theory), interest in applying the one parameter Rasch model to the analysis of language test data is growing rapidly. One of the main reasons Item Response Theory appeals to language testers is that, using IRT, it is possible to equate different forms of a test without having to give both forms to the same group of test takers. However, since the Rasch model assumes that the discrimination index for all items is the same, the possibility exists that ignoring differences in item discrimination could affect the way different forms of a test are equated.

The experiment reported here documents this drawback of the Rasch model. Using data from two forms of the multiple choice portion of The Businessman’s English Test and Assessment (BETA), the effects of using anchor items of differing discriminating power in equating the two forms were examined. The two sets of anchor items were matched for difficulty but one set contained items with lower discrimination index values and the other set contained items with higher discrimination index values. Equating of the two forms was done twice; once with each set of anchor items, and the assessments of individual test takers using the two equating processes were compared. The paper concludes with a discussion of the differences found, their significance, and suggestions for coping with the problem raised.

References:
The Businessman’s English Test and Assessment (BETA) International Language Centre, Tokyo
McNamara, Tim *Measuring Second Language Performance* Longman 1996
Otomo, Kenji *Komoku oto riron nyumon* (Introduction to Item Response Theory) Taishukan 1996
TOEFL CBT Listening Comprehension: An Analysis of LC Minitalk Stimulus and Item Characteristics. Mario Yepes-Baraya & Jean V. Yepes, *Educational Testing Service*

The minitalk is a task designed to evaluate academic listening comprehension on the Test of English as Foreign Language (TOEFL). With the recent change to a computer-based mode of delivery (TOEFL CBT), the stimulus length has been increased from 90 to (up to) 150 seconds. The purpose of this presentation is to present preliminary results of a study to investigate whether the increase in length results in a decrement in performance for memory-related reasons or whether other features of the task are the critical variables affecting listening comprehension performance. Several analyses of key features of the stimulus and questions will be presented. The goal of these analyses is to determine which features or combination of features have the most powerful effect on listening comprehension.

Forty minitalk stimuli were studied in order to determine the stimulus and item variables that account for item difficulty. Each stimulus was coded for syntactic complexity, density of information, and lexical difficulty. Each of the corresponding items was coded for type of information requested, type of match and plausibility of distractors. Three main types of analyses were performed: (1) analysis of the text used for the stimulus of each minitalk, (2) analysis of the questions or directives for each minitalk, and (3) a regression analysis linking item difficulty as measured by standard ETS statistical data to stimulus and item variables.

This presentation will provide valuable new information on certain features of disclosure and their effects on comprehension. This information will help test developers better predict the difficulty level of LC sets. The study will not only inform test development in foreign language LC, it will also provide information to educators about the characteristics of academic lectures that facilitate or interfere with comprehension. It has potential use for LC strategy training, and for applied linguists and educators of nonnative speakers interested in analyzing the accessibility of academic lecture input.

Language testers have increasingly advocated and practiced the derivation of rating scales from analysis of a sample of the performances to be rated (i.e., a data-based approach) (Chalhoub-Deville 1995; Fulcher 1996; Turner & Upshur 1996a). It has been proposed that scale development factors be considered as facets of method (Turner & Upshur, 1996b). Scales (and ultimately final scores) can be affected by the scale makers and the actual sample used in the scale development process. Scales constructed empirically have been illustrated in the literature along with their measurement qualities. A description of how the actual scale-making process occurs, however, has not yet been reported.

A team composed of five ESL subject-specialists and guided by two researchers worked together to create a six-level scale. Using a set of writing samples, they found salient characteristics to distinguish levels of writing ability. The purpose of this study is to describe the discourse among the team members as they interacted to “discover” a set of indicators to be used for assigning scale levels to performances.

A phenomenological approach will be used to describe themes and patterns which have been found through analysis of the data (Creswell, 1994). Procedures will include segmenting the information (Tesch, 1990), developing coding categories (Bogdan & Biklen, 1992), and generating categories (Marshall & Rossman, 1989). Particular attention will be paid to the following areas: perceptions held by individuals, individuals’ ways of thinking, evolving team process and consensus, individual and team strategies, content focus, and activity. The intent is to gain insight into the group process of empirically deriving scales for educational purposes. Data consist of audio-tapes of the complete two-day scale making process, observational notes, and retrospective commentary on the audio-tapes by the participants.

The Ministry of Education sponsored the development of empirically-derived rating scales for the ESL leaving exam at the secondary level. A major concern was that current scales did not reflect pedagogical practice. Writing samples for scale development were collected from leaving exams.

As the field of language testing has come to rely more upon ratings of examinees’ speech and writing, the effects of scale-related facets need to be investigated. A description of processes in the empirical development of rating scales can enhance our understanding of method effects in performance testing.

Background. A persistent problem in language testing is fair and efficient rating of performance. Particularly in large scale testing programs the balance between comprehensiveness and transparency of rating instructions is critical to obtain rater agreement at minimum cost.

Purpose of the research. This research was undertaken to simplify and improve the rating instructions and procedures used in a large scale testing program of second language oral proficiency.

Research design and method. The rating process was analyzed using a think-aloud protocols recorded on tape. Rating problems were also catalogued from the standard reports required from all raters. Subsequently, rating instructions were categorized according to aspect, task and scoring rules. Based on twelve test forms, effective weights of ratings per category were computed. Two samples of one thousand candidates, were used to evaluate rater agreement and measurement properties according to item response theory, for all categories of rating instructions.

Results. The study shows that content and form aspects should be modeled as distinct traits. Dichotomous scoring rules yield higher rater agreement and better data-model fit than polytomous scoring rules. Dichotomization of originally polytomously scored ratings leads to improvement of rater agreement and data-model fit.

Implications. Simplification of rating instructions has been recommended to improve efficiency and fairness of the rating process. In order to benefit from the demonstrated higher measurement qualities of dichotomous scoring rules, raters require intensive training to overcome their inhibitions to make dichotomous decisions.
The Development of Second Language Graduation Proficiency Tests in Reading and Writing. Cheryl Alcaya, Melody Jacobs-Cassuto, Marcos Holzner, Ursula Lentz, Gabriela Sweet, & Dan Reed, Center for Advanced Research on Language Acquisition (CARLA), University of Minnesota

The poster session will present work on reading and writing tests from an ongoing project. Already developed are contextualized reading, speaking, and writing entrance proficiency tests for students of French, German, and Spanish. These assessments are applicable for establishing baseline proficiency (targeted for Intermediate Low level, based on the ACTFL scale) in both the secondary and post-secondary contexts. An entrance-level listening proficiency test, currently under construction, will complete the four-modality battery of criterion-referenced, proficiency-based assessments, each with a cohesive theme. The format of the reading test and the projected listening test is multiple choice, while the writing and speaking tests are expert rated on student performance. For the latter two modalities, tasks are scored on a four-point scale. Scoring is on a pass/fail basis (0-1) for each criterion.

As a continuation of this work, reading and writing graduation proficiency tests are also being developed, for Intermediate High and Intermediate Mid levels, respectively. The format and scoring for each of these assessments mirrors that of the EPT instruments presented above. The poster will focus on preliminary analyses from pilot tests of these graduation proficiency tests. Descriptive statistics, classical item analyses, and Item Response Theory (IRT) model fit will be examined. After revisions have been made, based on pilot tests, continued field testing is planned with the objective of increasing item bank size to allow for multiple versions of the assessments.

This research is part of a long-term program of the Association of Language Testers in Europe (ALTE), the goal of which is to provide a multi-lingual framework of language proficiency.

Previous phases of the Framework project have produced a set of “can-do” scales: user-oriented statements of what a language learner at a certain level is able to do. These scales, covering three general areas of experience (‘Social and Tourist’, ‘Work’, and ‘Study’) have so far been translated into eight ALTE member languages, and through a process of task analysis have been related to a five-level system and thence to the examination systems of ALTE members.

This research concerns the collection of empirical data on the can-do statements, to transform them from an essentially subjective set of level descriptions into a calibrated measuring instrument. Data collection is based on self-report, the can-do scales being presented as a set of linked questionnaires.

The investigation has these aims:

1. To check the function of individual statements within each can-do scale.

2. To equate the different can-do scales, i.e., establish the relative difficulty of the scales.

3. To investigate the neutrality of the can-do scales with respect to language.

Questionnaires are to be administered in the subjects’ own first language, except at very advanced levels. Copies of the questionnaires are available in eight languages with data collected in a number of countries.

All data are included in a single analysis. Bias analysis is then used to identify language-specific problems with particular tasks. Bias may concern particular language backgrounds, or particular target languages.

The analysis may lead to textual revision of the can-do scales, as well as some adjustment of levels. The can-do statements when equated across scales and across languages are intended to provide a sensitive language-neutral rating instrument which will be important in the next phase of the Framework Project - the empirical equating of particular language qualifications to the common system of levels.
Evaluation of the Dimensionality of a Free Writing Task and Multiple-Choice Questions for ESL Writing Assessment. Yeonsuk Cho, University of Illinois at Urbana-Champaign

The aim of this study is to compare the dimensionality of multiple-choice (MC) questions with a free writing task (i.e., essay) as methods used to measure writing ability of non-native speakers of English (NNS) for academic purposes. Free writing tasks have long been considered the best method to assess writing ability of NNSs. However, the scoring of essays, which heavily relies on the judgment of raters, has not been able to escape the criticism of reliability and length of scoring time. To overcome these persistent problems, MC questions, which can be easily scored, have been suggested as an equally reliable alternative.

For this research, MC questions developed to measure writing ability will be added to the current English Placement Test of the University of Illinois at Urbana-Champaign (EPT). The dimensionality of the free writing task and the MC test will be investigated, using hierarchical cluster analysis and a nonparametric test of dimensionality (Stout, 1987). Essays will be scored componentially to match the subscore structure of the MC test items: The hierarchical cluster analysis will analyze the componential scores of the essays and the corresponding MC subscores. A computer program, DIMTEST (Stout, 1987), will be used to evaluate whether the subscores are dimensionally similar. If both assessments prove to measure the same composite of writing ability, the MC test could be claimed to be an equally valid measure of writing ability, and thereby, could be used to supplement the essay ratings to increase the accuracy of the overall writing assessment.

**WORK IN PROGRESS**

Curriculum/Testing Revision-A Synergetic Methodology. Christa Hansen & Christine Jensen, *University of Kansas*

**Objective:** The creation of a principled methodology for the development of proficiency/placement test specifications, design and revisions that is dictated by and interacts with major curricular changes in a language teaching program.

**Rationale:** A program that undergoes a radical change in curriculum will of necessity require radical changes in the proficiency instrument that is used for placement into the program coursework. The purpose of this project is to design and implement a principled methodology for revising tests based on the needs dictated by curricular changes.

**Methodology:** The current curriculum is a skills-based, context-independent program. Placement for initial and subsequent semesters is based on performance on a battery of proficiency-based placement tests in the following skill areas: listening, reading, structure and composition. Proposed curricular changes include a content-enriched curriculum incorporating academic skills and learner strategies with an integration of language skills, language support workshops and tutorials. Placement and promotion issues are in the exploratory stage.

The methodology developed will address the following issues:
- Identification of curricular changes and subsequent mismatch of the current test specifications and objectives
- Decisions about connection between placement and promotion and test performance
- Identification of curricular outcomes
- Identification of testing outcomes
- Test specification review process based on the above information
- Development of an external validation procedure
- Transition to the new test.
WORK IN PROGRESS

Assessing Speaking Ability: Two Aspects (Linguistic and Pragmatic) from Three Dimensions (Monologue, Dialogue and Multilogue), Yuji Nakamura, Tokyo Keizai University

Is speaking ability a single component or is it composed of multi-componential traits? This paper explores Japanese college students’ (n=50) English speaking ability by focusing on the linguistic and pragmatic aspects from three dimensions categorized based on the number of people involved and context of the test situations as follows: 1) Monologue such as a speech making test or a visual test, 2) Dialogue such as a live interview test or a tape mediated speaking test, and 3) Multilogue such as a small group (3-4 people) discussion test or a large group (more than 20 people) discussion test. The linguistic aspect includes grammar, vocabulary, discourse etc., and the pragmatic aspect deals with the level of sociolinguistic appropriateness in spoken utterances.

Through a close examination of the correlations of six different test results, the paper will indicate the specificity of students’ speaking ability in one test and the generalizability of their speaking ability among the different tests.

The full study has not been completed yet. The interim results, as a whole, demonstrate that some students who are good at a monologue type test, even a dialogue type test, cannot perform satisfactorily in the discussion test/activities even though possessing an outgoing personality, which is generally considered a positive factor in speaking. It is hoped that a great variety of tests for communication ability might give a positive washback to students and urge them to realize the framework of communication. It may also lead the students to communication-oriented learning habits.

EAP: What is Important in Testing Reading. Liliana Fortuny & Martha Botto de Pocovi, Universidad Nacional de Salta, Argentina

The partial results of a reset in progress carried out at Consejo de Investigacion de la Universidad Nacional de Salta, Argentina, will be presented by means of a poster dis.

We consider that the production of a summary is of the utmost importance in EAP reading comprehension. So, testing should reflect this ability. To achieve this aim, we take into consideration the organization of information in a hierarchical order and the recognition of discourse strategies.

Key words: testing reading, summarizing strategies, discourse strategies.
WORK IN PROGRESS

Nationality Group Differences in the ESL Placement Test and Students’ Satisfaction over Placement, Kyong Hyon Pyo, University of Illinois, Urbana-Champaign

To date little research has been directed to the possibility that language placement tests in English as a Second Language (ESL) programs may be biased against examinees from particular language or cultural groups. The purpose of this research is to investigate the English Placement Test (EPT) employed at the University of Illinois, Urbana-Champaign, to see possible unfair decisions based on the results of the EPT which consists of four subtests (i.e., grammar, reading, dictation, and oral communication). The major aim of the study is threefold: (a) to identify the test patterns of examinees of varying national and educational group membership; (b) to compare the test attitudes and perceptions of the examinees toward the placing procedures at the Intensive English Institute (IEI) at UIUC; and (c) to determine the degree of association between test attitudes and test scores. A test attitude inventory was specifically designed to gather data on examinees’ attitudes and perceptions toward both the EPT and its placing procedures. The inventory was administered to 164 IEI students enrolled in the spring of 1996. It is my goal to identify a possible bias made by the interaction between the language aspect facet and group differences and support the idea of bias with the inventory showing perceptions, attitudes, and preferences of examinees in an ESL setting.

Creating an Assessment Instrument for ESL Conversation. Diana Steele, University of Illinois, Urbana - Champaign

This work-in-progress would describe the development of a reliable placement test from scratch for a small, new ESL school which currently does not even have an established curriculum. The long-term goal is to map the assessment process onto a curriculum for each level of English ability. However, because the majority of students desire conversation classes so as to improve their integration into this university community, the immediate purpose of this project is to create an interviewing instrument that assesses the learner’s oral communication and ensures each class (usually 4-6 students) will be homogeneous in conversational ability. My scoring rubric describes a range of abilities in five specific areas: pronunciation, grammar, vocabulary, listening comprehension, and fluency. Each teacher would receive a profile of scores for each student assigned to the class, such that the course content could be modified to meet the needs of the students in that particular class. Because assessment is an iterative process, we also need criteria for the teachers to be able to evaluate their students’ performance during the course in order to determine whether students were correctly placed.

My poster session would have a flow chart of the components of this assessment process. I would like feedback from LTRC members on such aspects of test development as the ideal weighting of my five chosen subcategories, the most pedagogically useful way to present the profile of subscores to the instructor, and appropriate methods of measuring the predictive validity of the instrument.
A Cognitive Perspective: Validation of Task Performance as a Measure of Communicative Competence. Minako Yamada, The Pennsylvania State University

Since the advent of communicative language teaching in the 1970’s, performance-based assessments have been used increasingly. Although communicative competence models that have been proposed to date incorporate a metacognitive component, only a few validation attempts have taken into account individual test taker’s cognitive processing efficiency. The current study proposes to validate a performance-based assessment as a measure of communicative competence using tasks that are designed to reveal the nature of cognitive constraints imposed by the efficiency of working memory (Ericsson & Kintsch, 1995).

The participants in the current study are Japanese and Chinese college students (N=30) residing in the U.S. A map task was chosen to elicit talk-in-interaction on the basis of Anderson, Brown, Shillcock, and Yule (1984). As in Anderson and Boyle (1994), for example, it has been shown that the map task elicits significant variation in the ability to communicate effectively. Computerized tasks such as a picture and word matching task (Gernsbacher & Faust, 1991) and a think-aloud task (Trabasso & Suh, 1993) were used to measure participants’ cognitive abilities.

The evaluation of construct-related evidence of validity of the map task performance will be examined with respect to the theoretical relationship between communicative competence and cognitive abilities. Generalizability analysis of the tasks as well as analysis of discourse features (Schegloff, 1980, 1990) manifested in the talk-in-interaction and verbal protocols will be used to further support the validity.

References


Problems Associated with Setting Language Admission Standards for Teachers Wanting to Teach English or French as a Second Language.  Doreen Bayliss, University of Ottawa

Faculties of Education offering a teacher education program with an option to teach English or French as a second language receive many applications. The problem facing these institutions is that of setting standards for the level of language required of potential teachers and of developing tests which reflect those standards. This poster session will outline how tests were constructed and standards were set and the subsequent problems encountered because of this testing initiative.

Since the tests were meant to identify candidates at an advanced L2 performance level, we decided that the tests would concentrate on reading, writing and oral skills. Listening would be tested indirectly via a modified dictation task (designed around problem words and phrases associated with L2 learners) and via the oral interview. The test had to serve two purposes: (1) to identify successful candidates for admission and (2) to provide diagnostic information to any interested candidates. Standards were set based on the performance of the field population of L1 students already enrolled in the teacher training program.

The test program has been in place three years and because of the high failure rate (both first and repeat failures) especially in French, we have or are in the process of implementing the following initiatives:

1. the development of a self-administered kit to determine readiness for the written test.
2. the future development of a self-assessment kit of oral skills based on the model of the Texas Oral Proficiency Test (Stansfield, 1994).
3. the possibility of mounting a special summer intensive language program designed to bring candidates identified as being close to the required level to the pass level.
The Local Item Banking System (LIBS) - a PC-based Item Bank for Developing Language Tests.
Simon Beeston, University of Cambridge Local Examinations Syndicate

This poster session presents Item Banking software developed at the University of Cambridge Local Examinations Syndicate. It is a modern, user-friendly system, uniquely integrating both word processing and database software and has the functionality required to make test management as straightforward as possible.

The system has the following key features:

1. All test tasks are stored as formatted Word for Windows files which can only be accessed through the item banking software. Where tasks include more than one item, these are marked in the document and linked to the database thus allowing description of individual items as well as main tasks;
2. All items can be described by an attribute system that can be tailored to individual requirements;
3. A straightforward item selection procedure allows for the rapid generation of new tests based on any of the descriptive categories held against items;
4. LIBS can generate a number of reports that allows users to decide if a test has the right content and level of difficulty. LIBS uses item difficulties to produce raw score to ability transformation tables. This facility helps in test equating;
5. A test version is stored as a Word for Windows document and requires only minor formatting to produce a final copy;
6. The bank can hold a large number of tasks and items in a secure system;
7. The system is a 32-bit application that runs on Windows '95 or NT.
POSTERS

Listening Comprehension, Critical Elements, and Classroom Discourse: Development of a New Measure.  Alejandro Brice, University of Central Florida

Griffin and Hannah (1960) reported that children spend as much as one half of their time in school listening to teachers talking. School discourse places great demands upon listening comprehension. This appears to become even more pronounced as grade level increases. However, it has been said that listening skills are the most neglected area of language testing (Donahue, 1997; Pearson & Fielding, 1982). In addition, most language tests do not measure listening as it applies to classroom discourse situations.

Hence, the purpose of this paper is to present the results of a study investigating teacher discourse in different classrooms (K-5 English as a second language, second grade general education, and fourth grade general education). The research questions are: (1) do different classrooms place different listening demands upon their students, (2) is a difference to be found according to grade level?

The insights gained from this study led to:

(1) The development of a new procedure measuring the length and complexity of spoken language. Language testers may be interested in the development of a measure of critical elements involved in listening and its concurrent validity in relation to standard language sampling measures of length and complexity and,

(2) Insights into testing bilingual children with possible language disorders. These will be shared in a discussion about how the new measure may be applied in testing situations.
Assessment as a critical Component of a Successful Articulation Project. Micheline Chalhoub-Deville, University of Iowa

Goals 2000: Educate America Act (1993) elevated the status of foreign languages (L2) from a tangential subject to an integral component of the curriculum. Given this new status, funding agencies have supported projects that provide coherent instructional programs and encourage students to persist in and achieve higher levels of L2 learning, e.g., the Minnesota Articulation Project (MNAP). MNAP is a statewide initiative for developing a proficiency-based model for coordinating French, German, and Spanish instruction and outcomes across various levels and systems. The present paper will report on MNAP development and use of assessments to achieve these goals.

The paper provides an overview of issues to consider in developing contextualized, criterion-referenced, proficiency-based assessment instruments that are compatible with a communicative approach to instruction and can be administered on a large-scale. Illustrative samples are provided. Also, the paper reports the results of modified Angoff procedure sessions in which L2 educators participated to set cut scores for making pass/fail decisions at the ACTFL Intermediate Low level.

Additionally, the paper reports the results and discusses the implications of ANOVA and regression analyses performed on various background variables, including years of L2 study, achievement, motivation, etc., for language learning/instruction. For example, results show that the IL level attained by university students after one year of L2 instruction is typically not achieved by high-school students after three years, as originally hypothesized. Findings indicate that most high-school students need at least four years to perform at the IL level. Such findings are likely to (a) make students/parents question the value of studying L2 in high school rather than at the university and (b) produce repercussions in the high-school curricula.
**POSTERS**

**Linguaskill: a Multi-lingual Computer-adaptive Testing System.** Neil Jones, *University of Cambridge Local Examinations Syndicate*

*Linguaskill* is a computer-adaptive test developed by UCLES for a corporate client, first released in 1995 as a text-only DOS program, and now re-developed for Windows. *Linguaskill* offers a range of task types, based on extensive item banks. Listening is tested through short dialogues with single MCQs, as well as by longer passages with groups of MCQs. Graphic prompts are used in some tasks. There are short and longer texts for testing reading, too. There are both open and MCQ forms of cloze passages, while a sentence completion exercise offers a form of constructed response task which is unusual in computer-based testing. *Linguaskill* is a general proficiency language test with a business style.

Performance is reported on a 5-band scale corresponding to the general framework adopted by ALTE (the Association of Language Testers in Europe). This framework is appropriate, as *Linguaskill*, first released in English, is now available for testing French, Spanish and German. Dutch and Italian are also under development. The development of these versions, in collaboration with ALTE partners, has presented a considerable challenge, not simply to construct and calibrate large item banks in each language, but also to develop a framework for validation which would enable us to establish the equivalence of the levels reported by each test, in terms of functional language ability. Data collected during the trialling of the different language versions in a number of countries allows us to begin to address this issue.
POSTERS


Understanding what trained raters attend to while scoring oral proficiency tests (OPTs) is critical to understanding test scores as evidence of learner language ability (Douglas, 1994) and to establishing test score validity (Chaloub-Deville, 1996; Douglas, 1994). While a body of literature exists using verbal protocols to examine elements attended to and thought processes used by writing test raters (Milanovic, Saville, & Shen, 1996; Vaughan, 1991; Huot, 1988), similar studies on OPT raters are conspicuous by their absence.

This paper presents the findings of a study replicating a triangulated methodology (Milanovic & Saville, 1994) of think-alouds, written retrospectives, and questionnaires to capture data on raters of the Speaking Proficiency English Assessment Kit (SPEAK), a semi-direct, tape-mediated test. The aims of the study were to develop a transcription coding scheme, identify features in the speech samples that raters attended to, and identify rater cognitive processes. Six experientially diverse raters were trained to score the SPEAK and perform verbal and written protocols during and immediately following the scoring of one-item samples from 10 examinees; questionnaires were completed post rating. Verbal and written reports were coded and analyzed. Preliminary results suggest that raters primarily attend to linguistic and discourse elements and that raters trained together with the same rating materials employ dissimilar rating strategies and cognitive thought process. Details of the triangulated method and findings along with implications of the study for better rubrics, rating scales, and rater training will be presented.
POSTERS

Investigating the Impact of IELTS. Michael Milanovic & Nick Saville, University of Cambridge Local Examinations Syndicate

Tests provided by major testing agencies and examination boards have an impact on educational processes and on society in general. While washback (i.e., the effect of tests on teaching and learning) has received some limited attention in recent years, the more general concept of impact and how it may be investigated has not been systematically addressed.

This session outlines the principled approach to impact that has been adopted by the University of Cambridge Local Examinations Syndicate, and illustrates this in relation to the International English Language Testing System (IELTS) which is currently taken by about 80,000 candidates a year seeking admission to higher education (e.g., in UK, Australia and Canada).

In order to understand test impact better, a range of instruments and procedures are being developed which initially are being used to focus on the following aspects of IELTS:

- the content and nature of classroom activity in IELTS-related classes;
- the content and nature of IELTS teaching materials;
- the views and attitudes of user groups towards IELTS;
- the IELTS test-taking population and the use of test results.

The first two of these points concern washback in the more traditional sense - the effect of the test on teaching and learning. The second two are concerned with the effects of the test on other systems in the administrative and academic contexts where the tests are used, and on attitudes and behaviour. The project is now in its third year and the session will report on the development and application of the instruments to date.
POSTERS

The Development of English Language Tests for Young Learners (age 7-12). Nick Saville, University of Cambridge Local Examinations Syndicate

Is it possible to give a seven-year-old child an English language test that makes an accurate and fair assessment and has a positive impact on the child's future language learning?

This question has been addressed by the University of Cambridge Local Examinations Syndicate in the development of the Cambridge Young Learners English Tests which have been designed for young learners in the age range of 7 to 12 years. This session presents the development of these tests.

The teaching of English to children of primary school age has been increasing rapidly in recent years and in many countries it is now part of the national curriculum. Even in contexts where there is no national policy, there is active support for such teaching to take place. Very importantly, parental demand has led to rapid growth of young learner classes in private language schools and, in turn, this has led to a demand for international tests for young children similar to those taken by their older brothers and sisters.

In this session the considerations and constraints which were taken into account during the development of the tests are discussed, including the specific needs of the test users and the educational consequences of using such tests with children. Considerations include:

- curriculum and pedagogy;
- children's cognitive and L1 development;
- appropriate tasks;
- variation between L1 groups and cultures.

Samples of material testing four skills - Reading, Writing, Listening and Speaking - are used to illustrate how the issues have been addressed and what solutions have been reached.
POSTERS

A Study on the Role of the TOEIC IP. Ishikawa Shoichi, Department of Foreign Languages, National Defense Academy, Yokosuka, Japan

This presentation will examine the role of the TOEIC Institutional Program (IP) at National Defense Academy (NDA). The Department of Foreign Languages of NDA require all freshmen to take the TOEIC IP each spring and this IP has been used as a placement test for freshmen classes. The practical research analyzing the results of the TOEIC IP has been going on by the author for the past three years.

The particular focus of this presentation will be an examination of what the IP purports to be as an academic testing instrument, how its format and content match or mismatch the students experience and needs at NDA, and statistical relevance results obtained.

Providing Diagnostic Information for Understanding Students’ Learning Processes.
Naoki Taira, The National Center for University Entrance Examinations
Hisami Saito-Scott, University of Illinois at Urbana-Champaign
Masahiro Kasai, State of Florida, Department of Business and Professional Regulations, Bureau of Testing

Students’ learning processes and individual differences should be emphasized in language classrooms. However, in assessing students’ learning ability, usually only total scores are provided, and this assessment procedure does not seem to consider the complexity of the vocabulary acquisition process of individual learners. In achievement tests, diagnosing students’ mastery of skills is especially important.

This study will challenge the issue by applying the Rule Space Methodology (RSM) to a Japanese vocabulary test. The purposes of this study are: (1) to identify cognitive and linguistics characteristics (hereafter called attributes) involved in the Japanese vocabulary test, and (2) to provide possible diagnostic information regarding the examinees’ attributes mastery. Thirty items were selected from a high-school-level Japanese test developed by Taira et al. (1993). The original 20 attributes were identified by analyzing think-aloud protocols of three native Japanese speakers. Two high school Japanese teachers revised the attributes list, and 23 final attributes were identified. Some examples were: the difficulty level of Kanji (Chinese characters), inference, and the degree of vocabulary formality.

The RSM was applied to 1,500 students’ responses, and 80.5% of students’ performance was explained by the proposed attributes. The results show that most students did not master the attribute: inferring the meaning of a word when the same Kanji was used in the stem and options.

This study revealed that diagnostic information obtained from the test is very useful for understanding students’ learning processes and improving instruction.