Language Testing Research Colloquium 1994

Program and Book of Abstracts

Center for Applied Linguistics
George Washington University
National Foreign Language Center

Washington, DC
March 5-7, 1994
Institutional Acknowledgements

Center for Applied Linguistics (CAL)

Sara Meléndez, President
Charles W. Stansfield, Director, Division of Foreign Language Education and Testing
Nell Hyman, Laurel Winston, Institutional Liaisons

National Foreign Language Center (NFLC)

David Maxwell, Director
Betsy Hart, Institutional Liaison

George Washington University (GWU)

Stephen Joel Trachtenberg, President
Ray C. Rist, Director, Policy Research Center
Joel Gómez, Director, Institute for the Study of Language and Education
Irene Thompson, Chair, Department of Slavic Languages and Literature
Maurya Meiers, Institutional Liaison
Special Acknowledgments

LTRC '94 Steering Committee

Dorry Kenyon, LTRC '94 Program Chair
Lyle Bachman, LTRC '95 Program Chair
John de Jong, LTRC '93 Program Chair

Vettors/Vetting Site Coordinators

In addition to the members of the steering committee, the following individuals served as vettors of abstracts and, in many cases, coordinators of group abstract vetting at their respective institutions.

Eduardo Cascallar, Center for the Advancement of Language Learning
Marjorie Des Brisay, The University of Ottawa
Catherine Elder, The University of Melbourne
Christine Jensen, Christa Hansen, The University of Kansas
Liz Hamp-Lyons, The University of Colorado
Grant Henning, The Pennsylvania State University
Sari Luoma, University of Jyväskylä
Nick Saville, University of Cambridge
Elana Shohamy, Tel Aviv University
Charles W. Stansfield, Center for Applied Linguistics

Conference Registration and Materials

Laurel Winston, Center for Applied Linguistics
Nell Hyman, Center for Applied Linguistics
Meg Malone, Center for Applied Linguistics

Additional On-Site Assistance

Becky Bar-Lev, National Foreign Language Resource Center/CAL
Beverly Boyson, National Foreign Language Resource Center/CAL
Xixiang Jiang, Center for Applied Linguistics
Renee Jourdenais, National Foreign Language Resource Center/CAL
Brenda McClain, National Foreign Language Resource Center/CAL
Lisa Pinson, National Foreign Language Resource Center/CAL
Pre-Colloquium Workshops
In the Second Floor Conference Room at the Center for Applied Linguistics
(by pre-registration only)

Workshop 1  Beyond the Manual: Advanced Topics in Many-Facet Rasch Analysis
John Michael Linacre, University of Chicago
Thursday, March 3, 9:00 a.m. - 4:30 p.m.
Friday, March 4, 9:00 a.m. - 12:00 p.m.

Workshop 2  An Introduction to Many-Facet Rasch Analysis
John Michael Linacre, University of Chicago
Friday, March 4, 1:00 p.m. - 5:00 p.m.

Friday, March 4

Site: Center for Applied Linguistics

7:00-9:00 pm Registration and Informal Welcoming Reception
Saturday, March 5

Site: Auditorium, Funger Hall, George Washington University

8:00 am - 5:30 pm  Registration Desk open at the Auditorium, Funger Hall, George Washington University

9:00-9:05 am  Welcoming Remarks  
Ray C. Rist, Director, Center for Policy Research, George Washington University

9:05-10:50 am  Paper Session 1, Chair: Charles Stansfield

Process and Outcomes in Oral Assessment  
Ann Lazaraton, The Pennsylvania State University  
Nick Saville, University of Cambridge

The Use of Role Play in Assessing Oral Ability: The Effects of Situational Context on Measurement  
Alexander Teasdale, Thames Valley University

Raters and Scales in Oral Proficiency Testing  
Lucinda Hart-González, Foreign Service Institute

10:50-11:15 am  Coffee Break

11:15-12:25 pm  Paper Session 2, Chair: Dorry Kenyon

A Case for Plausible Rival Hypotheses in Language Testing Research  
Alan Davies, University of Edinburgh

Models of Performance in Second Language Performance Tests  
Tim McNamara, The University of Melbourne

12:25-1:55 pm  Lunch Break

12:25 - 1:55 pm  Working Lunch Meeting for Language Testing
Saturday, March 5 (con’t)

1:55-3:40 pm  **Paper Session 3**, Chair: *Eduardo Caccallar*

_The Nature of Multi-Dimensional Data_  
Gary Buck, *Educational Testing Service*

_Theory Building: Sample Size and Data-Model Fit_  
John H.A.L. de Jong, *CITO*  
Fellyanka Kaftandjieva-Stoyanova, *CITO/University of Sophia*

_Scales as Task and Rater Specific: Derived Dimensions and Weights_  
Micheline Chalhoub-Deville, *The Ohio State University*

3:40-4:10 pm  **Coffee Break**

4:10-5:20 pm  **Paper Session 4**, Chair: *Elana Shohamy*

_The Complementary Roles of G-Theory and Multi-faceted Rasch Measurement in the Development of Performance-based Assessments of the ESL Speaking and Writing Skills of Immigrants_  
Brian Lynch, *University of Melbourne*  
Tim F. McNamara, *University of Melbourne*

_Using FACETS to Model Rater Training Effects_  
Sara Cushing Weigle, *UCLA*

Site:  One Washington Circle

7:30 pm  **LTRC ’1994 Banquet**

Program:  
Welcoming Remarks, *Charles W. Stansfield*  
Acknowledgements, *Dorry Kenyon*  
Presentation by ETS of the Second Annual TOEFL Dissertation Award to Antony John Kunnan
Sunday, March 6

Site: Auditorium, Funger Hall, George Washington University

8:00 am - 12:30 pm  Registration Desk open at Auditorium, Funger Hall, George Washington University

9:00-10:45 am  Paper Session 5  Chair: Nick Saville

Personality Characteristics and the Assessment of Spoken Language in an Academic Context
   Vivien Berry, The University of Hong Kong

The Effect of Information Gap on Elicited Discourse of Candidates in an Oral Test of English
   Gillian Wigglesworth, University of Melbourne

Two Communicative Tests in Comparison: What do the Test Results Show and What do Test-Takers Say About the Tests?
   Anu Halvari, Language Centre for Finnish Universities, University of Jyväskylä

10:45-11:15 am  Coffee Break

11:15-12:25 pm  Paper Session 6  Chair: Tim McNamara

A Cognitive Approach to Assessing the Relationships Between Different Forms of Writing Skills
   Alison J.K. Green, University of Cambridge

An Investigation of Marker Strategies Using Verbal Protocols
   Michael Milanovic, Local Examinations Syndicate, EFL Division, University of Cambridge
   Nick Saville, Local Examinations Syndicate, EFL Division, University of Cambridge

12:25 - 6:30 pm  FREE TIME
Site: One Washington Circle Hotel

6:30-7:30 pm Annual LTRC Business Meeting

7:30-8:30 pm Presentation of Poster Sessions Chair: John de Jong

Language Testing with Speech Recognition: Methods and Validation (Demonstration)
Jared Bernstein, Entropic Research Laboratory

Criterion-Referenced Language Test Development (CRLTD): An Overview
Fred Davidson, Division of English as an International Language, University of Illinois
Brian Lynch, Department of Applied Linguistics and Language Studies, University of Melbourne
Dongwan Cho, Division of English as an International Language, University of Illinois
Susan Larson, Environmental Engineering and Science Program, Department of Civil Engineering, University of Illinois

T-LAP: Test of Listening for Academic Purposes
Christa Hansen, University of Kansas
Christine Jensen, University of Kansas

A Multistep Placement Test
Michel Laurier, University of Montreal

What is Good Enough? Exploring Some Aspects of the Validity of Testing Speaking
Sari Luoma, Language Centre for Finnish Universities

The Relationship Between Type and Frequency of Errors and ACTFL Level on the SOPI
Margaret E. Malone, Center for Applied Linguistics

A Study of Writing Tasks Assigned in Academic Degree Programs: A Stage II Report
Carol Taylor, Educational Testing Service
Gordon Hale, Educational Testing Service

TOEFL 2000 Report: Defining the Constructs
Construct Validation of Measures of Communicative Effectiveness and Grammatical Accuracy: A Multimethod Approach
   John A. Upshur, Concordia University
   Carolyn E. Turner, McGill University

Considerations in Writing Item Prompts for SOPI
   Pavlos Pavlou, Georgetown University
   Sylvia Rasi, Pacific Union College

Site: Center for Applied Linguistics

8:30-10:00 pm Viewing of Poster Sessions at CAL
Monday, March 7

8:00-8:45 am  Vans leave from the Inn at Foggy Bottom to the National Foreign Language Center (NFLC)

Site: 1st Floor Auditorium, National Foreign Language Center

8:00-11:00 am  Registration Desk open at NFLC

9:00-9:05 am  Welcoming Remarks
   David Maxwell, Director, NFLC

9:05-10:50 am  Paper Session 7  Chair: David Ingram

Predicting Item Difficulty in a Reading Comprehension Test with an Artificial Neural Network
   Kyle Perkins, Department of Linguistics, Southern Illinois University
   Lalit Gupta, Department of Electrical Engineering, Southern Illinois University
   Ravi Tammana, Department of Electrical Engineering, Southern Illinois University

The Effect of Varying the Immediate Recall Protocol Procedure on Recall Measures of Listening Comprehension
   Sheryl V. Taylor, Ohio State University

Developing and Administering a Cloze Test in American Sign Language
   Christine Monikowski, National Technical Institute for the Deaf/Rochester Institute of Technology

10:50-11:15 am  Coffee Break
Monday, March 7 (cont’t)

11:15-2:00 pm  Experimental Sessions

11:15-12:00 pm  WORKSHOP
    NFLC 5th Floor
    Independence: A Lurking Assumption of the Statistical Models used in Language Testing
    Bruno D. Zumbo, University of Ottawa
    Tim Pychyl, Carleton University
    Janna Fox, Carleton University

11:15-12:00 pm  ROUNDTABLE DISCUSSION SESSIONS
    (A-C) NFLC 5th Floor
    (D) 1st Floor Auditorium

(A)  Testing Aptitude Testing: The MLAT
    Madeline Ehrman, Foreign Service Institute
    Lucinda Hart-González, Foreign Service Institute
    F.H. Jackson, Foreign Service Institute
    Joseph N. White, Foreign Service Institute

(B)  Comparing Language Qualifications in Different Languages: A Framework and Code of Practice
    Michael Milanovic, University of Cambridge
    Nick Saville, University of Cambridge

(C)  Language Competence in the Federal Government: An Articulation of Proficiency and Performance Assessment
    Eduardo C. Cascallar, Center for the Advancement of Language Learning
    Marijke W. Cascallar, Federal Bureau of Investigation
    John L.D. Clark, Defense Language Institute
    Madeline Ehrman, Foreign Service Institute
    Pardee Lowe, Jr., ILR Language Testing Committee
    Julie A. Thornton, Center for the Advancement of Language Learning
    Bernard Spolsky (discussant), Bar-Ilan University

(D)  Research on the Properties of Alternative Assessment Procedures
    Elana Shohamy, Tel-Aviv University

12:00-2:00 pm  SITE VISIT AND TESTING DEMONSTRATIONS
    ORGANIZER, Eduardo Cascallar, Center for the Advancement of Language Learning (CALL)

    (by advance reservations at local registration only; buses leave from NFLC, brown bag lunch available from the Inn at Foggy Bottom at additional charge)
Monday, March 7 (con't)

12:45-2:00 pm  
**Lunch Break**

2:00-3:45 pm  **Paper Session 8**  Chair: *Gary Buck*

- **ESL/EFL Trait Structure Variation at Multiple Ability Levels**  
  Fred Davidson, *University of Illinois*

- **Differential Subset Performance and Test-Taker Characteristics**  
  James Dean Brown, *University of Hawaii at Manoa*

- **Optimal Indices of Gain Score Dependability for Criterion-Referenced Language Tests**  
  Steven Ross, *University of Hawaii, Manoa*  
  Te Fang Hua, *East-West Center/University of Hawaii, Manoa*

3:45-4:15 pm  **Coffee Break**

4:15-5:25 pm  **Paper Session 9**  Chair: *Alan Davies*

- **The Role of Cohesion in Communicative Competence as Exemplified in Oral Proficiency Testing**  
  Pávlos Pavlou, *Linguistics Department, Georgetown University*

- **Metalinguistic Knowledge, Language Aptitude and Proficiency**  
  J. Charles Alderson, *Department of Linguistics and Modern English Language, Lancaster University*  
  David Steel, *Lancaster University*

5:30-6:00 pm  Vans leave from NFLC to the Inn at Foggy Bottom

**Site: Center for Applied Linguistics**

7:30-9:30 pm  **Annual Business Meeting of the International Language Testing Association (ILTA)**

**TRANSPORTATION TO BALTIMORE/TESOL**  
(by pre-registration/confirmation at the local registration desk only)

Monday, March 7  
8:00 pm  Evening bus to Baltimore/TESOL leaves from the Inn at Foggy Bottom

Tuesday, March 8  
9:00 am  Morning bus to Baltimore/TESOL leaves from the Inn at Foggy Bottom
Abstracts of Papers

(in alphabetical order by first author)
Metalinguistic Knowledge, Language Aptitude and Proficiency

J. Charles Alderson, Department of Linguistics and Modern English Language, Bowland College, Lancaster University
David Steel, Lancaster University

Objectives of the Research
The aim of the research to be reported in this paper was a) to establish the extent to which British students entering university to study French had an awareness of and a sensitivity to grammar b) to explore the relationship between such awareness and the students' accuracy in their use of French and c) to relate both the above to an ability to use French for communicative purposes.

Background/Rationale
Developments in the communicative teaching of modern foreign languages over the past two decades are largely perceived to have succeeded in enabling British language learners to communicate in their chosen language. However, there is a growing concern among language specialists at the tertiary level that university students are increasingly inaccurate in their use of the foreign language. In addition, the perception is growing that students entering universities to study language have very little knowledge about language. Yet it is felt by university teachers that students need to know about language in order to deepen their proficiency in the language, to increase their linguistic accuracy, and in order to be sensitive to language use.

A parallel development within British language teaching has become known as the language awareness movement, which has its origins in a concern for the impact that explicit language knowledge has on the language and language learning. This movement is currently debating what language awareness is, and how it might be related to proficiency. There is currently no evidence that knowledge about language and language awareness enhances language proficiency. Nor have any tests been developed to explore this question.

Method
A battery of tests was developed, comprising a) a test of metalinguistic knowledge, for both French and English, b) a test of language aptitude, c) a test of grammatical knowledge in French, d) a self-assessment questionnaire, e) a test of French comprehension. In addition, a bio-data sheet collected information on previous learning and achievement in French.

Results and Implications
Levels of metalinguistic knowledge appeared to be low. However, no relationship was found between metalinguistic knowledge, accuracy in the use of French, and communicative ability. It is argued that developments in language teaching and applied linguistics have proceeded without paying attention to the need to operationalise constructs. The need for better tests of metalinguistic knowledge and language awareness will be discussed, and proposals for test design will be presented.
Personality Characteristics and the Assessment of Spoken Language in an Academic Context

Vivien Berry, *The University of Hong Kong*

**Theoretical Background and Rationale**
Paired interviews and group discussions are becoming increasingly popular as methods of assessing spoken language both in international proficiency tests and, more particularly, in university placement tests. Yet recent research has shown that extreme extroverts and introverts differ in how well they perform on oral test interviews depending on whether personality types are homogeneously or heterogeneously paired. It can also be predicted from relevant experimental research reported in the personality psychology literature (Hall et al., 1988; George, 1990) that a similar kind of bias will exist when extroverts and introverts are tested in groups.

**Purpose of the Research**
The aim of the study is to investigate the hypothesis that extroverts and introverts will perform differently on a group oral test, depending on the degree of homogeneity of personality characteristics within a group.

**Research Design and Methodology**
150 incoming undergraduate students were assigned to groups of five and tested on their ability to take part in an academic seminar/tutorial prompted by a short reading passage and a lead-in question from a tutor. Each student was rated by two experienced raters on a nine point scale for a) participation and relevance and b) articulation. The 90-item version of the Eysenck Personality Questionnaire, validated for use in Hong Kong (Eysenck and Chan, 1982) was administered to all students. Ratings of speaking performance of both extremes on the extraversion scale are compared to the homogeneity of personality types within each group.

**Results**
Results indicate that differences can be observed in the performances of extroverts and introverts under varying interactional conditions.

**Implications of the Results**
The findings from this research clearly demonstrate the importance of deriving hypotheses from the psychological literature when conducting research into the effect of personality variables on performance. The paper concludes with a discussion of the feasibility of oral testing in groups and of the stability of results obtained.

**References**
Differential Subset Performance and Test-Taker Characteristics

James Dean Brown, University of Hawaii at Manoa

Farhady (1982) argued that test-taker characteristics like sex and language background may be related to differential performance on various types of language tests. The purpose of the present project was to explore the issues raised by Farhady while avoiding the flaws in that study by addressing the following research questions:

1. Are there significant and meaningful differences in test performance on the TOEFL due to subsets, sexes, languages, or their interactions?
2. Which languages are significantly different from which and on which subsets?
3. What is the percent of the variance due to subsets, sexes, languages, or their relative interactions?
4. What are the relative contributions to test variance of the languages, persons, items, subsets and their interactions?

There were 24,500 subjects in this study sampled from the May 1991 worldwide administration of the TOEFL. The materials involved were administered under normal operational conditions and included all three subsets of the TOEFL: a) listening comprehension, b) structure and written expression, and c) vocabulary and reading comprehension.

The statistical analyses included descriptive statistics, repeated-measures ANOVA, and follow-up univariate ANOVAs (while controlling for experimentwise probability levels), Scheffé post-hoc comparisons, and eta squared analysis, as well as a series of generalizability studies conducted to isolate the variance components due to persons, items, subsets, and languages.

The results indicate that there were statistically significant differences in performance between the sexes and among the languages. However, further analyses showed that these differences were not very meaningful and, more importantly, that the interaction effects of the sexes with subsets and languages with subsets accounted for very little of the variance in the test scores. It may therefore be time to rethink any claims of important differential test performance based on test-taker characteristics.
The Nature of Multi-Dimensional Data

Gary Buck, Educational Testing Service

Snow and Lohman (1988) have pointed out that there is a fundamental mismatch between the two main types of models of cognitive performance: psychometric models and cognitive information processing models. Language testers use statistical procedures based on psychometric models to analyze data which linguists increasingly describe in information processing terms. It is pertinent to question the validity of such analyses. At LTRC 1992, the author presented an initial report which examined this issue, entitled "The Analysis of Multidimensional Data Sets." This research attempted to examine the extent to which different degrees of violation of the assumption of uni-dimensionality influenced statistical analysis of a test.

Using Coomb's (1964) model of multi-dimensional data, a number of data sets, with known characteristics, were constructed with various traits combined in conjunctive, disjunctive and compensatory composition. These included bi-dimensional, tri-dimensional and a number of multi-dimensional data sets. These data sets were then analyzed using both classical and Rasch techniques, both of which make the assumption of uni-dimensionality. Initial results suggested that violation of the uni-dimensional assumption does influence the validity of common methods of data analysis. However, the results also suggested that the extent to which this happens depends on the nature of the relationship between the various traits. Multi-dimensional data in which the traits are in compensatory combination seems to behave more like one uni-dimensional data set, whereas traits in conjunctive, or disjunctive combination result in a far more serious violation of the uni-dimensional assumption.

This paper reports the continuation of this research agenda. The effects of different degrees and different types of multi-dimensional data are examined. The properties of these data sets are known (reliability, ability of each testee on each trait, correlations between the traits, etc.). They will be subjected to a wide range of statistical procedures: classical, Rasch and IRT. Various subsets of items and persons will also be examined; along with commonly suggested tests of data dimensionality. The results of these analyses will be compared with the known ability of the persons and difficulty of the items on each of the traits. These comparisons will allow the researcher to draw conclusions about the validity of commonly used test analysis procedures. Results will be reported, and conclusions drawn. The implications for language testing, and the analysis of language test data will be discussed.
Scales as Task and Rater Specific: Derived Dimensions and Weights

Micheline Chalhoub-Deville, *The Ohio State University*

**Background**

Much current research focuses on the severity of raters' subjective judgments (as indicated by the number of papers presented at LTRC'93), an important issue in L2 oral assessment. Whereas these investigations examine scores and fit indices, they do not reflect what these scores mean. According to Brindley (1991), because different judges may operate with their own personalized constructs irrespective of the criteria they are given, it would be a mistake to assume that high inter-rater reliability [or proper rater fit] constitutes evidence of the construct validity of the scales or performance descriptors that are used” (p. 157). Research is, therefore, needed that derives the dimensions salient to different rater groups when assessing learners’ L2 oral proficiency. This would facilitate a better interpretation and use of L2 oral test scores.

**Purpose**

The initial phase of the present paper investigated the weighted components of L2 holistic oral scores of learners of Arabic performing three elicitation tasks (interview, narration, and read aloud) across different rater groups. Raters were divided into three groups: native speakers (NSs) who teach Arabic as a foreign language to adults in the U.S.; non-teaching NSs who have been residing in the U.S. for a period of at least one year; and non-teaching NSs living in an Arab country. Results showed that the three rating groups used dimensions differentially to assess L2 learner's speech.

Because of the documented variation in learners’ speech products due to task effect (Larsen-Freeman & Long, 1991), it was important to follow up on the above study by investigating whether different dimensions would emerge according to task type. The present study analyzed each task separately and examined the dimensions considered by each rater group on each of the three tasks. This study addressed the following research questions: (1) What are the dimensions that underlie L2 oral scores on each of the three elicitation techniques? (2) What are the relative weights of the derived dimensions across the three rater groups?

**Methodology**

An adaptation of the matched-guise technique (Lambert, 1967) was used to randomize 18 speech samples, which were then presented to the raters. Each rater provided a holistic score for every speech sample. Holistic scores were analyzed using multidimensional scaling (MDS). More specifically, individual differences scaling (INDSCAL) that "accounts for individual differences in the perceptual or cognitive processes that generate the responses [ratings]" (Young and Harris, 1990) was deemed the most appropriate MDS technique to specify the salience of each of the delineated dimensions for each of the three rater groups.

The averaged holistic scores provided by each of the three rater groups were used to construct three proximity matrices, the rows and columns of which represented the 18 speech samples. The three matrices were submitted for analysis using the INDSCAL model within the ALSCAL MDS program of SPSS (1990). Undimensional scale ratings, obtained from each rater, were also used to aid explication of the MDS solutions.

**Results**

Results of task analyses corroborated the documented effect of elicitation task on speech products (Beebe, 1980; Schmidt, 1977, 1988; Tarone, 1983), i.e., the three tasks resulted in different oral language dimensions. Analyses also showed that the three rating groups used dimensions differentially to assess the L2 learner's speech. Further analyses that looked within and across tasks, and rater groups will be reported in the actual presentation.

**Significance**

This study provides educators with a better understanding of the L2 oral construct and will hopefully lead to improvement in assessment of the L2 speaking skill. This study argues that L2 oral testing should not employ generic, a priori component scales; rather, scales need to be empirically derived according to the particular task and audience. In addition, results indicate that our knowledge of how NSs assess non-native speech is lacking, and thus in need of improvement for the purpose of training teachers and developing scales.
ESL/EFL Trait Structure Variation at Multiple Ability Levels

Fred Davidson, University of Illinois

Objectives
This study examines variability of language trait structure at several ability levels.

Background/Rationale
For many years now, researchers in language testing have investigated the nature of the language trait in second/foreign language learning. Typically, these studies center on statistical evidence for the multidimensionality or unidimensionality of second/foreign language ability, and most often this evidence is provided by factor analysis. Repeatedly, language testing research has uncovered either a single language factor, or several factors which are themselves dominated by a single, higher-order factor. This ubiquitous general language trait continues to haunt language ability modeling, because language learning theorists have insisted that language is componentially complex.

Method
The present study reports on new evidence in this area. Seven language test datasets from five world sources were analyzed by the first phase of item-level exploratory factor analysis (EFA), i.e. principal components analysis (PCA). Each dataset was analyzed at the whole-group level and at five normally-distributed subgroups representing the whole ability range. As a cross-check, analysis was also run for each test on a randomly-drawn subgroup spanning the entire ability range but of the same n-size as the ability subgroups. All analyses used item-level zero-one data as input, and both smoothed and unsmoothed inter-item tetrachoric coefficients were examined. Dimensionality was reported by studying the difference in magnitude between the first and second eigenvalues, and by examination of scree plots. Comparisons were examined across ability levels within tests and across all ability levels and tests.

Results
Results indicate two major findings. (It must be emphasized that ‘dimensionality’ in the PCA context refers to whether or not further EFA would be motivated to seek more than one dimension, rather than to claim that one particular test is more or less dimensional than another.) First, generally, the whole-group and random subsample analyses tend to give the greatest evidence of unidimensionality. Second, there is a by-test effect in that some tests tend to exhibit greater evidence of unidimensionality than others across all ability groups. A tertiary finding is that the use of smoothed vs. unsmoothed tetrachorics did not appear to affect these two findings.

Importance/Implications
This project suggests that future work on detection of language multidimensionality should take multiple ability levels into account.
A Case for Plausible Rival Hypotheses in Language Testing Research

Alan Davies, University of Edinburgh

Objective
To argue that language testing research can pay closer attention to construct validity by taking account of work in second language acquisition (SLA) research.

Rationale
Current research in language testing is much concerned with the elaboration of reliability procedures. The paper argues that a greater role should be given to examining the validity of our ideas (Tyler 1963) regarding language. Because validation is never complete, progress in validation requires a continuing shift of attention. Language testing research now needs to devote time to 'strong construct validation (which) is best guided by the phrase "plausible rival hypotheses"' (Cronbach 1988).

Procedures
This paper offers input from an extensive in-depth review of current SLA Research (Davies 1993) which can help explore the validity of our own ideas. Four major areas of consensus in SLA research are examined in terms of the need to take account of:

1) both unitary and divisible views of proficiency in eliciting and reporting on candidate performance
2) the salience of particular features of language performance (e.g. accuracy and fluency) in rater judgements of proficiency
3) the nature of interaction between interviewer and candidate and between candidates themselves and its effect on both performance and rater judgements of that performance
4) the importance of elicitation task characteristics and their role in exploring the construct of variable proficiency (Skehan 1992)

Conclusions
A programme of language testing research which makes use of the key features of recent work in SLA is proposed.

Implications
Language testing research must look outside itself to other areas of applied linguistics, such as second language acquisition research, for the strong plausible hypotheses it needs for progress in construct validation.

References
Skehan, P. 1992 'Tasks in language testing' plenary address LTRC Vancouver (unpublished)
Theory Building: Sample Size and Data-Model Fit

John H.A.L. de Jong, CITO
Fellyanka Kaftandjieva-Stoyanova, CITO/University of Sophia

Background and Rationale
Item Response Theory (IRT) is often used in language testing and language testing research. The properties of the IRT models such as the Rasch model offer many advantages in the construction of measurement instruments and in the building of theoretical models or constructs of language proficiency. An important requirement is that for these properties to apply, the degree of fit of a data set to the model must be assessed. In this approach, as in any statistical testing of a hypothesis, the objective of the researcher is to falsify the model. The researcher sets up a hypothesis, which (s)he believes to be true. Subsequently (s)he collects data in an attempt to prove the model wrong. If this attempt is not successful, i.e., the data do fit the model, then the researcher can safely assume that, at least for the data used, there is no need to reject the model.

Purpose of the Research
This research was undertaken to investigate the influence of sample size on (1) the estimate of model-data fit, (2) the evaluation of dimensionality in a data set, and (3) the ability estimates of subjects.

Research Design and Method
Data from a number of groups of a thousand or more candidates taking Dutch as a second language examinations have been collected. The examinations compromise separate tests for reading, listening, writing and speaking. In the present study, the groups of examinees are divided into several randomly assembled subsamples, differing in size. Independent analyses of the subsamples are run to assess the influence of sample size on the output variables.

Implications of the Study
The advantages of IRT are a direct result of the strong assumptions underlying IRT models. However, in language testing research, the sample sizes used by researchers are often so small that it is highly unlikely that the measurement model can be falsified. A serious implication of the present study then would be, that many reported studies in language testing research should be reconsidered as to their claims with respect to theory building.
A Cognitive Approach to Assessing the Relationships Between Different Forms of Writing Skills

Dr. Alison J.K. Green, University of Cambridge

Objectives
Most theoretical approaches to assessing language proficiency assume that language proficiency is not a unitary skill. Even taking one of those components, writing skill, it is clear that there are different kinds of writing skill. An important question, then, is what to assess and how to assess it?

Many assessments of linguistic skills focus on a student’s ability to demonstrate his/her ability through writing. Different writing tasks make different demands of students and are likely to require different sorts of skills. This paper sets out to identify more precisely what it is that a range of assessments of writing skills are actually assessing, and whether this is reflected in examiner behaviour.

Theoretical Background and Rationale
Writing is not a unitary skill - there are different classes of writing (Bereiter and Scardamalia, 1987). Assessments of writing skills tend to be problematic because of the subjective nature of the assessment. Whilst it is possible to define objective criteria for the assessment of mechanical aspects of writing skills (e.g. spelling, grammar, punctuation and so on), the more complex forms of writing depend on less well-defined criteria for assessment, e.g. "organisation", "clarity", "novelty". The more complex forms of writing also demand non-linguistic skills, e.g. reasoning skills. What are the relationships between different forms of writing and can these be understood by considering the range of cognitive processes involved in carrying out the task? How adequate are the criteria used for assessing performance on writing tasks and do these criteria accurately reflect the skills that are actually used? Which factors most influence an assessment of writing skill, and are these primarily linguistic factors, or are there other factors?

Method
Students attending a local Sixth Form College completed six essay tasks. The first four essay tasks required students to encode new information and then to base a composition on that information. The sixth essay task required students to encode new information containing an argument and then to evaluate that argument.

Cognitive task analysis was used to specify the cognitive demands of each task. Concurrent think aloud protocols were gathered from examiners in order to identify the factors examiners were focusing on in order to make their assessments.

Results and Conclusions
The essay tasks clearly demanded different combinations of relevant prior knowledge and linguistic skills. As predicted, greatest variations in performance occurred for those essay tasks which demanded the highest amounts of relevant prior knowledge. Results focus on the interactions between task demands, cognitive skills involved and appropriacy of assessment.

Implications of Results
Cognitive task analysis allows the researcher to better understand the ways in which different classes of knowledge and different cognitive processes interact in carrying out a given task. Having specified what is required in order to perform a given task, the researcher can then state what an assessment ought to focus on, and determine whether the assessment is actually based on those factors or not.
Two Communicative Tests in Comparison: What Do the Test Results Show and What Do Test-Takers Say About the Tests?

Anu Halvari, Language Centre for Finnish Universities, University of Jyväskylä

During the summer and fall of 1993, a comparability study involving the International Certificate Conference's Certificate in English and the Finnish National Foreign Language Certificate in English was carried out. Both tests are aimed at adult language learners of intermediate (approximately Threshold) level. The tests are communicative by nature but different in format. The aims of the study were 1) to find out how the tests compare to each other in terms of levels achieved by the subjects, 2) to find out how the subjects felt about doing two tests seemingly of the same level, but different in form.

30 subjects were asked to complete both of the tests and fill in pre- and post-test questionnaires. The tests include reading, writing, vocabulary, structure, listening and speaking. The quantitative part of the study comprises the subjects' scores and marks; the analyses include descriptive statistics. The qualitative part of the study comprises the analysis of the data obtained from questionnaires and some interviews.

The preliminary results show some interesting tendencies. The two tests seem to measure somewhat different levels of language proficiency. On the English Speaking Union's 9-level scale, the ICC Certificate in English seems to measure levels 3-4, while the intermediate level of the National Foreign Language Certificate in English measures ESU levels 3-5. The differences in levels achieved by the testees in each subtest are worth noting. There seems to be test-internal variation in the requirements for the level of proficiency in the separate subtest for achieving the same mark. For example, the vocabulary and structures subtest in the ICC Certificate in English demands a higher level of proficiency compared to the overall proficiency requirements, while in the National Foreign Language Certificate in English it is the listening comprehension tasks that are more demanding in relation to the other subtests.

The qualitative data show that the subjects felt that both tests had pros and cons: the ICC Certificate Level test was considered homogenous and effective, yet dull and one-sided mainly because of the frequent use of multiple choice. The National Foreign Language Certificate in English was considered communicative with real-world tasks and situations, but the test takers often questioned its function as a test. The open-endedness of the tasks made the testees unsure about what was required of them.

The practical purpose of this is to apply the information obtained from the data in the development of the National Certificates. More widely, this study may provide some new insights onto how results from the comparability studies can be used in test development.
Raters and Scales in Oral Proficiency Testing

Lucinda Hart-González, Foreign Service Institute

Oral proficiency testing necessarily depends on some system of rating speech samples. The most common form of such rating is on some sort of Likert scale of steps from limited to no proficiency up to some expected upper limit of proficiency. It has been suggested that the cognitive limit to the number of steps a rating scale can have is seven. If it is so, then this fact is in tension with the statistical principle that the greater the differentiation in a scale, the greater the power of the analysis (Klein-Braley 1991).

The language proficiency scale used in federal agencies, named for the Inter-Agency Language Roundtable (ILR), has the appearance and some characteristics of a eleven-point Likert scale, ranging from 0 to 5 with "+" levels in between. In other ways, however, it is quite different. This paper discusses the language proficiency rating process at one federal agency in detail, always from the point of view of the statistical implication of the scale.

The development of the current rating process at this agency is recounted essentially as a pendulum swing between the competing motivators, scale reliability and score differentiation. Parts of the history of testing at this agency have been recounted in different places (e.g. Rice 1959, Lowe 1988, Cornwell & Budzinski 1993), but this is the first attempt to examine the whole of that development specifically in terms of its scaling implications. While the ILR scale began as essentially a Likert scale, a checklist of rating factors (e.g. grammar, vocabulary, fluency) was developed by Rice and Wilds which transformed the scale from a pure Likert scale into something Guttman-like without fully utilising the item-countability feature of checklists that make the Guttman scalogram possible (Maranell 1974, Angoff 1984). The particular factors have been modified from time to time, but this has no effect on scaling.

The next important event for scaling was the development of an Index Score of 0-120 points to assist in the assigning of ILR proficiency levels. This Index Score is a clear swing toward score differentiation, and could in principle help answer such questions as "how nearly the student has approached the next higher proficiency level since the last assessment was made" (Clark & Lett 1988:78).

In a swing back toward recognizing the cognitive limitations against a 120-point scale, the Index Scores were divided into six Performance Levels. While the number of levels is the same as the six basic bands of the 0-5 ILR scale, the two are not entirely comparable, as will be shown. Once the Performance level is assigned, the next step is rating on a subscale of up to 6 points which will contribute to the total Index Score. In other words, the cognitive limits of rating are addressed by successive rating of smaller portions of the larger scale, which allows for greater differentiation.

This testing procedure is more complex than the procedures at any of the other agencies, but it offers promise for research, little of which has yet been done using the Index Scores (but see White 1991). Problems with using the Index Score for research purposes, e.g. the problem of independence, are also discussed.
Process and Outcomes in Oral Assessment

Anne Lazaraton, The Pennsylvania State University
Nick Saville, University of Cambridge

Objectives of the Research
In this research project, the aim was to identify features of oral examiner language and to search for outcomes in terms of candidate language behavior and its subsequent rating in relation to these variables.

Background and Rationale
In standardized oral assessment procedures it has normally been assumed that the language of the oral examiner and the task the candidate is asked to perform will be "neutral" in relation to the outcome (the rating awarded to the candidate). Recently, however, test developers and oral examiners have begun to recognize that the talk and interaction that takes place during oral assessments affects in some way the language that candidates produce, and by extension, the ratings they are awarded. One line of research in this area has focused on the affective factors that are in force in the interaction (Porter, 1991), other researchers have focused on contextual factors of speech event (Shohamy, et. al., 1993), and still others have looked at the oral examiner and the type of accommodative behaviour that occur (Ross, 1992; Ross and Berwick, 1992).

One avenue of inquiry that has not been explored empirically is how the instructions about assessment tasks are set out via the instructional language of the examiner who acts as interlocutor; how standardised must this language be to ensure that all candidates are given the same opportunities to actually complete the assigned task? That is, does the language that is used by the interlocutor in setting up the task and providing procedural support during it lead to different outcomes in term of candidate ratings? Is it possible to monitor this instructional language, and to control it to a sufficient extent to ensure that procedures are standardised?

Research Method
In order to address these questions, the method used in this study was to take an established oral procedure and use it for collecting data. Specifically, both process and outcome data were collected for one administration of the Cambridge Assessment of Spoken English (CASE) (Milanovic et al., 1992; Lazaraton, 1993); the data-set contains detailed transcripts and candidate ratings for 58 assessments, conducted by 10 trained examiners, that took place in Japan in September 1992. In particular, the researchers focused on the phases of the assessment procedure in which the interlocutor acts as a facilitator in order to set up testing tasks by following prescribed interlocutor frames.

The aim of the analysis was to identify features of examiner language which occurred in these phases of the CASE assessments (using conversation analysis to understand the recorded and transcribed assessments) and to examine the outcomes of candidate language and the ratings awarded in relation to these features of examiner behavior (using statistical procedures to analyze the ratings).

Results
Preliminary results indicate that some features of examiner language, whether it be for "instructional" purposes or to otherwise support the candidate, do have an impact on both the language the candidates produce in the local turn-by-turn interaction, and the more global ratings of their overall (and more specific) abilities. Although no firm conclusions can be reached, areas for future research are discussed.

Implications
A greater understanding of the processes involved in face-to-face assessments will lead to better design of assessment procedures, better training of the examiners who are involved in them, and better interpretation of the outcomes - the actual ratings of candidate language - that emerge.
The Complementary Roles of G-theory and Multi-faceted Rasch Measurement in the Development of Performance based Assessments of the ESL Speaking and Writing Skills of Immigrants

Brian K. Lynch, University of Melbourne  
Tim F. McNamara, University of Melbourne

This study investigates the potential roles of Generalizability Theory (Brennan 1983; Shavelson & Webb 1991) and multi-faceted Rasch measurement (Linacre 1993) in the development of a performance based assessment procedure. Second language performance tests, through the richness of the assessment context, introduce a range of facets which may influence the chances of success of a candidate on the test. Preliminary investigation of the relative contributions of Generalizability Theory and multi-faceted Rasch measurement to the development of such tests has begun (e.g., Bachman, Lynch, Mason 1993), but the extension of this research to a range of assessment settings is required.

Data for this study come from a trial of materials from the access test, a recently introduced test of communicative skills in English as Second Language for intending immigrants to Australia. Performances on the speaking sub-test of 93 candidates are multiply rated, and a number of facets of the assessment setting investigated using G-studies, D-Studies and multi-faceted Rasch analyses. Practical consequences of varying the number of raters and the structure of the sub-tests as measured by reliability coefficients (G-theory) and errors of estimate (Rasch) are reported and discussed, and recommendations made for test design. The advantages and specific role of the contrasting analytical techniques are considered in detail in the light of the analysis.

The study demonstrates the role of Generalizability Theory and multi-faceted Rasch measurement in test design and in more theoretically motivated research on second language performance tests. It also clarifies our understanding of the comparative advantages of each of these relatively recent methodologies in the analysis of performance test data, and contributes to the current ongoing debate in educational measurement generally about their relative characteristics.

References
Models of Performance in Second Language Performance Tests

Tim F. McNamara, The University of Melbourne

Objectives
To consider models of performance in second language performance tests and their implications for research on aspects of the validity of such tests.

Background/Rationale
Twenty five years after the advent of communicative language testing, and with the growing use of performance based assessment, we have only poorly characterized from a theoretical point of view the performance dimension of second language performance tests. Despite recent advances in our conceptualization of the 'knowledge' or 'competence' dimension of such tests (Bachman, 1990; Bachman and Palmer, in press), our relative incoherence about what Hymes terms 'ability for use' seriously jeopardizes test validity and inhibits our ability to conduct research on such tests.

Methods/Design/Procedures/Techniques
The paper summarizes work in the tradition of Oller, which sees performance tests as being characterized by internal processing requirements, and in the personnel selection tradition, which focuses on the successful performance of observable tasks, and discusses issues raised by each of these approaches. It relates the discussion to Hymes's distinction between two senses of performance: 'actual instances of use' and 'underlying models of performance', and to Messick's distinction between task-centered and construct-centered approaches to performance assessment. Problems peculiar to language performance assessment, where language is both vehicle and target of assessment, are identified. The issues raised are related to examples from current occupationally related second language performance test development projects.

Results/Conclusions
The current confused state of debate in this field, and the necessity for adequate models of performance for research on performance tests, are demonstrated.

Importance/Implications of the Results
The construct validity of second language performance assessments is currently not well established. The prevalence of this form of assessment requires us to more adequately model and research the nature of second language performance test settings.
An Investigation of Marker Strategies Using Verbal Protocols

Michael Milanovic, Local Examinations Syndicate, EFL Division, University of Cambridge
Nick Saville, Local Examinations Syndicate, EFL Division, University of Cambridge

Objectives of the Research
The investigation of the markers’ thought processes in the marking of examination compositions is an important issue in the assessment of L2 writing (Cumming, 1990). To date, little is known about the decision-making strategies which are employed by the markers in making their assessments, but a better understanding of this area would lead to better design of writing tests and improved marking procedures (Milanovic, Saville, and Shen, 1993).

Background and Rationale
In 1992, the researchers designed and carried out an exploratory study to investigate the thought processes of markers for the Cambridge FCE/CPE compositions. The study was based on a hypothetical model of the marking process and investigated markers’ decision-making processes while evaluating EFL compositions. It looked at the effects on decision-making behaviour of the markers background and the proficiency level of the scripts, and data was collected using 4 methods: retrospective written protocols; verbal protocols while marking; a questionnaire; and a group interview.

The current study was designed as a follow-up, based on the findings in the exploratory study. Unlike the exploratory study, however, data was captured using a single method and a single homogenous group of markers was used. The aim of the present study was to investigate: a) the sequence of different behaviours employed by the markers, and b) the elements within the compositions which the markers focused on.

Research Method
The research design used verbal reports (protocols) as data, i.e. through the recording, transcription and analysis of examiner’s verbal comments made while marking compositions. (Ericsson and Simon, 1993). 20 experienced markers for the Cambridge Certificate in Advanced English (CAE) were used in the study. They were each required to mark the same 20 CAE scripts which had been selected by the researchers to represent a range of candidate background and a range of proficiency in writing at the upper intermediate level. First, the markers were trained in recording verbal protocols. This was done through exercises carried out at home and a 3 hour centralised training session. The next stage involved them in marking the scripts while recording their thoughts on cassette tapes. Each marker marked all scripts in a prescribed order and at a single sitting (usually lasting 90-120 minutes). All cassettes were transcribed and a coding system was developed.

Analysis and Results
The coding system has four components, each broken down into varying numbers of subcomponents. Two codes were standardised in the use of the system in order to achieve high interrater consistency. All 400 protocols were coded and were analysed using various computer programs. The coding system and the results of the analysis are presented in this study.

Implications
It is hoped that research of this kind will lead to an empirically-based model of the marking process. This will have an impact on the design of writing prompts and rating scales, and on the training of markers to carry out the assessment.

References
Developing and Administering a Cloze Test in American Sign Language

Christine Monikowski, National Technical Institute for the Deaf/Rochester Institute of Technology

Objectives
Interpreters who work with American Sign Language (ASL) and English tend to seek assessment of their interpreting skills from a system established by the national Registry of Interpreters for the Deaf. Assessing their language proficiencies is difficult. Can a cloze test reveal similarities and differences between proficiency levels of second language signers and a control group of native language signers?

Rationale
Osgood and Sebeok (1965) suggested that a cloze procedure might be used for assessing the relative proficiency of a bilingual’s two languages. Subsequent research (Oller et al. 1972) presented evidence that cloze could be used to develop roughly equivalent tests in different languages. This was supported by Xiao (1993) in his research on English/Chinese bilinguals. Lambert (1992) suggested using the cloze technique to assess the level of L2 proficiency for translators and interpreters, in addition to using cloze for their selection and education.

Methods/Design
This paper discusses the development and application of a videotaped cloze test in ASL, including translation of the ASL narrative to English and the English narrative to ASL, the rationale for the ASL deletions, and test administration. A total of fifty-four subjects (both native ASL-users and second-language users, grouped according to interpreting skills) were given the ASL cloze test.

Results/Conclusions
Initial results support these hypotheses: cloze scores for the Deaf people would be higher than the scores for other groups; overall, the L1 scores (of all subjects) would be better than their L2 scores; a linguistic analysis of responses to the cloze tests would reveal richer, more sophisticated use of ASL by the native ASL users; within each of the three groups of hearing interpreters, those who had parents who were Deaf/native ASL-users (CODAs) would tend to score higher than the non-CODAs; and underlying all of the foregoing, interpreting ability is primarily a function of proficiency in each of the two languages.

Implications
Traditionally, signed language interpreter education programs at colleges and universities in the U.S. attempt to teach the interpreting process after only three or four semesters of the language. It is imperative that the skills be separated and students be given sufficient opportunities to develop basic proficiency in ASL. And, it must not be assumed that every native speaker of English has the proficiency in L1 to become an interpreter.
The Role of Cohesion in Communicative Competence as Exemplified in Oral Proficiency Testing

Pavlos Pavlou, Linguistics Department, Georgetown University

This paper examines the significance of cohesion as a part of a speaker's pragmatic competence and consequently of a speaker's overall communicative competence. It investigates whether the development of the ability to produce a cohesive oral text goes hand in hand with the development of overall proficiency in a foreign language.

A learner's competence in a foreign language is considered to be a multidimensional construct (Canale and Swain 1980, Canale 1983). Increased attention to communicative competence in a foreign language has raised interest in the testing of oral proficiency. Therefore, some recent tests of oral proficiency are graded by scales which reflect the multidimensionality of communicative competence (e.g. Bachman and Palmer 1983). However, many of the scales used widely to assess oral proficiency have been criticized because they assume that all the components of oral proficiency are progressing in a linear parallel fashion (Young 1993).

The cohesion of sixty oral reports from an equal number of Cypriot EFL students are analyzed, graded and correlated with an overall grade given to each examinee. The scale used for the grading is a modification of scales proposed by Bachman and Palmer 1983. The scale has three different components (grammatical competence, pragmatic competence and sociolinguistic competence), cohesion being a part of pragmatic competence. The reports are developed from a modular test of oral proficiency which has been developed and tested with EFL students is Cyprus. The modular nature of the test battery is based on the recognition that oral proficiency can be displayed in various speech interactions such as group discussion, oral report, role play and oral interview (Shohamy 1986, Trosten 1992). The analysis of the cohesion is based on the model proposed in Hassan 1976.

The analysis reveals quantitative and qualitative differences in the use of cohesive devices by the learners. Those differences imply that the ability to produce a cohesive text does not necessarily increase at the same pace as other components of proficiency. Grading scales should reflect this differential in the developing components of proficiency.

The results of this study can be useful in developing new and revising existing scales of oral proficiency. Such scales should be based upon the empirical data provided by the student samples rather than upon unsubstantiated theory and intuition.
Predicting Item Difficulty in a Reading Comprehension Test with an Artificial Neural Network

Kyle Perkins, Department of Linguistics, Southern Illinois University
Lalit Gupta, Department of Electrical Engineering, Southern Illinois University
Ravi Tammana, Department of Electrical Engineering, Southern Illinois University

Objective of the Presentation
The paper reports the results of using a three-layered backpropagation artificial neural network (ANN) to predict item difficulty in a reading comprehension test. The purpose of the study was to train an ANN to predict item difficulty values and to compare the actual item difficulties in order to determine whether the two sets of values were statistically similar or different.

Background/Rationale
There is a growing literature that suggests that ANNs outperform traditional statistical procedures such as multiple regression in prediction studies. A reason which has been offered to account for the better performance of ANNs is that backpropagation networks (one type of an ANN) are a form of nonlinear regression and are not bound to the functional fitting inherent in multiple regression which utilizes the least-mean-squared error to determine the best representative function in a data set.

The validity of multiple regression studies to predict item difficulty is not high when few variables account for the significant variance in the dependent variable, as was the case in the study from which the data were extracted for the current study. It was hypothesized that using variables in combination and introducing forms of non-linearity might improve the validity of item difficulty studies. Combining variables and introducing non-linearity can be accomplished in an ANN by manipulating the input variables and by employing the non-linear transfer functions of the neurons in an ANN.

Methods/Design/Procedures/Techniques
Two network structures were developed: one with the sigmoid function in the output processing unit and the other without the sigmoid function in the output processing unit. The data set (consisting of twenty four predictor variables including various counts of text surface structure, propositional analysis of the reading passages and item stems, and cognitive demand of the test items and the corresponding item difficulties) was partitioned into a training set and a test set in order to train and test the ANNs. To determine the consistency of the ANNs in predicting item difficulty, the training and testing runs were repeated four times starting with a new set of initial weights. Additionally, the training and testing runs were repeated by switching the training set and the test set.

Results/Conclusions
The mean squared error values between the actual and predicted item difficulty values demonstrated the consistency of the ANNs in predicting item difficulties, and the Kruskal-Wallis test indicated no significant difference in the ranks of the actual and the predicted values.

Importance/Implications of the Results
The results obtained from the two backpropagation neural networks can consistently predict item difficulty with a high degree of success. The results of training an ANN have direct application to test development and for generalizing about how selected variables affect item difficulty.
Optimal Indices of Gain Score Dependability for Criterion-Referenced Language Tests

Steven Ross, University of Hawaii, Manoa
Te Fang Hua, East-West Center/University of Hawaii, Manoa

Much of the recent work on criterion-referenced language testing addresses the issues of item writing and cut-score dependability. Criterion-referenced item writing is centrally concerned with determining the content congruence and learnability of each item's content. Cut-score dependability focuses on the consistency of decisions in repeated testing or assessment of language learner performances. A more general issue related to language program development also involves empirical rationalization of cut-score decisions. In this general case the issue is of determining the optimal index of gain score dependability in the pre-instruction and post-instruction approach to assessing the language learning gains.

The present paper examines three commonly used approaches to assessing gain score dependability, observed differences between pre-instruction scores and post-instruction scores, an approach using residual scores, and a base-free gain index. The optimal index of gain score dependability is derived from examining the cut-score dependability of the pre-instructional administration of the criterion-referenced test, as well as the post-instructional CRT, in relation to differences in the ratio of pre and post instruction variances.

The database for the present paper comes from a pre-instruction administration of an academic listening test, followed by a counterbalanced post-instruction administration of an alternate form of the same test after one semester of instruction. The subjects were 213 advanced ESL learners at a large American university English language institute.
The Effect of Varying the Immediate Recall Protocol Procedure on Recall Measures of Listening Comprehension

Sheryl V. Taylor, Ohio State University

Objectives
This study was designed to investigate the effect that the generally accepted task sequence of the Immediate Recall Protocol Procedure (IRPP) has on recall measures of listening comprehension. This task sequence was altered in this investigation by the number of listenings permitted, the sequence and number of written recalls completed, and the listener’s opportunity to have access to a "practice" recall during the second listening/second written recall. The listener's language ability (its effect in listening comprehension, as well as its effect in each task sequence variation of the IRPP) was also considered as a secondary factor.

Background/Rational
Due to the absence of a widely accepted testing instrument that is compatible with the current knowledge about second language (L2) listening, many researchers have begun to use the Immediate Recall Protocol Procedure as a measure of listening comprehension. Until recently, researchers were utilizing the recall format that was accepted in L2 reading research (i.e., two consecutive exposures to the reading passage followed by a written recall). Recommendations have been made and implemented to vary this format when it is used as a measure of L2 listening comprehension by including an additional "practice" written recall to be completed after the first two listening exposures. Despite researchers' agreement to use the IRPP as a measure of listening comprehension, they lack a consensus concerning the most suitable task sequence of the procedure. They also have neglected to consider or question the effect that varying this procedure has on measures of written recall. Consequently, the empirical results of listening studies have been generalized regardless of the IRPP task sequence procedure that was employed in data collection.

Methods/Design
A group of 120 third-quarter university students of Spanish were randomly assigned to one of the following four task sequence procedures of the IRPP: (1) listen, recall; (2) listen, listen, recall; (3) listen, recall (recall collected), listen again, final recall; (4) listen, recall (student keeps recall), listen again, final recall (student is able to refer to preliminary written recall during the second listening and final recall). The recall scores were statistically analyzed by an analysis of covariance in order to account for (1) the influence of IRPP task sequence procedure, (2) the influence of language ability on students' listening comprehension scores, and (3) the effect of interaction between the two aforementioned factors.

Results/Conclusion
Results indicated that the main effect of IRPP task sequence procedure was statistically significant. In fact, group means for the procedure demonstrated that students of one variation in particular had the highest scores, and subsequent multiple comparisons tests also revealed a significant difference for one of the four procedure variations. The secondary factor of language ability also reached statistical significance. Interaction between the two factors, however, was not significant.

Importance/Implications
Findings of this study provide strong evidence that varying the task sequence procedure of the IRPP does affect recall measures of listening comprehension. Consequently, procedure selection of the IRPP clearly becomes an important consideration in data collection. It is also imperative that researchers investigate the specific task sequence procedure of the IRPP utilized in studies of listening comprehension before they attempt to generalize the empirical results across studies. Results of this study indicate that students scores were higher for one particular task sequence procedure of the four employed. Perhaps this procedure is the most suitable when the IRPP is employed as a measure of listening comprehension. Indeed, findings of this study provide a point of departure regarding researchers' determination of the most appropriate variation of the IRPP to be used as a measure of listening comprehension.
The Use of Role Play in Assessing Oral Ability: The Effects of Situational Context on Measurement

Alexander J.H. Teasdale, Thames Valley University

Theoretical Background and Rationale
Role play is frequently used in the oral testing of English for Specific Purposes. Because of difficulties in simulating authentic contexts, this approach often requires testees to utilize their imagination, thus introducing potential for bias in the measurement.

Purposes of the Research
The research questions addressed are:

- is there a quantifiable difference between oral interaction items which provide a sufficient context for testees and those which are context-reduced and therefore require the testee to use their imagination
- does this variation within the technique of role play allow for consistent measurement across all testees
- is there evidence of bias in favor of candidates who are able to accommodate their performance more successfully to context-reduced interaction

Research Design and Methods
A semi-direct test of English for Occupational Purposes (in which the testee responds to taped and visual cues) was investigated with sample sizes of n = > 100. Each performance contains 25 items and was rated on two dichotomous rating scales covering Appropriacy and Performance. Each item was also judged as +/- sufficient context. The data were then analysed using multi-faceted Rasch measurement (Linacre, 1989).

Results
The results investigate the relative effects on measurement of context-sufficient and context-reduced interactions in role play.

Implications of the Results
The findings have implications for the design of tests which use role play in the assessment of oral ability, and particularly for performance tests of English for Specific Purposes in which context has a strong determining influence.

Reference
Using FACETS to Model Rater Training Effects

Sara Cushing Weigle, UCLA

Rater training is often cited as one of the most important factors in getting composition raters to rate essays reliably. Although a number of studies have shown that trained raters are more reliable than untrained raters (e.g. Shohamy 1992), some researchers have begun to express skepticism over the stated goals of rater training, saying that no amount of training can overcome inherent differences among raters in terms of severity. The function of training should instead be to get raters to judge with internal consistency (Stahl & Lunz 1991). If raters learn to be consistent in their ratings, then any differences in overall severity among raters can be compensated for mathematically, using, for example, the computer program FACETS (Linacre 1989, 1992), an implementation of a multi-faceted Rasch approach to measurement which takes into account different facets in a measurement situation such as rater severity and task difficulty when estimating examinee ability. However, the extent to which training affects either the severity or the consistency of individual raters has not yet been demonstrated empirically through a pre/post design.

This paper describes a study investigating some effects of training on raters of ESL compositions at UCLA. Sixteen raters, eight experienced and eight inexperienced, rated subsets of 30 essays from a set of 60 ESL placement compositions (30 each on two topics) using a three-part scoring rubric which included content, rhetorical control, and language. All raters then went through the regular rater training process and participated in a live rating session at which some 800 placement compositions were scored. Following the live rating, the raters rated a different subset of 30 compositions from the original set of 60.

The data were analyzed using FACETS, with the facets of Examinees, Raters, Topics, and Subscales for both the Pre-training and Post-training data. The research questions for the study were as follows:

1. To what extent did rater training affect the severity of individual raters?
2. To what extent did rater training affect the spread of rater severities (i.e. make raters more like each other in terms of severity)?
3. To what extent did rater training make raters more consistent in their judgments?

Among the results to be presented are the following:

(a) The inexperienced raters tended to be more extreme (more harsh or more lenient) than the experienced raters before but not after training.
(b) Although there was a slight decrease in the rater spread after training, significant differences in severity among raters were still present after training.
(c) As predicted, inexperienced raters were more inconsistent in their ratings than experienced raters before training, but not after training.

Some theoretical and practical issues regarding the use of FACETS for this kind of study will be discussed, and implications of the study for rater training and composition testing will be presented.
The Effect of Information Gap on Elicited Discourse of Candidates in an Oral Test of English

Gillian Wigglesworth, University of Melbourne

Background/Rationale
The elicitation of interactive discourse in a test situation is the aim of many oral interaction tests. Task type has been shown to elicit variable patterns of language (Duff 1986, Long 1989), and there is a need for closer examination of the effects of variation in task type in the context of oral tests.

Objectives of Research
This study investigates the effect of task type (information gap or lack of information gap) on the resultant candidate discourse, the input of the interlocutor, and the results obtained by the candidate.

Method
The test, designed for prospective immigrants to Australia, was undertaken in a trial situation by 110 candidates, distributed equally between two conditions in which two tasks were manipulated. In the first task, the candidate was required to describe the differences between two pictures. For this task, half the interviewers had access to both pictures, while half had access to only one of the pictures. In the second task, the candidates were required to describe a process from a series of eight pictures. Again, half the interviewers had the pictures whilst the other half did not. The experimental design was such that all interviews had an information gap in one of the tasks. The resultant discourse was analysed to provide measures of candidates’ linguistic output and the interactive nature of the discourse.

Results
It was found that where there is an information gap, the nature of the candidate discourse may differ in quality from the discourse elicited where no information gap exists. In addition, the detailed analysis demonstrates that the interactive nature of the discourse between tester and testee may vary according to the condition.

Implications
The implications of these findings for test task design and for the choice of rating criteria for oral interaction tests are discussed.

References

Abstracts of Poster Session Presentations

(in alphabetical order by first author)
Language Testing with Speech Recognition: Methods and Validation

Jared Bernstein, Entropic Research Laboratory

Purpose
The research objective has been to test the feasibility of automatically evaluating the pronunciation of English sentences read aloud by non-native speakers of English. We sought a method for automatically deriving pronunciation scores for spoken sentences that correlated well with those assigned by human expert listeners. In principle, such a method would be useful not only for testing students' spoken language skills, but also as an element in a more comprehensive system providing diagnosis and/or instruction.

Background
Speech recognition systems improved significantly in the 1980s. One outcome of the new level of recognition performance, particularly speaker independence and vocabulary independence, has been that some new applications become much easier to develop.

Method
A series of studies (carried out at SRI International between 1987 and 1993) tested the performance of automatic systems that score the speech production of non-native speakers of English. The systems undertook increasingly more difficult tasks and signal conditions: first, Japanese adults recorded with a high quality microphone; then Japanese middle-school children speaking over normal telephone channels, and finally, speakers of diverse backgrounds recorded under heterogeneous conditions. The automatic grading procedures first align the speech with a model and then compare the segments of the speech signal with models of those segments that have been developed from a large database of native speech. In the final stages of the Japanese middle-school project, the system was validated on a data base of 5,000 utterances. In all the studies, we measured the performance of the automatic systems by calculating a correlation between automatically generated scores and the average of the scores assigned by three expert human listeners.

Results
With high quality speech signals, correlations of overall speech production quality have been as high as 0.9 (n = 200). We have achieved a correlation of about 0.8 with telephone speech (n = 2000). The poster will discuss the relation of signal duration and signal quality to grading reliability, and its implication for the diagnosis of specific speaking characteristics.

Implications
Primarily, these studies offer hope that meaningful measures of spoken language performance can be generated automatically at low cost and with wide geographic coverage. Among several challenges remaining, the most central is to generalize the methods to work with a wider range of first and second languages.
Criterion-Referenced Language Test Development (CRLTD): An Overview

Fred Davidson, Division of English as an International Language, University of Illinois
Brian Lynch, Department of Applied Linguistics and Language Studies, University of Melbourne
Dongwan Cho, Division of English as an International Language, University of Illinois
Susan Larson, Environmental Engineering and Science Program, Department of Civil Engineering, University of Illinois

Objectives
This poster session illustrates a process called 'Criterion-Referenced Language Test Development' (CRLTD).
This process allows flexible, iterative language test development among a group of educators. It is intended to enhance content evidence of validity.

Background/Rationale
Criterion-Referenced Measurement (CRM) has been examined extensively in the language testing literature. Several studies have explored CRM statistics as a means to analyze CRM tests. However, language test development does not yet have a concise formalization of CRM test development. That formalization is the goal of CRLTD.

Technique
CRLTD principles contend that test development tends to disempower many interested educators; it is a process of external forces, e.g. social, educational, and personal mandate. To counter this, CRLTD is conducted by groups of concerned educators who have a stake in the production of reliable and valid measures. In the CRLTD process, the groups each produce a CRM specification (modeled after Popham, 1978). The groups then swap specifications and attempt to write an item/task from each specification. Upon reconvening as a whole group, typically, the items/tasks do not match the specification. The entire process is repeated until the groups communicate more readily their intended measurement goals.

This poster session includes several short paragraphs summarizing further the consensus-based empowerment of CRLTD. It also includes several examples of the evolution of test specifications, from earliest miscommunication to later agreement among and within test development groups. These examples include specifications developed in language testing classes and testing workshops. Also included is the evolution of the specification for the University of Illinois Video/Reading-based Essay (VRESSAY) exam. The VRESSAY is an ESL placement measure which uses as prompt material a videotaped lecture and associated reading passage. Examinees write an essay drawing from both sources. Samples from the evolution of this specification include video-taped excerpts.

Importance/Implications
There are several ways in which CRLTD has important implications for language test development:

Through its group consensus technique, CRLTD allows a voice to persons not typically involved in test development, most critically, classroom teachers (as when devising a placement or achievement test for a language institute).

It provides a means to instruct teachers in language testing. The CRLTD group development process can be run in a language testing class or seminar, even when language students are not available for a practicum. The language testing teacher trainer can impart a sense of the flexibility and feedback of all good test development by having the trainees experience the feedback nature of the CRLTD spec-item writing loop.

Finally, by appeal to the nature of a CRM specification, CRLTD should enhance content evidence of validity.
T-LAP: Test of Listening for Academic Purposes

Christa Hansen, University of Kansas
Christine Jensen, University of Kansas

General Test Description and Purpose
In this session, we would like to introduce the T-LAP, a content-based listening test we have developed as a screening and placement test. It is used to screen students for university classes and to place students in appropriate levels of listening instruction if they are not ready for university classes.

Population
The test population includes both undergraduate and graduate students from as many as 60 different countries.

Theoretical Constructs
The T-LAP is a listening comprehension test designed to take advantage of the cognitive constructs that listeners use as they make meaning from sound. According to van Dijk and Kintsch (1983), the decoding of the audio input is begun in the short-term memory with the aid of information retrieved from the long-term memory from the textbase (semantic representation of input) and the situational model (experiential aspects of the text).

For listeners to be successful in their academic lectures and real life situations, they need to be able to use (integrate) all these parts of the process. Therefore, on the test, listeners are asked to perform tasks that go beyond recording the short-term audio input. Listeners are asked to extract information from within clauses, recognizing key ideas and the interrelationships of ideas.

Test Description
The T-LAP has 2 major components, an academic and a nonacademic. In the nonacademic section, listeners hear a series of dialogues on a single theme. In the academic section professors deliver excerpts of their lectures. Each form has one lecture in a technical and one in a non-technical discipline. The context is set for each part of the test before listeners hear the discourse. In addition, time is allowed for listeners to preview questions to allow them to use predictive listening skills by accessing semantic and situational information in the long-term memory. Listeners write their responses to short-answer questions in real time to eliminate memory storage problems. In the lecture section, the questions are detail and global questions. Test takers answer detail questions on a first play through, while global questions are responded to on the second time through.

Test Analysis
We have run descriptive statistics (including Cronbach alphas) and item analyses for the pilots of five forms using SPSS, revised those forms, and equated two revised forms using Quest. These analyses show the test to be a reliable instrument.

Structure of Session
We plan to describe the T-LAP briefly in our five-minute introduction. The display will include a test description, a sample of the test, and the results of the statistical analyses of the different forms.

Reference
A Multistep Placement Test

Michel Laurier, University of Montreal

Objectives
This poster session will present a placement test in French as a second language intended for secondary and post-secondary students enrolled in an intensive language program. The presentation will focus on:

1) the implementation of adaptive testing concepts on a paper-and-pencil test,

2) the use of video technology to provide a natural context and give uniform directions,

3) the adaptation of the ACTFL approach to a placement situation in a Canadian context.

Background
The purpose of the project was to provide the institutions participating in the program with a common placement test that could cope with the administrative constraints of a placement decision, take into account the student's French background and meet the objectives of the program which aims at developing oral skills.

Design
The whole test consists of two parts: a group administration that uses a video program and an oral interview with two students. In order to find the beginners, the first part begins with a short sentence comprehension section. The beginners will immediately go for the oral interview, whereas the others will stay for the rest of the group administration which includes aural comprehension questions based on semi-authentic video passages and multiple-choice questions where the student is asked to select the most appropriate statement. Based on the group administration results, the students are paired for the confirmatory oral interview that can be conducted at the beginners' level, the intermediate level or the advanced level. The first part has been created using IRT techniques. For the rating of the second part, a seven band system has been set up.

Importance
Two parallel forms are now being used. Although some institutions question the practicality of the multistep approach, it is considered as a very helpful and innovative placement tool. This test has also set up standards for the language program.
What is Good Enough? Exploring Some Aspects of the Validity of Testing Speaking

Sari Luoma, Language Centre for Finnish Universities

The objective of the study is to find the optimal combination of tape-mediated and face to face tasks for validly measuring intermediate (upwards of Threshold) level speaking in a national four-skills proficiency test. The test sought would cover use, usage and knowledge of conventions of spoken English, and provide enough evidence for reliably assessing where on a range of three (or in some cases five) bands of a nine-level system a testee is with regard to speaking. Apart from providing enough evidence for the assessor to make a reliable rating, the solution should be practical and cost-effective, and acceptable to the testees. The possible effects of the test on teaching must also be considered.

The study was triggered when, during the piloting of a previous version of the test, it was found that the tape-mediated speaking test did not provide all the data necessary for validly reporting proficiency in speaking as the construct was defined in the test specifications. The test was essentially non-interactive and measured language knowledge and knowledge of usage. Ability to use the language interactively was, however, the first definition of speaking ability in the specifications. Therefore, exploring the potential benefits of adding a face to face interaction test (paired candidates) was felt to be advisable, while the possibility of changing the tape-mediated test to better correspond with the definition was also recognized.

During the investigation, the two tests, the face to face interaction test and the improved tape-mediated test, are compared and contrasted with each other and with the definition of speaking as it now stands. The analysis is divided into three parts: response analysis, feedback analysis and assessment analysis. Audio and video recordings and transcripts of the test performances as well as questionnaires and some interviews are used as data. The assessments are studied both quantitatively and qualitatively through a structured interview procedure.

Initial results indicate that the linguistic differences between the kinds of speech elicited center on the degree of interactivity of the two test modes. The foci of elicitation, and consequently the assessment procedures, also differ between the two tests: the tape-mediated test concentrates on language functions and their fulfillment, and assessments are mostly made turn by turn, while the discussion test focuses on ability to handle content areas interactively, and the assessment task is approached more holistically. There seems to be a high correlation between the assessments made in each test type, but it also appears that the test-takers much prefer the face to face test to the tape-mediated one. There are some cases where ability level assignments differ between the two test modes. Assessment analysis is still in progress in January 1994.

The results speak for the inclusion of a face to face interaction test into the test battery, but not for the exclusion of the tape-mediated test unless substantial increases into the test price can be accepted. Suggestions for improvement of the tasks and rubrics were obtained from feedback by testees and assessors.
The Relationship Between Type and Frequency of Errors and ACTFL Level on the SOPI

Margaret E. Malone, Center for Applied Linguistics

Objectives
This poster investigates the relationship between type and frequency of errors and an examinee's ACTFL rating on the simulated oral proficiency interview (SOPI). Five types of errors were coded: verb tense errors; subject/verb and noun/modifier number agreement errors; noun/modifier gender agreement errors; register errors and vocabulary errors. Transcriptions of 15-item SOPIs of nine examinees at three ACTFL levels (Intermediate, Advanced, and Superior) were analyzed.

Background/Rationale
The ACTFL Guidelines frequently refer to errors made by examinees at different ACTFL levels. The extent of errors made, however, is never explicitly defined, nor are raters trained to count the number and types of errors made by examinees when rating SOPI tapes. This poster, then, investigates whether a relationship exists between frequency of error and ACTFL level and type of error and ACTFL level.

Methods
Tapes of nine examinees taking the Texas Oral Proficiency Test (TOPT), a Spanish simulated oral proficiency interview, were transcribed. The transcriptions were then coded for the five errors described above. A Chi-Square was performed to determine the relationship between frequency and type of error and ACTFL level.

Results
This poster presents the results of the Chi-Square, which indicate that certain error types are more among speakers at low-levels than others. For example, vocabulary errors, agreement number/agreement errors and gender/agreement errors showed the only significant relationship with ACTFL level. Verb tense and register errors showed no significant relationship with ACTFL level.
A Study of Writing Tasks Assigned in Academic Degree Programs: A Stage II Report

Carol Taylor, *Educational Testing Service*
Gordon Hale, *Educational Testing Service*

**Theoretical Background and Rationale**
In order to develop a valid writing test intended to measure the kinds of academic writing skills college-level students are expected to demonstrate, it is important to identify the kinds of writing tasks that are actually required of college and university students.

**Purpose of the Research**
The purpose of the present study, conducted by Gordon Hale, Carol Taylor and Brent Bridgeman in consultation with Joan Carson, Barbara Kroll, and Robert Kantor, was to collect and examine assigned writing tasks in order to expand the understanding of the types of writing tasks that are actually required and to provide a basis for developing valid essay topics and tasks for inclusion in an academic writing test.

**Research Design and Methods**
Data for the research were obtained from university teachers from one Canadian and seven U.S. universities, both at the undergraduate and graduate levels. The selected institutions varied with respect to geographic location, size and control (i.e., public vs. private), with particular emphasis on the types of institutions that draw large numbers of international students. Within each institution, disciplines were selected to represent general areas in which the greatest numbers of international students study or areas that form part of the core curriculum. The research team examined the materials and developed a writing classification scheme that characterized the different types of writing encountered across disciplines and then used the scheme to classify all writing assignments.

**Results and Implications**
One hundred seventy-six teachers participated in the study and provided 184 sets of course materials. This represented approximately one-third of those invited to participate, after excluding those who responded that no writing was assigned in their classes. The data were examined by the project team and consultants. The first stage of the project, data collection and development of a classification scheme was reported at LTRC 1993. The classification scheme included five areas of classification: locus of writing, length of product, genre, cognitive demand, and rhetorical specification. The main focus of this poster is on stage two efforts of the project: application of the scheme to all assignments, data analysis and results. Findings and implications for the assessment of academic writing will be discussed.
Construct Validation of Measures of Communicative Effectiveness and Grammatical Accuracy: A Multimethod Approach

John A. Upshur, *Concordia University*
Carolyn E. Turner, *McGill University*

Objectives
This poster session presents the results of a 3-way validation study of L2 tests of communicative effectiveness (CE) and grammatical accuracy (GA). The purpose of the study was to validate tests developed to measure one general L2 trait, CE, and one putative component of that trait, GA. The approach taken was construct validation in which sets of tests of different abilities should exhibit both convergent and divergent validity.

Background
In instructional settings, test formats must be compatible with teaching practices. Aspects of language or language use that teachers are interested in and have taught are generally the focus of testing. These traits of interest are likely to be highly correlated. Classroom tests are frequently characterized as criterion-referenced and need to be sufficiently brief to accommodate school class schedules. They typically rely upon evidence of face and content validity. At times, however, teachers use test information diagnostically to identify students who may lag on some trait of interest and then prescribe individualized instruction. This use of scores requires evidence of divergent validity of the measures used in diagnosis.

Method
In this study, tests were administered to 130 French speaking grade 5 learners in Intensive ESL classes in Montreal, Quebec. Three construct validation procedures were then employed:

1. Campbell-Fiske (Campbell & Fiske, 1967) with traits and tests as method;
2. LISREL (Jöreskog & Sörbom, 1989) with traits and factors of method;
3. Outlier analysis for traits (Boldt & Oltman, 1993).

Results
(1) High correlations were found among measures. (2) Campbell-Fiske criteria for divergent validity were not satisfied because of high r’s in comparison with reliabilities. (3) The data did not fit any well-motivated LISREL model; high intercorrelations led to m (phi) matrices that were not positive definite. (4) The outlier analysis provided evidence for two traits.

Implications
The results suggest that the most commonly used methods for construct validation of L2 tests may prove insensitive to divergent validity when tests are highly correlated and variance is restricted, as in instructional settings.

References
Abstract of Experimental Sessions:
Roundtables, Workshops, Site Visit

(in alphabetical order by first author)
Language Competence in the Federal Government: An Articulation of Proficiency and Performance Assessment

ROUNDTABLE DISCUSSION LEADERS

Eduardo C. Cascallo, Center for the Advancement of Language Learning
Marijke W. Cascallo, Federal Bureau of Investigation
John L. D. Clark, Defense Language Institute
Madeline Ehrman, Foreign Service Institute
Pardee Lowe, Jr., ILR Language Testing Committee
Julie A. Thornton, Center for the Advancement of Language Learning
Bernard Spolsky (discussant), Bar-Ilan University

Objectives
In this session, LTRC participants will have the opportunity, within a small-group format, to discuss current directions in federal language assessment procedures with testing representatives from the Center for the Advancement of Language Learning and of those federal agencies most involved in cooperative language testing efforts. The discussion will center around the issues of proficiency and performance assessment needed to satisfy the requirements of personnel selection, assignment, and training to fulfill the missions of the respective federal agencies. The goal is to have a highly interactive session with the participation of interested attendees.

Background
The Center for the Advancement of Language Learning (CALL) was created in 1992 as a part of a congressional effort to improve the foreign language capability of the US government. While its first aim is to strengthen government language teaching and testing, CALL is also the government’s link to the academic and business language communities. In the area of language testing, CALL and the government language schools have set up a national-level Language Testing Board to coordinate test development, conduct reliability studies, and serve as a resource on testing methods for government and non-government organizations.

Session Design
The session will open with a presentation on the theoretical basis for the development of a profile approach for language competence assessment and will provide a framework into which each of the other participants’ presentations may be fit. Next, each of the remaining participants will outline cooperative projects among federal agencies collaborating at CALL as well as describe agency-specific testing activities in which they are currently involved. Within the last year, a number of cooperative activities in the area of language testing have begun, as well as ongoing and future projects which include the participation of various academic institutions and individual researchers. These projects address topics in a wide variety of areas, and two projects completed in the past year include a set of new level descriptors for translation and a system for the computerized analysis and scoring of oral proficiency in non-native English speakers. Ongoing and future projects include the following:

- nationally-administered language needs/use survey for government agencies,
- interagency comparability study of speaking testing,
- multimedia computerized testing tool development,
- listening summary translation test development and validation (in Chinese & Arabic),
- background questionnaire research,
- ESL writing proficiency test development and validation, and
- language aptitude research into cognitive, affective, and motivational factors.

At the conclusion of the session, the discussant will summarize the presentations and offer a critical review, within the framework of proficiency and performance assessment of language competence theory, of the tasks accomplished and planned by CALL.
CALL Site Visit and Testing Demonstrations

ORGANIZER

Eduardo Cascallo, Center for the Advancement of Language Learning (CALL)

Session Design
In a site visit, LTRC participants will travel from the LTRC convention site in Washington to CALL's facilities in Arlington, VA. They will receive a tour of CALL's facilities, a brief introduction to CALL's goals and activities, and demonstrations of CALL's ongoing testing procedures for speaking. Visitors will observe mock oral proficiency interviews conducted in face-to-face interviews and over the telephone as well as via CALL's distance/video teletesting equipment via satellite connection with other locations within the US. In addition, they will participate in the electronic monitoring and recording of actual testing sessions, learn about the procedures for the development and collection of exemplary oral proficiency interviews, and view the operation and format of CALL's testing database.
Testing Aptitude Testing: The MLAT

ROUNDTABLE DISCUSSION LEADERS

Madeline E. Ehrman, Foreign Service Institute
Lucinda Hart-González, Foreign Service Institute
Frederick H. Jackson, Foreign Service Institute
Joseph N. White, Foreign Service Institute

The Modern Language Aptitude Test (MLAT) was developed in the late 1950s, largely with federal government language training in mind (Carroll, 1959). One government agency, which was a test site during its development, has used the MLAT ever since. In the thirty-five or so years since its original implementation, so much has changed in the theoretical assumptions underlying the MLAT that the agency initiated a broad-based re-examination. That initiative coincides with an ongoing wider interest in language aptitude testing at all agencies in the federal language testing community and in the teaching community as a whole (e.g., Parry and Stansfield, 1990).

This roundtable will serve as a forum for integrating several avenues of thought on language aptitude in general and the MLAT in particular.

1. review of issues in the literature on aptitude in general, and language learning aptitude in particular, such as: (a) is language learning aptitude different from general cognitive aptitude (e.g., Schneiderman and Desmarais, 1988); (b) does aptitude change; is it learnable (e.g., Angoff, 1988); (c) is aptitude a valid and useful unitary construct or is it more useful to look at it in terms of the interaction of various traits (e.g., Wesche et al., 1982, Gardner, 1991)?

2. studies completed and in progress on the MLAT in federal language training, include (a) profiling of students at the high and low ends of the aptitude scale in terms of background, learning styles, and training outcomes; (b) following the progress throughout training of learners at different levels of aptitude as measured by the MLAT; (c) assessing the power of the MLAT in comparison with other known predictors such as general level of education, previous language background, etc., to predict levels of success under different teaching methods and for different language types.

3. discussion of the questions: based on the kinds of information that are available from aptitude testing, what kinds of applications can be made of aptitude testing in the federal training context, e.g. diagnosing learner problems, student counseling?

The roundtable presenters are members of the agency MLAT Research Committee, each of whom has contributed to one or more of the studies in (2) above. Each presenter will discuss on one of the discussion questions of a theoretical nature from (1) above, or of an applied nature from (3) above. Research projects and findings will be presented in the context of the discussion questions, and audience participants will be invited to add other findings to the discussion. The end result should be the beginnings of a consensus on the scope and direction of response to each question, together with an exchange of research findings and sources.
Comparing Language Qualifications in Different Languages: A Framework and Code of Practice

ROUNDTABLE DISCUSSION LEADERS

Michael Milanovic, University of Cambridge
Nick Saville, University of Cambridge

There is a considerable amount of interest currently in the notion of a framework of levels of proficiency for foreign languages. This was the subject of a workshop, specially convened by the Council of Europe, in Strasbourg in October 1993.

Developing any such framework is fraught with difficulties. Do we compare on the basis of usage, recognition, content or statistical equivalence? How do we take into account different educational traditions and practices? Is there a requirement for an internationally accepted code of practice? If so, what should it focus on?

The initial presentation in this discussion will describe and explain the work of the Association of Language Testers in Europe (ALTE). This is an association of 10 European providers of foreign language qualifications who are actively engaged in the creation of a comparative framework and a code of professional practice. The work that they have completed to date and their proposed plan of action will be reviewed.

It is hoped that the roundtable discussion will provide critical feedback on the work carried out so far by ALTE and help in the formulation of future directions. A discussion paper will be available in advance for those interested.
Research on the Properties of Alternative Assessment Procedures

ROUND TABLE DISCUSSION LEADER

Elana Shohamy, Tel-Aviv University

Alternative assessment procedures have been the most recent development in the testing field in the past few years. Procedures such as portfolios, self assessment, documents, interviews and observations have been widely used to assess students language development and language proficiency. Yet, in spite of their appeal in terms of content validity, there have been serious questions, and often strong criticism, as to their various properties, especially those involved psychometric properties such as reliability and validity. The purpose of the roundtable is to raise a number of issues and discuss alternative ways for investigating these properties as they relate to the ratings, diagnosis, decisions and recommendations based on such procedures.

The specific assessment battery that will be discussed in the roundtable as an example will be a new battery just introduced nationally to assess the language acquisition and proficiency of immigrants children in grades 3 to 12. The assessment battery includes four assessment methods: a test, a portfolio, a student self assessment, and teachers' evaluation based on observations of the students' linguistic performance in class in a number of school subjects. All the above information is then analyzed by the teacher/tester in an assessment conference and is converted into diagnostic profiles and a set of pedagogical recommendations. The battery is currently being trialed on an experimental basis. A framework based on Nevo and Shohamy (1986), which focuses on four criteria (utility, accuracy, feasibility and fairness), will be used to collect data on the quality of the different assessment methods as well as on the battery as a whole. Thirty schools are participating in the experiment, sixty immigrant students, of elementary, junior-high and high school. The plan is for the different analyses to be performed on each of the four assessment methods and on the whole process of assigning a rating, giving the diagnostic profile and providing a set of recommendations. The discussions and issues raised in the roundtable will focus on the type of information needed for examining the quality of alternative assessment methods and on the suitability of traditional and classical test analyses procedures to these new assessment procedures and at the same time to come up with innovative ways of analyses which may be more suitable to these types of procedures.
Independence: A Lurking Assumption of the Statistical Models used in Language Testing

WORKSHOP LEADERS

Bruno D. Zumbo, University of Ottawa
Tim Pychyl, Carleton University
Janna Fox, Carleton University

The objective of this experimental session is to familiarize participants with a fundamental and lurking assumption of the statistical models used in language testing. Of all the assumptions of the statistical models, independence is possibly the most unforgiving for both measurement theory and inference during data analysis. Rozeboom (1966) expresses the centrality and pervasiveness of this assumption when he states that for practical applications, assuming independence is akin to "putting on a clean shirt to rassle a hog" (p. 415). This session will present recent findings in both measurement theory and data analysis (e.g., Yen, 1993: Zimmerman, Williams & Zumbo, 1993; Zimmerman, Zumbo & Lalonde, 1993; Zumbo & Zimmerman, 1991; Zumbo, 1993) which show that (1) coefficient alpha is overestimated (hence altering the standard error of measurement) under violation of the assumption of uncorrelated subtest error scores, (2) the information function and standard error of measurement of item response models (IRT) are altered, and (3) in language testing research, the type I error rate and confidence intervals of statistical tests of hypotheses are drastically altered by nonindependence. The problem of nonindependence is an important issue in the consideration of the statistical models used in language testing. Scenarios in language testing which are likely to produce nonindependence as well as methods for managing the nonindependence will be discussed.

This session will be didactic in nature. The format of the presentation will resemble a workshop. We will begin the presentation from basic principles assuming little to no knowledge of the issues of independence and little knowledge of statistical models. It will be assumed, however, that the participants have some familiarity with classical reliability theory and item response models. To facilitate the discussion, handouts will be available and scenarios from language testing will be used to demonstrate the potential problems. The session is planned for approximately 75 minutes.

References