19th Annual
Language Testing Research Colloquium

Theme: Fairness in Language Testing

Program and Abstracts

Holiday Inn Select
Orlando International Airport
Orlando, Florida, USA
March 6-9, 1997
Acknowledgments

LTRC '97 Steering Committee

Antony John Kunnan, LTRC '97 Program Chair
Mary Spaan, LTRC '97 Program Chair
Ari Huhta, LTRC '96 Program Chair
John Clark, LTRC '98 Program Chair

LTRC '97 Organizing Committee

Antony John Kunnan, Program Chair, California State University, Los Angeles
Mary Spaan, Program Chair, University of Michigan
Sarah Briggs, Associate Chair, University of Michigan
Barbara Dobson, Associate Chair, University of Michigan
James E. Purpura, Associate Chair, Teachers College, Columbia University

Organizations

English Language Institute, University of Michigan, Ann Arbor, MI
TESOL Program, California State University, Los Angeles, CA
TOEFL Program, Educational Testing Service, Princeton, NJ
EFL Division, University of Cambridge Local Examinations Syndicate
International Language Testing Association

Individuals

Lyle Bachman, University of California, Los Angeles
José Galván, California State University, Los Angeles
Robbie Kantor, Educational Testing Service
Michael Milanovic, University of Cambridge Local Examinations Syndicate
Simeon Slovacek, California State University, Los Angeles
John Swales, University of Michigan

Volunteers

From the TESOL Program, California State University, Los Angeles:
Yutaka Kawamoto, Beryl Meiron, Carmen Velasco-Martin

From Teachers College, Columbia University:
Susan Stempleski

From the English Language Institute, University of Michigan:
Theresa Rohlck, Eric Ström
Acknowledgments

Abstract Evaluators

Lyle Bachman, University of California, Los Angeles, USA
Eduardo Cascallar, City University of New York, USA
Carol Chapelle, Iowa State University, USA
Caroline Clapham, Lancaster University, UK
Alister Cumming, Ontario Institute of Education, CANADA
Fred Davidson, University of Illinois, Urbana-Champaign, USA
Dan Douglas, Iowa State University, USA
Pat Dunkel, Georgia State University, USA
Liz Hamp-Lyons, Hong Kong Polytechnic University, HONG KONG
Grant Henning, Pennsylvania State University, USA
Dorry Kenyon, Center for Applied Linguistics, USA
Brian Lynch, University of Melbourne, AUSTRALIA
Michael Milanovic, University of Cambridge Local Examinations Syndicate, UK
Jacqueline Ross, Educational Testing Service, USA
Bernard Spolsky, Bar-Ilan University, ISRAEL
Carol Taylor, Educational Testing Service, USA
Carolyn Turner, McGill University, CANADA
John Upshur, Concordia University, CANADA

and

The LTRC '97 Organizing Committee
PROGRAM
Thursday, March 6

9:00 – 1:00  Pre-Colloquium Workshop I
International Ballroom A

An Approach to the Design and Development of Language Test Tasks

Presenters:
Lyle Bachman, University of California, Los Angeles
Adrian Palmer, University of Utah

2:00 – 6:00  Pre-Colloquium Workshop II
International Ballroom A

Statistical Data Handling: A Principled Process

Presenter:
Fred Davidson, University of Illinois, Urbana-Champaign

4:00 – 7:00  On-Site Registration
Atrium Hallway outside International Ballrooms A & B

7:00 – 9:30  Welcoming Reception
Lower Atrium

Co-hosted by the English Language Institute, University of Michigan and
the International Language Testing Association (ILTA)
PROGRAM
Friday, March 7

8:00 – 5:00  REGISTRATION
Atrium Hallway outside International Ballrooms A & B

8:30 – 10:00  Session 1 – Opening Plenary
International Ballroom A

Welcome
Antony John Kunnan, California State University, Los Angeles
Mary Spaan, University of Michigan

Introduction of Plenary Speaker
Jacqueline Ross, Educational Testing Service

A “Postmodern” View of the Problem of Assessment or Why Do I Get Such a Headache Thinking About Test Design?
Henry Braun, Vice President for Research Management, Educational Testing Service, Princeton

10:00 – 10:15  Break

10:15 – 12:00  Session 2 – Panel with Open Discussion
International Ballroom A

Panel: Fairness in Language Testing

Chair and Moderator: Antony John Kunnan, California State University, Los Angeles

Panelists
Lyle Bachman, University of California, Los Angeles
Liz Hamp-Lyons, Hong Kong Polytechnic University
Bonny Norton-Pierce, University of British Columbia
Elana Shohamy, Tel Aviv University

12:15  Group Photograph

12:30 – 1:30  Lunch Break
Friday, March 7

1:30 – 3:00  Session 3 – Papers  
International Ballroom A

Chair: James E. Purpura, Teachers College, Columbia University

1:30 – 2:00  Assessment of language impairment across cultures  
Rosemary Baker, University of Melbourne

2:00 – 2:30  The effect of first language background on the cognitive and linguistic attributes underlying performance on the TOEIC  
Gary Buck, Kumi Tatsuoka, C. Tatsuoka & Irene Kostin  
Educational Testing Service

2:30 – 3:00  Multidimensional analysis and oral proficiency sampling: Performance effects of task and elicitation context  
Jeff Connor-Linton & Elana Shohamy, Georgetown University & Tel Aviv University

3:00 – 3:15  Break

3:15 – 4:45  Session 4 – Papers  
International Ballroom A

Chair: Carolyn Turner, McGill University

3:15 – 3:45  Questioning an early start – the transition from primary to secondary foreign language learning  
Alan Davies & Kathryn Hill; Jenny Oldfield & Nadine Watson  
University of Melbourne & Presbyterian Ladies’ College

3:45 – 4:15  The relation between reliability and validity: An issue of test fairness  
John de Jong & Fellyanka Stoyanova, CITO, Arnhem & University of Sofia

4:15 – 4:45  Constructing and validating parallel forms of performance-based writing prompts in academic settings  
Hui-Chun (Angie) Liu, University of Illinois, Urbana-Champaign
Chair: Sarah Briggs, *University of Michigan*

4:45 – 5:30  **Introductions to Works in Progress**

5:30 – 6:00  **Works in Progress**

1. Amma Kazuo. A multidimensional approach to learner’s grammatical proficiency.


7. Yuji Nakamura. Involving factors of fairness in language testing.


10. Randy Thrasher. Toward understanding the nature of communicative competence: An attempt to make sense of think aloud protocol data from reading test subjects.


7:30 – 9:30  **Reception**

Poolside

Co-hosted by Test of English as a Foreign Language (TOEFL) and University of Cambridge Local Examinations Syndicate (UCLES)
PROGRAM
Saturday, March 8

8:00 – 5:00  REGISTRATION
Atrium Hallway outside International Ballrooms A & B

10:00 – 6:00  Book Exhibits
International Ballroom B


8:30 – 9:30  Session 6 – Invited Speaker
International Ballroom A

Introduction of Invited Speaker
Mary Spaan, University of Michigan

Reading research, the development of reading abilities, and reading assessment
William Grabe, Associate Professor, English Department, Northern Arizona University, Flagstaff

9:30 – 9:45  Break

9:45 – 12:00  Session 7 – Colloquium 1
International Ballroom A

Computers and language testing: Evaluating access and equity

Chair and moderator: Simeon Slovacék, California State University, Los Angeles

Presenters
Carolyn Taylor, Educational Testing Service (Organizer)
Irwin Kirsch, Educational Testing Service
Joan Jamieson, Northern Arizona University
Dan Eignor, Educational Testing Service

Discussants
Charles Alderson, Lancaster University
William Grabe, Northern Arizona University

12:00 – 1:00  Lunch Break
1:00 – 2:30  **Session 8 – Papers**  
**International Ballroom A**

**Chair**: Charles Stansfield, *Second Language Testing, Inc.*

1:00 – 1:30  Is it fair to assess both NS and NNS on school ‘foreign’ language examinations?  
Catherine Elder, *University of Melbourne*

1:30 – 2:00  Investigating bias over time  
H. Gary Cook, *Michigan State University*

2:00 – 2:30  The relationship between interviewer style and OPI ratings at three levels of proficiency  
Daniel Reed & Gene Halleck, *Indiana University & Oklahoma State University*

2:30 – 2:45  **Break**

2:45 – 4:45  **Session 9 – Papers**  
**International Ballroom A**

**Chair**: Carol A. Chapelle, *Iowa State University*

2:45 – 3:15  Raters’ understanding of oral scales as abstracted concept and as instruments of decision making: A phenomenographic study  
Constant Leung & Alex Teasdale, *Thames Valley University*

3:15 – 3:45  Methods of assessing and displaying salient features of coherence in ESL writing  
Swee-Heng Chan, *University of Agriculture, Western Malaysia*

3:45 – 4:15  Computer-based oral proficiency assessment: Field results  
Jared Bernstein, *California Metrics*

4:30 – 5:00  **Meet the Language Testing Editors**  
J. Charles Alderson & Lyle Bachman

5:00 – 6:30  **ILTA and LTRC Joint Business Meeting**

6:45  Shuttle Bus leaving from hotel entrance going to AAAL at Holiday Inn International Drive

10:00  Shuttle Bus returns to Holiday Inn Select
PROGRAM  
Sunday, March 9

8:00 – 4:00  REGISTRATION  
Atrium Hallway outside International Ballrooms A & B

8:30 – 9:30  Session 10 – Invited Speaker  
International Ballroom A

  Introduction of Invited Speaker  
  Sarah Briggs, University of Michigan

  English triumphant, ESL leadership and issues of fairness  
  John Swales, Director, English Language Institute, University of Michigan, Ann Arbor

9:30 – 9:45  Break

9:45 – 12:00  Session 11 – Colloquium 2  
International Ballroom A

  Examining test taker characteristics and second language test performance using a multi-group structural equation modeling approach

  Chair and moderator: Frances Butler, UCLA

  Presenters  
  James E. Purpura, Teachers College, Columbia University (Organizer)  
  Antony John Kunnan, California State University, Los Angeles  
  April Ginther, Educational Testing Service  
  Goeff Brindley, Macquarie University  
  Steve Ross, School of Policy Studies, Sanda

12:00 – 1:00  Lunch Break
1:00 – 3:00  Session 12 – Papers
International Ballroom A

Chair: Eduardo Casciiar, City University of New York

1:00 – 1:30  The validity of the paired interview format in oral performance testing
Noriko Iwashita, University of Melbourne

1:30 – 2:00  Balancing fairness and authenticity in performance test: An alternative rating
approach for ITAs
Carol Moder & Gene Halleck, Oklahoma State University

2:00 – 2:30  Assessing the communication skills of veterinary students: Whose criteria?
Dan Douglas & Ron Meyers, Iowa State University

2:30 – 3:00  Non-native varieties and issues of fairness in testing English as a world language
Peter Lowenberg, San Jose State University

3:00 – 3:15  Break
Sunday, March 9

3:15 – 5:15  Session 13 – Posters
           International Ballroom B

Chair: Barbara Dobson, University of Michigan

3:15 – 4:00  Poster Previews

4:00 – 5:15  Poster Presentations

1. Gary Buck, Kumi Tatsuoka & C. Tatsuoka. The effects of coding reliability and the
validity of theoretical assumptions on the Rule Space procedure

2. Vivien Berry. Gender and personality as factors of interlocutor variability in oral
performance tests.

3. Alan Davies & Catherine Elder. Language distance as a factor in ESL test performance.

4. Rudiger Grotjahn & Christine Klein-Braley. Avoiding bias in C-test construction
and C-test.

5. Gene Halleck. Creating an ESL VOCI: From script-writing to field testing.

6. Fellyanka Kaftandijieva & Sauli Takala. Men are from Mars, Women are from
Venus: The case of an English vocabulary test

7. Sungwoo Kang. Effects of native language and learning experience on the
University of Illinois ESL placement test.

8. Gholam Reza Kiani. Extroversion and pedagogical setting as sources of variation in
performance in English proficiency tests.

9. Rana Mathew. Washback of tests on teaching: Is it fair to students?

10. Beryl Meiron, Mary-Erin Cook, Yutaka Kawamoto, Laurie Schick & Carmen
Velasco-Martin. Ratings, Raters, and Test Performance: An Exploratory Study

11. Barry O’Sullivan. The interlocutor as a variable: Affective factors in language
testing.

12. Kyle Perkins. The effect of proficiency on item difficulty, reliability,
reproducibility, and dimensionality in a second language reading comprehension
test.

13. Charles Stansfield & Marijke Cascallar. A listening summary translation exam for
southern Min.

14. Constance Tsagari. Testing reading comprehension: How fair is our choice of
method?
Sunday, March 9

5:15 – 6:00  Session 14 – Summary & Closing
International Ballroom A

Summary of Conference Themes
Chair and Moderator: Mary Spaan, University of Michigan

Panelists
Caroline Clapham, Lancaster University
Brian Lynch, University of Melbourne
Randy Thrasher, International Christian University

Closing remarks
Antony John Kunnan, California State University, Los Angeles
Mary Spaan, University of Michigan

7:00  Banquet and
TOEFL and ILTA Awards
International Ballrooms A, B & C

The 5th Annual TOEFL Outstanding Dissertation Research Award in Second or
Foreign Language Testing for 1997

Presented to Dorry Kenyon, Center for Applied Linguistics, Washington DC

Previous winners:
Gary Buck (1993)
Antony John Kunnan (1994)
Craig Deville (1995)
Caroline Clapham and Sarah Cushing Wiegle (1996)

The 3rd Annual ILTA Best Article Award for 1994

winner to be announced

Previous winners:
Bonny Norton Peirce (1992)
J. Charles Alderson and Dianne Wall (1993)

The 2nd Annual Best Graduate Student Presentation for LTRC ’97

winner to be announced

Previous winners:
Vivien Berry (1996, tie)
Tom Lumley and Annie Brown (1996, tie)
ABSTRACTS

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

PLENARY

A “Postmodern” View of the Problems of Assessment or Why Do I Get Such a Headache Thinking About Test Design?

Henry Braun, Vice President, Research Management, Educational Testing Service, Princeton, NJ

Building good tests has never been a simple matter, but to those who take it seriously the task seems to get harder rather than easier with each passing year. How does one explain the phenomenon? I will look at test construction as a problem of optimal design under constraints and ask how changing conceptions of validity and advances in technology impact the nature of the problem as well as the range of possible solutions. This will lead to some predictions about the long range trajectory of test design generally, with some consideration to the special case of language testing.

Henry Braun is vice-president for Research Management at Educational Testing Service. He holds a doctoral degree in mathematical statistics from Stanford University. Before coming to ETS in 1979, Braun taught and did research at Princeton University and the Sloan-Kettering Institute for Cancer Research. Braun was the co-principal investigator for the ACE/AACRAO Study of College Freshman Athletic Eligibility. In 1986, Braun received the Palmer O. Johnson Award of the American Educational Research Association and in 1991 was elected a Fellow of the American Statistical Association. Braun has served as associate editor of the Journal of the American Statistical Association and chaired the Management Committee of the Journal of Educational Statistics. He has also taught at the Woodrow Wilson School of Princeton University, Department of Statistics, Rutgers University, and served as a statistical advisor to the Office of Population Research at Princeton University; and the Centers for Disease Control in Atlanta.
INVITED SPEAKERS

Reading Research, the Development of Reading Abilities, and Reading Assessment
William Grabe, Associate Professor, English Department, Northern Arizona University, Flagstaff, AZ

An important concern for assessing reading skills is establishing a reasonable interpretation of the construct of reading abilities. Applied linguists, however, have not always indicated a thorough grounding in the more recent research on reading, and this limitation can seriously impact perspectives on reading assessment as well as the practical development of measures of reading ability. This talk will provide an overview of recent reading research from both L1 and L2 perspectives, describe briefly the influence of this research on reading instruction, and highlight issues for reading assessment which could someday change the ways that reading abilities are tested.

Bill Grabe is Associate Professor in the English Department at Northern Arizona University. He received his PhD from USC in 1984 and went to Northern Arizona in 1984. He has taught courses and lectured in various countries: China, Japan, Morocco, France, Czechoslovakia, Brazil, Argentina, Israel, and Mexico, and he was a senior Fulbright Lecturer in China (1980–1981) and Brazil (1990). He has research interests in reading, writing, literacy, and discourse analysis, and his most recent book, with Robert B. Kaplan is Theory and practice of writing: An applied linguistics perspective (Longman 1996). He has also served on consulting committees for TOEFL for a number of years. He is editor-in-chief of the Annual Review of Applied Linguistics (Cambridge).

English Triumphant, ESL Leadership and Issues of Fairness
John Swales, Director, English Language Institute, University of Michigan, Ann Arbor, MI

As the century nears its close, English is consolidating its position as the language of international communication in practically all spheres. While some view this advance in triumphalist terms, others view it as insidious linguistic imperialism, as the commodification of a cultural product, or as a threat to linguistic diversity. I suggest that today’s ESL leadership can no longer avoid this debate. I argue, in consequence, that a little of the vast resources (of all kinds) which underpin ESL teaching might now be redeployed to support the teaching and maintenance of “lesser” languages. The same argument holds, I conclude, for the well-supported global activity of ESL Testing.

John Swales is Professor of Linguistics and Director of the English Language Institute at the University of Michigan. Previous posts include Reader in ESP at the University of Aston in Birmingham (UK) and Director of ELSU, University of Khartoum, Sudan. His interests lie in discourse analysis, genre theory and advanced EAP materials. He was co-editor of English for Specific Purposes from 1986 to 1994. Recent books include Genre Analysis (1990), Academic Writing for Graduate Students (with Christine Feak) (1994), and Other Floors, Other Voices (in press).
PRE-COLLOQUIUM WORKSHOPS

Workshop I: An Approach to the Design and Development of Language Test Tasks

Lyle F. Bachman, University of California, Los Angeles
Adrian S. Palmer, University of Utah

Session I. In order to use the scores from a language test to make inferences about individuals' language ability, and possibly to make various types of decisions, we must be able to demonstrate how performance on that language test is related to language use in specific situations other than the language test itself. In order to demonstrate this relationship, we need a conceptual framework that enables us to treat performance on a language test as a particular instance of language use. That is, we need a framework that enables us to use the same characteristics to describe what we believe are the critical features of both language test performance and non-test language use.

When we design a language test we need to consider two sets of characteristics. We need to consider characteristics of individuals, particularly their language ability, in order to demonstrate the extent to which these characteristics are involved in language use tasks and test tasks. We need to consider characteristics of the language use situation and tasks in order to demonstrate the ways in which our test tasks correspond to language use tasks. These correspondences pertain to three essential qualities of assessments—construct validity, authenticity and interactivity—that contribute to the overall usefulness of the test for its intended purposes.

In this workshop we will focus on the second correspondence, that between characteristics of language use tasks and test tasks. We will discuss the notions of language use task and target language use (TLU) domain. We will then describe a framework of task characteristics that provides the link between the domain of test tasks and the domain of TLU tasks, and that permits us to select or design test tasks that correspond in specific ways to language use tasks.

Session II. In this session we will provide the participants with hands-on experience with procedures for designing and developing test tasks with emphasis on making language assessments useful for their intended purposes. These procedures include identifying the relevant TLU domain ('real-life' domain, language instructional domain), selecting TLU task types for consideration as test tasks, describing TLU task types, and developing test task specifications. This session will consist of whole group discussions of important concepts and procedures and small group or pair activities related to preparing a design statement and task specifications for one or more test tasks for a particular testing situation. Participants are encouraged to bring specific test development projects from their own educational settings, and these will provide the basis for the workshop.

Lyle Bachman is Professor of TESL and Applied Linguistics at the University of California, Los Angeles, where he can be found teaching courses in language testing, research methods, and second language acquisition. His extensive overseas experience includes raising chickens in the Philippines, singing in operettas in Hawaii and Thailand, dodging tear gas canisters in Iran and negotiating the shifting winds and currents of Tolo Harbour and of the Chinese University of Hong Kong.

Adrian (Buzz) Palmer is an Associate Professor of English at the University of Utah, where he directs the TESOL training program. His main professional areas of interest include language testing, teacher training and professional development, and language teaching methodology. He received his PhD in Linguistics from the University of Michigan. Buzz was one of the original founders of LTRC and has been involved in many of the annual colloquia, both as a participant and organizer. He has always believed that the colloquia should include entertainment along with academics, and opportunities for social interaction along with educational exchanges.
Workshop II: Statistical Data Handling: A Principled Process

Fred Davidson, University of Illinois, Urbana-Champaign

Based on Principles of Statistical Data Handling (Sage, 1996, by Fred Davidson), this workshop explores statistical “data handling”: input, manipulation, debugging, preparation (for analysis), and archival. This is the “getting ready” part of statistical analysis; it is the tasks we must follow to go from a stack of hardcopy on our desk (e.g. survey forms) to a computer dataset that is trustworthy. ‘Data handling’ can also be actual data analysis—e.g. running a particular statistical procedure to answer a particular research question—but only in terms of preparing the data for that analysis. The workshop also touches on tips and tricks of managing datasets, particularly those which are analyzed by routine programs on a regular basis (e.g. a periodic educational admissions analysis).

The workshop is organized around a series of vignettes. These are short stories about people managing data and encountering and solving problems—or not solving them, as the case may be. The issues involved in each vignette distill to certain principles of data handling. Following are a few sample principles (taken from Davidson, 1996):

- You cannot analyze what you do not measure.
- Always save a computer file copy of the original, unaltered data.
- Use the computer to check for impossible and implausible data.
- Take control of the structure and flow of your data. (This is arguably the most important principle we will explore.)

The principles should apply to many different contexts, and so some of them will apply to several of the vignettes which we will discuss. A handout will be provided containing the vignettes and a list of data handling principles from Davidson (1996). We will also have time to discuss data handling vignettes of your own—a sort of data-story-sharing round table—and we will explore the data handling principles which might be at operation in some of your settings.

Some of the vignettes (and the principles which they imply) will involve actual computer operations. These will be illustrated on the handout using the command languages of two leading software packages: SPSS and SAS. However, this workshop is not a SAS or SPSS training session. The key goal of this workshop is to detect principles at operation in the data handling process, and like much of our work, that process is largely human and not mechanical.

Fred Davidson is an Associate Professor of English as an International Language at the University of Illinois at Urbana-Champaign. He teaches courses in language testing, second/foreign language acquisition, reading and writing in a second language, and statistical data analysis. He has published in various journals, conference proceedings, and collections, and he is the author of Principles of Statistical Data Handling (Sage, Inc., 1996). His scholarly interests include second/foreign language assessment, writing in a second/foreign language, data analysis and management in applied linguistics, and the history of education with particular reference to language teaching.
Developing computerized tests for international language assessments leads to questions of access and equity if one considers that individuals with little or no prior computer experience may be taking the same test as students who are "computer literate." ETS and the TOEFL program have announced plans to introduce a computer-based TOEFL (TOEFL CBT) in 1998. Given the concern that the measurement of English language proficiency may be confounded with computer literacy, ETS and TOEFL have funded the development of two research efforts to examine, internationally, (1) EFL students' access to and experience with computers and (2) their performance on a set of computerized language tasks following a tutorial session designed especially for this population.

During the first section of the colloquium, three research reports will be presented. The first paper will report on an international survey of approximately 100,000 TOEFL test takers regarding their access to and experience with computers. A short scannable questionnaire (23 items) was developed and administered, and statistical analyses were performed with approximately 90,000 usable questionnaires. Based on the results of a factor analysis, 11 items were identified which loaded on a common factor that could be viewed as a computer familiarity factor. A score on these 11 items was constructed, and individuals were classified into one of three categories — low, middle, and high — based on their responses. Computer familiarity was also examined with respect to the following subgroups: gender, age, reason for taking the test, test center location, native country, native language, and TOEFL paper-and-pencil scores.

The second paper will report on a study which examined the relationship between computer familiarity and performance on computerized language tasks. For this study a set of computer-based materials was developed and administered to 1,200 EFL students in 11 counties. These materials included a set of specially designed CBT tutorials, 60 CBT test items, and a short questionnaire. Based on candidate responses to the questionnaire in the first study, approximately 1,200 TOEFL candidates that had been classified as either high or low computer familiar were administered the CBT tutorial, test items, and questionnaire. Results included a comparison of performance on the CBT items for computer familiar and unfamiliar groups after accounting for English proficiency. The final paper reports on the effectiveness of the tutorial designed especially for EFL students with little or no previous computer experience. The tutorial was examined in terms of time spent and level of performance on a set of structured exercises for both the total group and selected examinee subgroups.

During the second part of the colloquium, two language teaching and testing experts will provide critical feedback about the research reports and participate in a discussion on issues of access and equity in moving toward large-scale computer-based testing and implications for language teaching and testing.
COLLOQUIUM 2

Examining test taker characteristics and second language test performance using a multi-group structural equation modeling approach

What can multiple-group structural modeling offer to language testing research? Antony John Kunnan, California State University, Los Angeles

Population heterogeneity and thus population generalizability are well known concerns among language testing researchers and test developers of large scale standardized EFL tests. Unfortunately, these concerns have not resulted in researchers pursuing a line of research to isolate the most complex problems that surround population heterogeneity. Multiple-group structural modeling (MG-SM) is a method that can be used by researchers to study the differences in factor structure among different populations as well as structural models that model test taker characteristics and test performance together. Such an attempt could begin a discussion of the complex issues that surround population heterogeneity and generalizability, and thus of fairness in tests and test development and score interpretation.

The studies that follow investigate the factor structure of tests by using the multiple-group structural equation modeling approach. Several factor structures are examined in each of the contexts and interpretations regarding the models are presented. Implications for test development and test score interpretations based on the findings are discussed.

An investigation into the effects of strategies on SL test performance with high and low ability test takers: A structural equation modeling approach. James E. Purpura, Teachers College, Columbia University (Organizer)

Implicit in the research on learning strategies is the assumption that strategy use exerts a causal effect on the performance of high and low ability second language (SL) learners. For this reason, considerable research has been devoted to understanding how successful and unsuccessful learners utilize strategies to learn languages. One set of studies on “good” language learner strategies utilized observation, interviews and other verbal report protocols to provide an account of the type, variety and frequency of strategy use with low and high-ability learners and showed that strategy use did, in fact, relate to differential performance (Rubin, 1975, 1981; Naiman et al., 1978; Cohen & Aphel, 1981; Gillette, 1987; Chamot et al., 1988; Anderson, 1989). Another set of studies sought to find a statistical, as well as a substantive, relationship between strategy use and SL performance, but the results provided only partial support for the above claims (Bialystok, 1981, 1983; Politzer & McGroarty, 1985; Mangubhai, 1991).

Purpura (1996) investigated these differing claims by examining the construct validity of the putative effects of strategy use and SL test performance by using structural equation modeling as a primary analytical tool. He found that in a single-sample analysis, certain clusters of strategies showed no effect on performance, given the measures used, while others yielded significant, positive or significant negative effects. Additionally, some strategies produced a direct impact on SL performance, while other effects were indirect.

The overall purpose of this study is to examine the effects of strategy use and SL test performance with students of differing ability levels. More specifically, this study establishes separate baseline models of strategy use and performance for each ability-level group and then tests the equivalence of these models across ability level.
In this study, 1,363 students were given questionnaires to measure their metacognitive and cognitive strategy use. They were also given a University of Cambridge FCE Anchor Test aimed at measuring their language ability. The test takers were then divided into high and low ability groups and the components of both the measurement model and the structural model were tested by means of equality constraints for invariance across groups.

In the first part of the paper, I will summarize Purpura’s (1996) baseline single-sample model of strategy use and performance, commenting briefly on both the measurement and structural models. I will then describe the separate baseline models for each ability-level group. Finally, I will discuss the results of the multi-group analysis in which models of strategy use and performance for both high and low-ability groups will be compared.

Content characteristics of language examinations, background characteristics of examinees and examination structure. April Ginther, Educational Testing Service

Current research in cognitive psychology and language acquisition suggests that increasing competence in a language involves much more than the acquisition of new skills, but rather requires reconfiguration of knowledge structures. Furthermore, theorists concerned with the nature of communicative competence have long argued that language acquisition is best understood as a dynamic process rather than as a state (see Hymes, 1972; Stern, 1983; Savignon, 1983; Spolsky, 1989). If we hope to understand performance on language exams and what that performance reveals about the process of second language acquisition, interpretation of patterns of results on language exams must include analyses of the relevant background variables of examinees and address how such variables influence the relationships among the factors involved.

In this presentation, the results of two separate studies will be discussed. Both studies use a nested-hierarchical design with confirmatory factor analysis to examine the structure of language examinations for examinees who differ in language backgrounds and ethnicity. These studies share the same methodological approach, but differ in the content of the exams.

In the first, the structure of the Advanced Placement Spanish Language Examination was examined for five groups. The structure of the exam for (1) a Latin Spanish-speaking sample served as the starting reference. The structure of the exam for (2) a Mexican Spanish-speaking group, (3) a Mexican Spanish/English bilingual group, (4) a White group of Spanish foreign language learners, and (5) a Black group of Spanish foreign language learners Spanish/English bilingual group was compared to that of the reference group.

In the second, the structure of the Test of English as a Foreign Language (TOEFL) was examined for six groups. An analysis of the exam structure for examinees whose native language was (1) Arabic and who were classified as having low proficiency (<475) served as the starting reference. The structure of the exam for (2) a group of examinees whose native language was Arabic but were classified as having high proficiency (>550), (3 & 4) examinees whose native language was Chinese and who were of low and high proficiency, and (5 & 6) examinees whose native language was Spanish and who were of low and high proficiency was compared to that of the starting reference group.

In both studies, when examinee background characteristics were taken into account, differences in examination structure were found. These differences have implications for our conceptualizations of proficiency and should be taken into account when interpreting examination results. In addition, as one might expect, the differences found are clearly influenced by the content characteristics of the measured variables involved. The relationships among content characteristics, background variables, and examination structures will be discussed.
Trait-method comparisons across three language test batteries using exploratory, MTMM and structural equation modelling approaches. Geoff Brindley, Macquarie University, Steven Ross, School of Policy Studies, Sanda

Language testing research has over the past three decades used different methods of analysing evidence for construct validity. Early approaches relied on exploratory factor analysis and multi-trait multi-method analysis. More recently, language testers have used structural equation modelling (SEM) to examine evidence of hypothesised relationships among language tests. The present research uses these three different approaches in a hierarchical manner to investigate the construct validity of a number of language test batteries used in Australia. These are the ASLPR (Australian Second Language Proficiency Rating) interview procedure used for placement and exit assessment in government-funded language programs; the ACCESS (Australian Assessment of Communicative English Skills) test used to assess the proficiency of certain categories of prospective immigrants to Australia; and the STEP (Special Test of English Proficiency) test used to access the proficiency of some applicants for permanent residence.

This paper presents the results a study which used exploratory factor analysis, MTMM and structural modelling to compare evidence for discriminant and convergent validity of four traits (reading, writing, listening and speaking) across the three test approaches described above. Test scores of candidates (n = 110 to 234) were first compared in a pair-wise manner with the use of exploratory factor analysis. Following this, a three-method by three-trait multi-method multi-trait matrix was constructed in order to examine evidence for trait convergence. Finally, a structural equation model was constructed and tested so as to estimate path coefficients indicating trait convergence relative to paths showing test method artifacts. Although results provide some evidence of a writing trait across all three methods, they also suggest that the interview-based assessment approach creates more method than trait variance.
Assessment of language impairment across cultures. Rosemary Baker, *The University of Queensland, Brisbane*

*Background and Purpose.* This presentation examines the extent to which language tasks of the types commonly used for the assessment of language impairment may be considered valid for administration to people of different linguistic and cultural backgrounds. Discussion focuses on language assessment in age-related disease, specially aphasia and dementia, and is set in the Australian context. This issue is of increasing importance to health professionals in this ageing, multilingual society, because of its implications for correct diagnosis and management.

The work reported here sought to provide essential baseline data to inform the design of appropriate assessment procedures in this context, by investigating the performance of ‘normal’ elderly people, i.e. people with no known neurological impairment, from a range of ethnolinguistic backgrounds, on a variety of widely-used tasks.

*Method.* Over a three-year period, data have been gathered from a total of more than 250 Australian residents aged 60 and over, representing speakers of Dutch, German, Vietnamese, Latvian, Italian, Chinese (Cantonese and Mandarin), Polish and Greek.

The tasks administered were largely drawn from (or adapted from) published aphasia tests and from a battery designed to assess language disorders associated with dementia. They included word retrieval tasks, repetition and recall, and reading and listening comprehension tasks. Where possible, tasks were administered both in participants’ first languages and in English, in separate sessions. Details of participants’ personal and language background were also gathered, and, in view of the potential effect of age-related hearing loss on performance on language tasks, particularly those where little or no contextual information is available, hearing was tested using pure-tone audiometry.

*Results and Implications.* This testing has yielded rich data sets which allow identification of (i) tasks which appear to be problematic, whatever the language of administration, and (ii) tasks which one could expect non-neurologically-impaired members of these populations to perform without difficulty, given adequate proficiency in the language in question. Clearly, only tasks of the latter type could be considered potentially suitable for use in assessing language impairment in these groups.

The results also indicate the variation to be found between and within the different language background groups in terms of their overall performance on these tasks, and the nature of the responses obtained on specific tasks such as the generation of words by semantic category. In view of the use of such tasks in the diagnostic process, information on what constitutes a normal pattern of responses for individuals of different backgrounds is of immediate practical concern.

Overall, the discussion will point to some important group and individual characteristics that need to be taken into account in selecting or adapting procedures for the assessment of language impairment across cultures.
Computer-based oral proficiency assessment: Field test results. Jared Bernstein, California
Metrics

The presentation will discuss the rationale and the procedure used to validate a computer-based oral proficiency assessment that is being field-tested with international students at several universities. Speech is elicited, recorded and scored by computer, then subscale scores are combined to optimally predict oral proficiency scores assigned by human listeners. The project started in April 1996 and will continue through January 1997. The presentation will explain the elicitation procedures, play examples of spoken responses, and review methods used by expert human judges to score the students' proficiency.

Significant advances in speech recognition technology make it possible to improve and to automate some aspects of the speaking/listening sections of language tests. The new test of oral proficiency can be administered by computer over the telephone. It presents the examinee with a set of interactive tasks that require English oral comprehension and production skills at conversational speeds. A unique exam sheet is given to each field-test examinee. Each sheet has instructions on how to call and the aspects of performance that are scored. When an examinee calls, the system answers and reviews the procedure. The system then presents material and elicits spoken responses from the examinee.

In order to use this system for screening or placement of university ESL/EFL students, one gathers an appropriate sample of students and establishes their relative oral proficiency through expert human judgments that can be accepted as a criterion measure. The speech samples have been processed and scored in parallel, but independently, by a computer-based system that can rate the word accuracy, pronunciation, prosody and fluency of spoken responses. The correlation of the computer-based scores with the concurrent expert human criterion scores serve as validating evidence that the testing method can yield useful estimates of oral proficiency as it has been exercised and sampled during this over-the-telephone examination. The field test has produced samples of speech data that can be used to validate and calibrate the experimental testing system.

The field test has obtained interactive speech samples (as well as recitations and short answers) from a population of speakers located at several universities. This should provide a wider range of oral proficiency than is available at any one school. The resulting test may be particularly appropriate for screening or placement decisions when large numbers of students or candidates are to be tested. The long term goal is to provide a fair and efficient test that can serve as an alternative to (or a screening instrument for) a more traditional language proficiency interview. [This work is partially supported by a grant from the National Science Foundation.]
The Effect of First Language Background on the Cognitive and Linguistic Attributes Underlying Performance on the Test of English for International Communication.
Gary Buck, Kumi Tatsuoka, C. Tatsuoka, and Irene Kostin, Educational Testing Service

This paper addresses the question of how fairly a set of cognitive and linguistic attributes which have been derived from test-takers of one L1 background can be used to assess the performance of test-takers from a different L1 background.

Using the rule-space methodology, a set of cognitive and linguistic attributes were derived on a sample of 2000 Japanese test-takers, for each of the seven different parts of the Test of English for International Communication (four sets of listening items, and three sets of reading items). These seven sets of attributes (i.e. a separate set of attributes for each part of the test) were then applied to a sample of 2000 Korean test-takers who took the same form of the TOEIC. Performance of the two groups on each of the seven attribute sets will be compared in a number of ways: (i) the classification rates and the total score variances explained by the attributes will be compared; (ii) non-parametric regression functions using the spline smoothing method will be estimated on the TOEIC total scale scores, and the conditional means at various scale score levels will be compared for each attribute; (iii) the distributions of attribute mastery probabilities for each attribute for both the Japanese and Korean samples will be estimated and compared. These comparisons will be made for each of the seven parts of the test.

The results will be presented and implications discussed in terms of (i) how L1 background influences the attributes underlying performance on listening and reading tests, and (ii) how fairly attributes derived from test-takers of one background can be used to assess performance of test-takers from a different background.
Methods for assessing and displaying salient features of coherence in ESL Writing. Swee-Heng Chan, University of Agriculture, Western Malaysia

In the Malaysian ESL classroom, the traditional evaluation procedure used for the assessment of writing has been to break writing skills into five basic components. They are content, organization, language, mechanics, and grammar. Teachers would write these five words on the compositions and award marks for each of the criterion to arrive at a composite score. Explanations about the scoring system are seldom given. Such a system is neither meaningful nor helpful for writing improvement, especially if it is seen in the light of progress assessment.

In the search for a more meaningful system in the evaluation of ESL writing, a study was carried out whereby the coherence construct was focused upon. A Coherence Scale together with a display system of coherence problems were devised to allow students to realize the aspects of coherence that needed to be looked into in order to lead to writing improvement. These two systems, the Coherence Scale and the display system, which essentially involves the tracing of the theme-rheme progression and coherence breaks, are shown in context of scoring 136 university ESL student compositions.

The Coherence Scale defines six parameters for the assessment of coherence and the scores obtained were correlated with scores obtained from the ESL Composition Profile. The scores for the same compositions were obtained from two panels of raters. The correlations show a divergence in emphasis in the use of the two scales which supports the contention that coherence could be more sensitively evaluated on its own and merits a scale for its assessment.

However, a scale alone is insufficient if an assessment system is meant to be meaningful in terms of identifying the problems more concretely. As such the salient features of coherence needs to be further pin-pointed to serve as a framework of reference for improvement. The coherence display is both visual and descriptive. The tracing of the theme-rheme progression is demonstrated for sample compositions to show the manifestation of coherence patterns. Coherence breaks are indicated to mark coherence problems encountered and these breaks are explained with reference to the salient coherence features in writing.
Multi-Dimensional Analysis and Oral Proficiency Sampling: Performance Effects of Task and Elicitation Context. Jeff Connor-Linton, Georgetown University, Elana Shohamy, Tel Aviv University

Most frequently-used tests of oral proficiency assume that face-to-face interviews sufficiently sample a testee’s competence, but little empirical evidence exists of the relative homogeneity of NNS oral performances elicited under different conditions or of the relationship between performances elicited via different tasks/prompts. This paper reports a multi-feature/multi-dimensional analysis of register variation (form-function cooccurrence patterns) in NNS oral performances in English varying by elicitation context and task/prompt.

Ten Israeli university EFL students, representing a range of L2 English proficiency, responded to each of three speech act elicitation prompts (telling about oneself, complaining, and requesting) in five elicitation contexts: (1) in a traditional face-to-face interview with a tester, (2) face-to-face with a peer, (3) responding to an audiotaped prompt and (4) a videotaped prompt, and (5) by telephone. Frequencies of the 67 lexical and syntactic features selected by Biber (1988) were measured in each of the 150 samples, and the ‘factor scores’ of different prompts, elicitation contexts, and testee proficiency levels on Biber’s dimensions of linguistic variation are compared and correlated with ratings of the samples.

Analyses of variance show significant effects of task/prompt type, but not elicitation context, on subjects’ use of language. Implications of the results for elicitation and sampling of oral proficiency, as well as for our understanding of communicative competence, are discussed. Uses of multi-feature, multi-dimensional analysis in test development—and in interlanguage/proficiency level modelling—will be discussed.
Investigating bias over time. H. Gary Cook, Michigan State University

The most common way of identifying fairness (bias) in testing situations is by performing some sort of test or item bias analyses (see Reynolds, 1979 or Scheuneman & Bleistein, 1989; or for a more recent treatment in Language Testing, Ryan & Bachman, 1992). All of these methods are based on single administrations of exams. In fact, many test statistics rely upon the concept of single administration sampling, i.e., one examinee per sample. One noteworthy counter example would be test-retest reliability. There is nothing inherently wrong with this approach, but these approaches do not look at language growth over time. In essence, they take a slice of time and analyze bias based on that slice. It may be the case that test or item bias is occurring over time. Until recently, there has been no adequate procedure for analyzing the performance of language tests over time. This is primarily due to the nested nature of the data. Recently, in the field of educational research, a new procedure has been introduced which can adequately account for nested data. The procedure is called hierarchical linear modeling (HLM). Using HLM, Bryk & Raudenbush (1987, 1992) introduced a method for assessing change in educational environments.

Two tests, a reading and a listening comprehension placement test, will be analyzed. In a previous HLM analysis using linearly equated reading and listening comprehension tests, Cook (unpublished) found that different language groups did have significantly different growth curves. Drawing on this analysis and using Rasch modeling, this study will examine how different language and age groups perform on these measures. Several questions can be investigated using this procedure, e.g., To what degree do developmental differences contribute to growth curves?; Do language groups exhibit similar growth curves across proficiency levels?; Are growth curves similar across skills?; Do language groups perform similarly across skills and proficiency levels?; What are the social consequences of knowing that different language groups will perform differently on exams over time? The goal of this study is to further investigate the nature of these exams, especially as it relates to bias over time. It is hoped that this study will begin a dialog about growth and the testing of growth in the language testing field.
Questioning an early start — the study of foreign languages at primary school. Alan Davies and Kathryn Hill, University of Melbourne; Jenny Oldfield & Nadine Watson, Presbyterian Ladies' College

In line with the recommendations of a recent Australian Government report (Rudd 1994), the Victorian Government has set targets for all students to study a foreign language from the beginning of primary school to Year 10 by the year 2000 (LOTE Strategy Plan 1993). Such a policy assumes not only the value of language learning per se but that students gain an advantage from an early start, an assumption which seems to stem partly from the idea that the longer a language is studied the better and partly from the view that young children pick up new languages more easily than older learners. Clearly there is an expectation that children continuing a language from primary school have an advantage over those with no previous instruction in the language.

This paper reports on a study conducted at a secondary school in Melbourne. The school currently separates children entering Year 7 French into two groups on the basis of whether or not they have studied the language in primary school. The present study was set up to investigate whether the current placement procedures are justified and, if so, how they could be improved.

A total of 200 students, randomly selected from Year 7, 8 and 9 French classes, participated in the study. Half of the students at each year level studied French at primary school (Group A) and half had no previous tuition in the language (Group B). There were approximately 30 students in each of the six groups. The students were tested in reading, writing, listening and speaking using the Australian Language Certificates (Beginner French), an achievement/proficiency test designed for students from years 7 to 10.

This paper will address the question of whether the study of French in primary school gives students an appreciable advantage over other students, the nature of such an advantage and whether it is maintained over time. The study also raises important issues relating to the transition from primary to secondary foreign language learning, including how best to measure the benefits of an early start. Discussion focuses firstly, on whether what students learn in primary school language programs is adequately measured by the assessment instruments currently in use and secondly, whether it is fair to assess students who are continuing the language from primary school and students commencing study of the language in secondary school in the same way.
Assessing the communication skills of veterinary students: Whose criteria? Dan Douglas, Iowa State University, Ron Myers, Iowa State University

Jacoby and McNamara (1996) have proposed the notion of “indigenous assessment” to describe the criteria employed by subject-specialists when evaluating the communicative performances of their peers, as for example in rehearsals of conference presentations or reviewing manuscripts for publication. They suggest that analyses of such assessment practices may prove productive for establishing criteria for the assessment of specific purpose communicative language ability.

This study is an investigation of the assessment criteria developed and used by professors in a large college of veterinary medicine in the evaluation of veterinary students’ communication skills in interviewing a simulated client (animal owner) to obtain an appropriate case history before performing a physical examination of the patient. Each student’s interview performance was evaluated by a trained subject specialist assessor using a standard rating form. The focus of the present study is on the relationship between the assessment criteria on the rating form and the “indigenous” criteria that guided the judgments made by the assessors, as revealed in retrospective commentary. Data consist of videotapes of three interviews, representing three performance levels, the evaluation forms filled out by the assessors, and retrospective commentary on the videotapes by the students, the simulated client, the assessors, and other subject specialist informants. A discourse analysis was conducted on the retrospective commentary to obtain information about the "indigenous" assessment criteria in the minds of the various informants that may overlap with or vary from those on the standard rating form.

The presenters will describe the data and analysis, and discuss the notion of indigenous assessment in terms of fairness to the test takers and the veterinary profession, and its implications for specific purpose language testing. Assessment of language impairment across cultures
Is it fair to assess both NS and NNS on school ‘foreign’ language examinations?  Catherine Elder, University of Melbourne

The argument for teaching minority languages in Australian secondary schools is a social justice one, namely: to allow those who use these languages in the home to gain formal recognition for their existing competence. However, it is often the case that these languages are studied not only by speakers of minority languages (NS) but also by foreign language learners (NNS) with no access to the target language outside the classroom. Moreover it remains common practice for the two types of learner to be assessed in common on end-of-school ‘foreign’ language (FL) examinations.

It is generally assumed that NS, because of greater opportunities to use the target language, will perform better than NNS and that the end-of-school FL examinations are biased in their favour. This assumption underpins a recent policy initiative by the university selection authority to compensate NNS for what is perceived to be unfair competition from the NS.

To further investigate this issue, this study compares the performance of NS and NNS on the 1995 Year 12 Victorian Certificate of Education (VCE) examinations in three languages: Chinese (n = 1100), Italian (n = 794) and Modern Greek (n = 930). Two questions are posed: 1) Do NS perform better than NNS on the relevant ‘foreign’ language examination? 2) Can the university selection authority’s claim of bias in favour of NS be substantiated?

Candidates for each examination are assigned to one of four categories (on a continuum from native to non native speaker) on the basis of responses to a detailed language background questionnaire eliciting information about levels of exposure to the target language outside the classroom. Kruskal Wallis and ANCOVA procedures are applied to determine whether 1) there are significant differences between the scores of the NS and NNS groups on both internally and externally-assessed examination components for each of the three languages 2) there are differences in regression coefficients and intercept value across groups when the scores obtained on the relevant foreign language examination are regressed against those for VCE English/EFL (deemed by the university selection authority to be an equivalent measure of academic ability).

Although findings are somewhat different across languages, in general it appears that home use of the target language is not a sufficient condition for superior performance on school foreign language examinations. Nevertheless, the results of the bias analysis (using the English/ESL examination as benchmark) suggest that on each of the relevant foreign language examinations the scores of the NS candidates have been overestimated. It is argued, however, that the choice of criterion against which test bias is estimated and the subsequent interpretation of test scores is itself subject to bias. The current institutional practice of applying a correction factor to NNS test scores is therefore inappropriate. It is advocated that in this, as in many other testing situations, solutions other than (or in addition to) bias analysis must be considered to ensure fairness to all candidates.
The validity of the paired interview format in oral performance testing. Noriko Iwashita, 
*University of Melbourne*

Studies of oral test discourse have been mainly concerned with interaction between native 
speaker and non-native speaker (e.g. Lazaraton 1992; Ross 1992), but there is little research on 
interaction between non-native speaking candidates in a paired oral interaction test. Not much is yet 
known about variations in the quality and quantity of language which is produced by non-native 
candidates interacting with other non-native candidates or about the impact of this variation on test 
scores and its implications for test fairness.

In relation to the paired oral interaction test, this study addresses the following research 
questions: 1) Does the candidates’ discourse differ according to the proficiency of the speaking 
partner? 2) Does their score differ in relation to the proficiency of the speaking partner?

The data is drawn from performances in a task based oral interaction test. 20 candidates 
undertook the test twice, once with a partner of the same proficiency level and once with a partner at 
a different proficiency level. Each interview was rated twice for both candidates and analysed using 
FACETS, an IRT based model, to compensate for rater harshness. The tapes were transcribed and an 
analysis of specific discourse features (e.g. intelligibility, accuracy, vocabulary) was carried out. 
Candidates were also asked to complete a questionnaire eliciting their reactions to the two test 
conditions.

The findings of the analysis revealed that while the proficiency of non-native speaking 
interlocutor has some impact on the quality of the discourse, there was little difference in the test 
scores across the two testing formats. However, testtaker feedback suggests that candidates prefer 
NNS-NNS interaction mode to NS-NNS mode as they find it less threatening. The study has 
implications for testing in general and for classroom assessment of foreign language learners in 
particular.
The relation between reliability and validity: An issue of test fairness. John H.A.L. de Jong,

CITO, Fellyanka Stoyanova, University of Sofia

Background and Rationale. The requirements of fair testing imply that scores assigned to
candidates reflect their position with respect to the predefined trait only, and that no other attributes
of candidates, such as nationality, age, or gender impact on their score. The field of language testing
recurrantly endorses the debate on the relative importance of reliability and validity. Inspection of
recent discussions within the electronic discussion network LTEST-L shows that it is still commonly
suggested that there is a necessary trade-off between reliability and validity. This view is partly due
to a misinterpretation of the concepts of reliability and validity, but from the discussions it is also
apparent that the basic principle of abstraction — necessary for measurement — is sometimes badly
understood and measurement dimensions are confused with concepts of reality.

Purpose of the research. This research was undertaken to investigate the relationship
between the dimensionality in different subsets of items from a language test and indicators of the
reliability and validity of these subsets. In particular the effect of these indicators is studied on the
scores for specific groups of subjects taking the test.

Research design and method. Data were gathered from samples from two distinct
populations. The first sample (N > 1000) was taken from a population of adult learners of a second
language and represented a number of ethnic minorities. The second sample (N = 200) was taken
from students in secondary education and contained representatives from the national majority group
and from a number of ethnic minorities. Both samples were given a test to assess listening
comprehension in the majority language at a level assumed to be requisite for second language
learners to participate in studies that are open without further testing to the population of students of
the second sample. In addition the second was given a standardized test of verbal intelligence. Data
were analyzed using measurement models from Item Response Theory (IRT), the one parameter
Rasch model and a model allowing variation among items discrimination. Differential items
functioning (DIF) was investigated for various subgroups of the candidate population.

Results. The study shows that obtaining fit for a two parameter IRT model is insufficient to
retain the assumption of unidimensionality. The discrimination parameter allows to fit larger sets of
items to the measurement model, but absorbs dimensionality in the data thus leading to biased
estimates of ability. DIF analyses suggest relationships between nuisance variables and background
variables such as age, intelligence and ethnicity. Selecting a set of items to fit the one parameter
model leads to test results that can be interpreted as the expression of a unidimensional underlying
trait and that are invariant across subsets of items. The research design allows this underlying trait to
be interpreted as an ability that native speakers master at the target level and which is not related
significantly with intelligence, This is taken as an indicator of the validity of the set of items for the
intended measurement, whereas it is also shown that this set of items has higher relative reliability
than the total test.

Implications. The dimensionality discussion in language testing research may profit from
new insights in the relationship between the robustness of the assumption of unidimensionality and
indices of both validity and reliability. From a test fairness point of view, this study reaffirms that
deviations from unidimensional measurement should be considered with extreme caution.
Raters' understandings of rating scales as abstracted concept and as instruments for decision making: A phenomenographic study. Constant Leung and Alex Teasdale, Thames Valley University

Theoretical background and rationale. In the majority of assessment contexts raters have not been involved in scale development. Recent research has investigated rater variability but the focus has been principally on seeking to identify and, where possible, redress, the effects of variability between raters.

Purposes of the research. The current study, conducted in the context of the multilingual primary classroom investigates the understandings of generalist and ESL trained teachers of a statutory rating scale for oral performance and relates these to subsequent decisions made by the same teachers about children in actual assessments.

Research design and methods. Using a phenomenographic approach (after Marton 1981), teachers will be interviewed about their categories of concept related to the domains covered in the oral rating scale. The teachers will maintain a diary of their assessment practices in arriving at assessment scores for the children. The teachers will observe video sequences of themselves interacting with their classes and will be interviewed to establish which elements of their students' performance relate to the assessment criteria. Subsequently, they will be interviewed again with the transcripts of both previous interviews and invited to account for the consistencies and inconsistencies between the record of their abstract concept of the scales and their actual use of the criteria in their situated assessment practice.

Results. The analysis of transcript data will focus on the identification and comparison of the teachers' espoused theories and their theories-in-action (Arghris and Schon 1978). A fine-grained analysis of the content, processes and context of situation in the video data will also be conducted.

Implications and the results. The results feed into a larger context-specific study of assessment practice, the ultimate aim of which is to investigate the possibility of data-driven scale building in a relatively unstandardised mass assessment context. Issues of fairness in relation to scales for learners for whom English is not a Mother Tongue are a key concern. More generally, there are implications for rated test formats. Where high inter-rater reliability is present in a test, it is often assumed that the mapping and calibration (North 1995) of abilities in the rating scale are those operationalised by the raters. Differences between raters' representations of scale criteria are, however, a threat to construct validity, even where the statistical requirements for inter-rater reliability have been met.
Constructing and Validating Parallel Forms of Performance-based Writing Prompts in Academic Settings. Hui-chun Liu, University of Illinois, Urbana-Champaign

When a writing test is administered more than once, multiple forms that are comparable are often demanded due to practical concerns of test security and fairness (Petersen, Kolen, & Hoover, 1993). However, constructing “parallel” forms of performance-based writing tests is extremely complicated because observed ratings are the unique fusion of measurement instruments (i.e., writing prompts, raters, scoring rubrics and procedures) and examinees’ writing ability. Over the years, researchers have been studying the impact of these instruments on examinees’ performance (Pollitt & Hutchinson, 1987; Henning & Davidson, 1987; Spaan, 1993; Hamp-Lyons & Mathias, 1994) and ways to control the variability caused by these intervening variables (Phillips, 1985; DeMauro, 1992; Lee, 1993; Engelhard, 1994); yet questions such as “How should parallel forms of performance-based writing prompts be constructed?”, “What makes writing prompts parallel to each other?”, “Under what conditions can meaningful score generalization be attained across authentic academic writing tasks?”, and “Do ‘parallel’ prompts maintain their comparability across distinct subgroups?” remain inconclusive.

This study is designed to examine these unsettled issues by empirically constructing and validating parallel forms of a performance-based ESL Placement Test (EPT) in a major U.S. university. Under current context, “parallelism” across writing prompts are defined on three levels — test specification level, decision level and skill profile level. Three authentic academic writing prompts were first developed based on the EPT Video-Reading Academic Essay Specification, of which two were intentionally constructed to be parallel while the third one to be nonparallel. Content experts were asked to evaluate the parallelism among the three developed prompts on the basis of the EPT Video-Reading Academic Essay Specification and the results of their judgment were used as empirical evidence for the content validation of the EPT. 305 international students whose TOEFL score is below the university minimum (i.e., 607) or the departmental minimum (if higher) attended the first two administrations of the EPT in fall, 1996. They were treated as two separate groups such that both groups responded to one anchor prompt plus one different prompt. During the test, examinees first watched a ten-minute video-lecture and then read an article which addresses the same topic but with different information. Their task was to demonstrate that they have the adequate writing ability to incorporate both sources of information in a 1-2 page essay using the general writing format in academic settings.

The essays were then evaluated both holistically and componentially by qualified raters. The many-faceted Rasch-model-based FACETS program (Wright & Linacre, 1996) was employed to calibrate examinees’ writing ability. The variability caused by rater severity, prompt difficulty, aspect difficulty (i.e., organization, content, conventions, vocabulary, and style) was analyzed and controlled to adjust the estimate of the “true” writing ability. The results were used as evidence for evaluating parallelism in the levels of “decision reproducibility” (Lung, Stahl, & Wright, 1994) and skill profiles. The differential interaction between prompts, rater severity and subgroups of examinees in terms of their fields of study and academic status was also investigated. Examinees and raters who behave differently from the expected majority were identified and interviewed to further explore the underlying mechanism which leads to their aberrant behavior patterns. The implications of the findings of this project for the future development of criterion-based test specifications and the construction of parallel performance-based authentic writing prompts in academic settings are also discussed.
Non-native varieties and issues of fairness in testing Standard English as a world language.

Peter Lowenberg, San Jose State University

In the assessment of non-native proficiency in standard English, an implicit, and frequently explicit, assumption has long been that the norms for Standard English around the world are limited to those which are accepted and followed by educated native speakers of English.

This paper challenges that assumption, and thus the fairness of tests based on that assumption, by presenting data from domains of Standard English in the non-native varieties of English, which have developed in many former colonies of Britain and the United States, such as the Philippines, Singapore, India, Sri Lanka, Kenya and Nigeria. In these countries, Standard English is used daily as a second, often official, language in a wide range of intranational domains, including government, the legal system, commerce, the mass media, and as a medium of instruction in education.

Analysis of data from these varieties in domains of Standard English reveals numerous widespread and systematic divergences in these varieties from “native-speaker” varieties of English (e.g. British and American). These divergences occur at all linguistic levels, from morphosyntactic and lexical features that also vary across native-speaker varieties (e.g. count/mass distinctions in nouns, verb + preposition collocations, semantic shifts), to the transfer from other languages of pragmatic, discoursal, and stylistic conventions.

Additional data from attitudinal research conducted in two countries with widely attested non-native varieties, India and Singapore, indicates that approximately 50% of college educated English users believe that certain of these divergent features should be local norms for Standard English usage and the models for English language teaching.

In light of this evidence, acceptable models for Standard English can be seen to vary between native-speaker and non-native norms, depending primarily on the usage and attitudes of educated English speakers in each speech community where English is used.

A major implication of this research for language testing is illustrated by examining selected items from high stakes tests of English as a world language. The keys to these items, though in accord with norms of the native speaker varieties, violate norms for Standard English in one or more non-native varieties which have been described to date. These items could therefore be particularly difficult to answer for users of non-native varieties, thereby putting them at an unfair disadvantage in comparison with other examinees.
Balancing Fairness and Authenticity in Performance Tests: An Alternative Rating Approach for ITAs. Carol Lynn Moder and Gene B. Halleck, Oklahoma State University

Current trends in authentic language testing have led to an increased use of performance tests, especially for assessing those who will use language in a professional context. Although such tests are generally viewed as having greater content validity, their use has raised a number of issues of reliability and fairness (McNamara, 1996). Brown & Lumley (1996) have focused on problems related to the variability of interlocutors in the performance task, while Shohamy (1996) and Tarone & Chalhoub-Deville (1996) have raised questions concerning the participation of stakeholders in the rating of performance tests. Such issues are particularly relevant to the testing of international teaching assistants (ITAs).

The most common test format used in ITA testing is a task which requires candidates to teach a lesson to a group of raters (Hoejke & Linnell, 1994). In most cases the raters include trained faculty raters as well as stakeholders, faculty from the candidates’ prospective department and undergraduate students. While the inclusion of such stakeholders may improve the acceptance of the ratings, the judgments of the stakeholders is often problematic. For example, preliminary investigation into the use of student raters in the assessment of these teaching tests has found that their ratings are often more intolerant than those of faculty raters (Moder & Halleck, 1995). Such ratings may be based more on student biases than on the actual performance of the ITA. The present paper examines the effect of using an alternative assessment procedure which includes student stakeholders, but alters the way in which they participate.

Vidéotaped performances of ten ITAs were rated by two groups of student raters using the same rating sheet. To add to the nature of the authenticity of an ITA performance test, we required one group of student raters not only to sit in a classroom and observe the lesson, but also to take notes, and subsequently to answer questions about the “lecture” thus making the criteria more congruent with those of users in real tasks (McNamara, 1996). The other group followed the normal procedure of rating the performance without having taken any notes. The notes of the first group of students were coded according to the number of main ideas and sub ideas which were accurately represented. Two analyses were made. The first analysis compared the ratings of the two groups of rates. The second analysis compared the accuracy of the notes taken by the first group of raters with their ratings of the ITAs’ performance.

We hypothesized that if raters had to take notes they would be more likely to view the performance as students in an actual class would and that this extra measure of authenticity might enable the raters to more fairly rate such aspects of the performance as organization of the lecture, and comprehensibility of the discourse. The results of the study will be discussed together with their implications for future ITA Test administrations and for performance testing in general.
The relationship between interviewer style and OPI ratings at three levels of proficiency.
Daniel J. Reed, Indiana University, Gene B. Halleck, Oklahoma State University

Reed and Halleck (1996) found a relationship between proficiency rating obtained in an oral interview and the pitch of the interview (as determined by the interviewer). To further investigate this relationship between interviewer behavior and rating, an analysis of test-retest pairs of interviews at three levels of proficiency was undertaken. Each candidate was interviewed by two different ACTFL-certified testers with different styles. One interviewer spent more time in an attempt to firmly establish the floor of the interview, while the other interviewer focused more quickly on probes in an attempt to establish a ceiling.

Transcript analyses were carried out both for cases in which the retest resulted in a different rating for a particular individual, and for cases in which the same rating was assigned. Results suggest that interviewer style is a factor that can affect test outcomes even when style variation occurs within the constraints of the test's regular procedures, and that, furthermore, the nature of the consequences of the style differences depends on the general level of proficiency of the examinee. Limitations of this study's design for addressing these important issues are discussed along with recommendations for further research.
Genre as a determining influence on post-graduate students' grades: using Swales' model to investigate the assessment of student writing. Benny Teasdale and Alex Teasdale, Thames Valley University

Theoretical background and rationale. There are a number of genre-based studies which investigate the nature of academic writing (Crookes 1986; Dudley-Evans T. 1986; Swales 1990). For the majority of students studying at British universities academic writing makes considerable demands and they spend much of the duration of their course developing their abilities in the genre. A common feature of MA courses in ELT in Britain is the dissertation - a work typically of between 12 and 20,000 words. For many students it is the longest and most rigorous piece of academic writing they will have undertaken. It is ostensibly assessed in relation to subject discipline rather than language criteria.

Purposes of the research. The research investigates the degree to which MA ELT dissertations graded "high" (i.e. A or B on a 5 point letter grade system) and those graded "low" (i.e. D or below) conform to the CARS (Create a Research Space) discoursal model proposed by Swales (1990). Since post-graduate assignments are focused at opportunities for candidates to show evidence of content domain knowledge and critical reasoning skills, discoursal competence need not necessarily be associated with successful performance. This study investigates whether there is, in fact, an association between discoursal competence and grade awarded, with the implication that discoursal competence may be as much a contributing factor to grade as the knowledge and skills which the assessment is targeted at.

Research design and methods. Ten 10-12,000 word dissertations (7 "high" and 3 "low") were selected, 4 written by native speakers of English and 6 by non-native speakers. Dissertations which involved data collection and analysis were used since it is the research article which is the focus of Swales' CARS model. The introductions of each dissertation were analysed for conformity with the "move" and "step" structure outlined in Swales' academic writing model.

Results. The results indicate that higher graded dissertations displayed greater conformity to Swales' model than did lower graded dissertations. Not all of the higher graded dissertations were the work of Native Speakers of English.

Implications and the results. A number of interpretations of the findings are possible. One interpretation is that the higher graded individuals have been more successfully inducted into the academic discourse community and have as a consequences arrived at a better understanding of the values and methods of research in the subject discipline. An alternative view is that the requirements of academic discourse favour those who, by educational background, are familiar with related literacy practices. The findings suggest that specific assessment criteria for language and academic discourse would increase the validity of the assessment and have a strong formative impact on students' understanding of the task requirements.
A multidimensional approach to learner's grammatical proficiency. Amma Kazuo, University of Reading

This study attempts to focus on the learner's misconception of grammatical forms detected by multiple-choice items. It helps the learners as well as the teacher to get a more accurate feedback from incorrect responses than an ordinary feedback by dichotomous scoring.

Subjects are Japanese university students in their first and second years. They were asked to choose a grammatically correct sentence or paraphrase of a model sentence in a multiple-choice grammar test. The incorrect options chosen by the same test-takers across items were collected. The similarity scales of these pairs of options were calculated and this similarity matrix was converted to a two- or three-dimensional map by multidimensional scaling. A cluster of options on the map represents a favourite tendency of the test-takers' misconception of a grammatical form. Although clusters are interpreted in the same fashion as in factor analysis, a contrast is found with some clusters between test-takers of different ability levels. It means we can judge the threshold ability level of learners under which they are weak in that cluster category.

One of the advantages of this approach is that it combines individual differential item functions into one set of general tendency. It also clarifies notable differences of misconceptions according to different proficiency levels. One such strategy attributed to the low-proficiency group is to interpret a sentence by the lexical meaning of content words irrespective of the word order — a strategy parallel to that in early L1 acquisition. Meanwhile a high-proficiency group tends more often to use structurally oriented strategies. Thus the difference, not development, in L2 proficiency can be regarded as a difference in abstract structuralisation.

This approach has also a potential applicability to other areas of language tests: (1) 'skill' tests as in reading, listening, and vocabulary, and (2) comparison of performance of subjects of a particular background, e.g. L1, experience of early L2 education, attitude/aptitude, etc.

This research addresses the difficulties of confirming the appropriacy of the language proficiency criteria applied by a British University for entry to its post-graduate courses. It also explores methodologies which are sufficiently sensitive to investigate the relationship between initial language proficiency and eventual academic performance.

It is common practice, in British universities, to require an international student to provide proof of language proficiency. Tonkyn (1995) summarises the many types of test accepted and the criteria established. Not unexpectedly, institutions across Britain vary (in some cases widely) in the levels set for entry to similar courses. Apart from questioning which institution has established the most appropriate and equitable level, this variation raises the issue of how informative language proficiency is vis-à-vis a student’s eventual academic performance.

Previous research into the relationship between students’ initial language proficiency and their final performance on an academic course has been problematic. Difficulties have arisen from a number of factors, including: small n size; truncated samples; the establishing of a criterion for success; and, the interaction of language proficiency with other factors. In fact, little more than a tenuous link between language proficiency and academic performance has been established.

This study arises from an ongoing 15 month pilot study and has tracked the academic progress of 38 post-graduate international students studying at a British university. Data has been collected from both the students and their tutors and data analysis will be complete when the final degree results are issued in November 1996.

The paper will present the results of the pilot study and address the following methodological considerations:

1. How might representative and reliable data be gathered from teaching size?
2. How might a comprehensive list of other influences on performance be established and might it also be possible to establish their relative ‘weights’ in influencing performance?
3. It is unusual for post-graduates to fail the course for which they register. Consequently, it is unhelpful to establish ‘passing the degree’ as the criterion for success. The pilot study has, therefore, investigated ‘cost’ to students and their departments as an alternative measure. How might this ‘cost’ be investigated and established?
Task-based language test specifications designed for an adult TEFL context in Morocco. Samia Belyazid, University of Illinois

This study follows the development of test specifications that generate content-based communicative achievement tests especially designed to suit the needs and interests of adult students in the context of an EFL course in Morocco. The tests are meant to bridge a gap between the teaching and evaluation of the course. Insights were drawn from Task-Based-Language Teaching (TBLT) for the application of a Criterion-Referenced Language Testing Development (CRLTD) model.

CRLTD is a proactive process-method that has allowed a group of expert TEF/SL instructors, including this researcher, to achieve the following: develop eleven test specifications that include test items called "Test-Units", and work on the gradual evolution of the content validity of these test-units. This study was done because of two important motivations: a significant literature review, as well as, the context of the English in Other Departments (EOD) course. The literature review has covered two different angles; it briefly explains how TBLT draws insights from Sociolinguistics (Hymes 1967, Halliday 1972), Second Language Acquisition research (Preston 1989, Brown 1994, Ellis 1986 and 1990), CLT (Wilkins 1972, Savignon 1983, Nunan 1984 and 1989, Richards and Rogers 1986) as well as classroom research (Doyle 1983, Crookes 1986). Second, it provides a discussion of a selected body of Task-based research studies which has led to the following realizations: TBLT is multidimensional in scope in the sense that tasks are often used in the F/SL classroom to help learners acquire competence in multiple aspects of language use ranging from linguistic, sociolinguistic, pragmatic, cognitive to educational levels; and the second motivation is that in the EOD course there is a discrepancy between teaching and evaluation which has provided a relevant context familiar to the researcher providing motivation for this study. Familiarity with the course objectives and content is one of the prerequisites of the CRLTD framework. This CRLTD project has allowed the researcher to apply the TBLT multidimensional scope to the development of eleven test specifications (specs) especially suitable for EOD learners. The specs generate "test-units" which especially cater to the needs and interests of adult EFL learners in the context of Mohammed V University in Morocco. Each test unit includes an ultimate task with subtasks, all used to evaluate the learners on various competencies and skills in the new language.

In the course of spec development, every single spec benefited from the revision of a cadre of twelve expert second and foreign language teachers from different cultural backgrounds. The nationalities represented included: five Americans, three Moroccans, one French, one British, one Japanese and one South American. Fax, e-mail and postal service mail were used to communicate with reviewers in the researcher's area. During the evaluation process, they provided written and oral comments that helped refine the specs. In addition, a regular journal was kept by the researcher who noted major stages of the writing and revision. The reviewers' contribution included critique, correction, comments, and suggestions that provided the grounds for every single one of the eleven specs to increase their content validity. There were two revision cycles over the course of three and a half months of summer 1996. This paper makes an extensive report on the developmental process of the eleven specs, as well as the results from their development and pedagogical implications of this research study. Samples of the current versions of the specs will be presented.

These eleven specs are meant to provide some groundwork for ES/FL instructors to generate task-based language test units for adult EFL learners. One important outcome is that the CRLTD model is clearly adaptable to the use of the multidimensional scope of Task in TBLT in the context of test construction. It is this adaptability which has allowed for bridging a gap between the teaching and evaluation of this EFL course. Developing tests that not only match the course objectives but also meet these adult learners' specific language and educational needs and interest is what bridges this gap.
The bilingual classroom discourse protocol: Its development and use. Alejandro Brice, Mankato State University

Teachers worldwide repeatedly face the challenge of how to best provide services for those who are learning English as a second language (ESL). Communication in English is a major barrier for many ESL students. One means of overcoming the communication (language barrier) is for bilingual students to employ code switching, or the alternation of two languages, as a bridge between the languages they speak (Flats, 1989). Code switching in the classroom is obvious and unavoidable with bilingual children (Adjure, 1988). School professionals should regard code switching as examples of communicative strategies employed by students rather than a sign of a language disability.

The purpose of this paper is to present the development of the Bilingual Classroom Discourse Protocol (BCDP). The purpose of the BCDP is to increase fairness in testing of language abilities in children, especially those children with possible language disabilities. The BCDP is currently being developed via ethnographic validation. Two ethnographic studies, that is qualitative measurements, and a meta-ethnographic summative study serve as the basis for this protocol. The first study involved ethnographic observation of an ESL elementary classroom with bilingual-Latino children. The second study, currently under investigation, is a case study of two students in both the ESL classroom and the regular education classroom. The third study is a meta-ethnographic summative review of other studies involving code switching in bilingual classrooms (Gephart, 1988). Implications for classroom code switching as teacher strategies and use of the BCDP will be given.
Differential testlet functioning in EFL reading test. Yong-Won Lee, Pennsylvania State University

Testlet is an interrelated and integrated group of items, always presented as a single unit. In reading comprehension test, a group of interrelated items are constructed from a single common text passage and are sometimes called passage-dependent items. Until recently, such a passage dependent item set has caused a serious problem in test data analysis, for the assumption of local independence underlying both the classical and item response theory (IRT) is believed to be violated in a testlet or an item bundle. However, the recent application of IRT polytomous models to test data has made it possible to analyze a set of these locally dependent items as a single unit and construct a computer adaptive testing (CAT) based on the pool of precalibrated testlets.

Along with this trend, test fairness or bias detection has also become an important issue in educational testing and measurement. A variety of sophisticated techniques have been developed and applied in the detection of item bias, such as Chi-square, Mantel-Haenszel, and IRT procedures. Nevertheless, the application of these techniques has been mostly focused on a single item as a unit of analysis as is implied in the acronym DIF (differential item functioning). Given the fact that using passage-dependent item sets in EFL reading tests is prevalent and a group of item dependent items can have a biasing effect as a whole, it is quite urgent to investigate ways to determine DIF at a testlet level and minimize it.

As part of a research project for my doctoral dissertation on testlet, three to four different versions of testlet-based EFL reading comprehension tests will be developed and administered to high school students in major cities in South Korea. The collected data will be analyzed using an IRT polytomous statistical package (MULTILOG 6.4) to detect differential testlet functioning for different gender subpopulation (male and female students). The polytomous IRT model that will be employed in the analysis is Bock's (1972) nominal response model, which has been already applied in testlet bias detection in a reading comprehension test by Wainer, Sireci, and Thissen (1991).

As a result of this analysis, first, I hope to find out whether there is a biasing effect that evades detection at an item level but is visible at a testlet level. Second, it will be tested whether Wainer et al's (1991) techniques for testlet bias detection is a powerful and economical tool in the EFL reading test context as well in comparison with other popular DIF analysis techniques. Various ways will be explored to deal with an item set that is founded to function differentially for a focal and reference group. Finally, it is also expected that the findings from this research project could provide valuable feedback to item writers and CAT developers regarding how to construct a test and an item bank in such a way to minimize the biasing effects of reading passages on examinees' performance.

In sum, the current research project could pave a new avenue for a wider application of an item set in the assessment of language proficiency with more facility and precision.
The interaction between a computer adaptive test and self-assessment of second language ability. Lawrence Myles, Concordia University

The research will investigate the interaction between a computer adaptive test (CAT) and a self-assessment (SA) questionnaire of oral ability in a second language. It is hypothesized that this interaction will be synergetic: the CAT will augment the validity of the SA responses and these in turn will yield information for instructional decisions that is not accessible to the current generation of computerized tests of language ability. The project consists of two phases: 1) the development of an SA instrument and 2) the administration of this instrument experimentally to subjects before and after they have taken a CAT of second language ability. Regression analysis will be used to analyze the relationship of SA responses and the CAT scores to other measures of oral proficiency (interviews, teacher rating, or, possibly, the SPEAK produced by ETS). A sample of respondents whose SA responses are either congruent or incongruent with the criterion measures will be interviewed to gain insight into their SA strategies.

The advantages of CAT are well documented. However there have been questions whether the complexity of communicative language use is amenable to the unidimensionality assumption of item response theory, the psychometric model that makes CAT possible. Fears have been expressed that this crucial assumption may lead to the trivialization of language ability constructs in order to enjoy the advantages of CATs. In a similar vein, the fact that current CATs use objective items almost exclusively (the assessment of constructed responses is not presently feasible) limits the inferences we can make from CAT scores. One solution to this problem is to supplement the CAT with costly direct measures of speaking and writing in order to make instructional decisions. This solution neutralizes many of the advantages of CATs. Another solution is to incorporate an SA component in the CAT.

The capacity for SA is recognized as a key element in any competence category and in autonomous learning. Moreover, the metacognitive ability to assess linguistic needs, resources, and communicative success is seen as a basic component of language competence. The research is motivated by a cognitive model describing appraisal processes. The main research question will investigate the hypothesis that a prior encounter with a CAT will increase the validity of SA responses in 2 ways. 1) The language processing demands of the test will activate declarative and procedural knowledge which, while not identical, is related to the language task demands that will be articulated in the SA items. This experience will increase the respondents' ability to accurately appraise their competence by providing each with an internal benchmark. 2) The CAT will serve as an anchor that will provide a criterion measure of concurrent validity and will psychologically reduce any tendency for learners to deliberately inflate or deflate their self-assessments. A secondary research question relates to the manner in which task demands (a vital component in the appraisal model) are rendered in the SA questionnaire. Scores from SA instruments have had varying correlations with direct performance measures, the highest occurring when SA items refer to very specific and familiar language use situations rather than global statements about skills. Prior research also indicates that the wording of SA items can induce response affects: "can do" formats may elicit acquiescence, while the negative wording of "difficulty with" items may in itself introduce error. The verbal representation of task demands will be investigated in the development and calibration of the SA scale and through subsequent interviews of respondents as part of the validity study.
Involving factors of fairness in language testing. Yuji Nakamura, Tokyo Keizai University

Theoretical background and rationale. To what degree can we give a fair test to students as a teacher or tester? What can or should teachers do to give a fairer test to students? Some students are good at interview tests, while others are skillful in tape-mediated speaking tests. Some tend not to speak about families, whereas others are likely to. Still some students are given lower grades just because they are less skillful in oral summary even if they can comprehend the reading material. There seems to be many problems to be solved in the issue of fairness.

Purpose of the research. In this paper, fairness will be investigated mainly from three viewpoints: 1) test task 2) familiarity with an interviewer 3) test method. The eventual goal of this research is to show how these factors influence the testers and testees, and to help us give a much fairer test to students.

Research design and methods. This research is conducted in terms of three aspects: test task, familiarity with an interviewer and test method. In the task factor the question is: what topic students most want to talk about? or what topic they do not want to speak about? Also they are asked whether they ask the same type of questions or complaints in Japanese which are used in the English test. In the interviewer factor, the following questions are asked: 1) Is the interviewer a classroom teacher or not? 2) Can the teacher or interviewer share the common topic for conversation with students? 3) Are the interviewers interested in the topics the students can easily respond? In the method factor the questions are: 1) Even a the speaking test, direct tests are more favored by some students, while semi-direct tests are welcomed by others. Is the test method effect influential? 2) In a speaking test, monologue type test, dialogue type test, discussion type, debate type test, all of these test methods are influential factors in the ability of speaking. Is the test method effect important? 3) In reading, both intermediate and advanced level students show almost similar performance in reading comprehension test, while they behave quite differently in reading tests in a classroom where it includes summary even if they both can read well. Does the test method affect students’ performance?

Results and Implications. The full study has not yet been completed; however, it has been found that different level students perform differently in direct speaking tests and semi-direct speaking tests. It has also been noted that interviewers’ choice of test questions influences students’ performance and may even affect the raters’ rating.

In this in-progress session, the researcher will give an interim report on the results of the three aspects. The eventual goal of the research will help understand the necessity of multiple assessment of students’ language performance in terms of fairness in testing.
WORKS IN PROGRESS

Comparing native and non-native reactions to EFL speech samples. Pavlou Pavlos, University of Cyprus

This paper compares the evaluative criteria of native and non-native EFL teachers to determine similarities and differences in instructional goals and to familiarizes foreign students with principles of spoken English discourse.

Ten American EFL teachers (L1 English) and 10 Greek Cypriot EFL teachers (L1 Greek) evaluated 10 speech samples produced by adult L1 Greek Cypriot students of English. All speech samples were in the form of an oral report and were part of an oral proficiency test battery in English. The teachers were asked to rate the 10 oral reports holistically. Next, they justified their ratings by describing the three most influential features in each oral report.

Most Greek Cypriot L1 students enrolled in universities in English speaking countries have had several years of EFL instruction before starting their studies. Since standards and areas of emphasis for Greek Cypriot EFL and American EFL speaking instruction often vary, a comparison of the two should assist EFL instructors in their efforts to inform Greek Cypriot students of the demands of oral discourse in English. This involves: (1) a description of American oral academic discourse, and (2) a description of the similarities and differences between speaking skills already learned in Cyprus and those expected at a university in an English speaking country.

Validation of holistic scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. Alfred Appiah Sakyi, Ontario Institute for Studies in Education

Concerns about the validity of holistically scored writing samples were raised in 1970 by William McColly. Now after more than 25 years, the concerns are still echoed by many researchers in second language writing assessment (Connor-Linton, 1995; Hamp-Lyons, 1991; Sweedler-Brown, 1993; Vaughan, 1991). It has been argued that second-language writers in general are particularly likely to show varied performance on different traits (Hamp-Lyons, 1995; Reutten, 1991), and holistic scales therefore make it difficult to determine the actual quality of writing concealed under a single score. Even when similar scores are obtained from different raters, it is difficult to tell whether raters are using the scale descriptors exclusively to classify ESL compositions and how they distinguish between different levels of ability (Connor-Linton, 1995).

Most correlational studies have reported the dominance of linguistic errors on holistically scored ESL compositions (Homburg, 1984; Perkins, 1980; Rafoth & Rubin, 1984; Sparks, 1988; Sweedler-Brown, 1993). However, these studies have failed to develop any theoretical basis for the continued use of holistic scoring in ESL writing assessment.

The present study goes beyond correlational studies to describe the thinking processes used by holistic raters and to identify factors that influence their decisions during the rating process. The study uses think-aloud protocol analysis (Ericsson & Simon, 1993; Cumming, 1990) to seek answers to two main questions: (1) what aspects of ESL compositions do experienced raters pay attention to while reading to evaluate them? and (2) what strategies and criteria do experienced raters use to distinguish between different levels of writing ability on a holistic scale?

The study is being conducted in two phases. The first phase was a preliminary study to define and refine variables and to model parameters to be used in the second phase. It involved 6 experienced raters who rated 12 compositions written by first year university students. The raters were guided by a locally developed 5-point holistic scale. In the second phase 10 raters will rate 12 samples of Test of Written English (TWE) essays using the TWE scoring guide. The paper will report on findings from the first phase and describe plans for the second one, now in progress.
Toward understanding the nature of communicative competence: An attempt to make sense of think aloud protocol data from reading test subjects. Randy Thrasher, International Christian University, Tokyo

The taxonomy of communicative competencies proposed by Canale and Swain (1980) and developed by Bachman (1990) and others has given a sense of theoretical legitimacy to the movement, in language teaching and testing, away from a narrow focus on structure to serious consideration of other aspects of communication. Yet the taxonomy may have less theoretical adequacy than it first appears to have. In spite of Bachman's presentation of the various competencies as a tree, we do not really understand what the relationship is between these competencies. We do not know how the competencies combine to produce full blown communication. Are the various competencies additive? Does communication equal grammatical competence plus textual competence plus illocutionary competence plus sociolinguistic competence? And, if so, are all of equal weight? If the competencies are not additive, what is the relationship among them?

This paper argues that we do not have answers to these questions because we lack an adequate theory of communication and suggests that think aloud protocol data from 15 subjects taking a 'passage with questions' reading test can be best explained using the theory of communication proposed by Sperber and Wilson (1986). The examination of this data also indicates that Sperber and Wilson's Relevance Theory provides an intuitively satisfying explanation of what constitutes an authentic test task which does not depend on a prior real world occurrence of the utterance or text used in the task or on a vague appeal to the similarity of the test task language to possible real world language use.

The paper closes with a discussion of the fairness issue raised by the crucial importance of the background knowledge of the test taker in Sperber and Wilson's view of communication.
Testing vocabulary: The effect of method on test scores and grades. Timo Tormakangas,

*University of Jyväskylä*

The Finnish National Certificate examination has been administered three times. The certificates include 7 different languages (English, Finnish as second language, French, German, Russian, Spanish and Swedish) at three levels (basic, intermediate, advanced). The test is criterion referenced, measuring levels 1-8 (roughly comparable to ESU levels 2-8), including subtests of reading and listening comprehension, vocabulary and structures, writing and speaking.

Vocabulary testing is an important part of the examination and therefore it is necessary to know what knowledge of vocabulary actually means. How many words does the candidate have to know to be awarded a certain grade? What is the effect of the testing method on this estimate of vocabulary size? On the qualitative side, how does the type of words in the sample affect the test score? What kind of words should be tested?

The research questions are addressed through a carefully designed experiment. Second language learners on approx. ability levels 1-5 are asked to do vocabulary tests in several of the aforementioned languages. The plan is to collect a vocabulary bank of a few hundred words which will be sampled from an approx. 10,000-word basic/intermediate vocabulary. Then this sample is translated into several of the examination languages in several different task types. Several techniques will be used, e.g. words in sentence context, decontextualised words, productive and receptive tasks (multiple choice, filling in gaps, translation). The sample vocabulary is also grouped thematically to examine how word categorisation affects the scores. This indicates the impact of language teaching and learning on knowledge of vocabulary and specifies the focus on content areas that are thought central for language teaching/learning. All this will help to determine the extent to which the test task words represent an appropriate and sufficient sample for awarding the different grades.

The study will provide information about the differences involved in using various testing methods and kinds of vocabulary. Also the magnitude of the method effect will be determined. This enables test designers to better determine cut off scores for different grades and how to appropriately combine tasks for use in the examinations.

By the time of the LTRC '97, there will be results from the pilot studies and their analyses. These studies are carried out to a sample of candidates with different types of test items. The pilot sample will be acquired from candidates in the English examination. Presenters wish to discuss problems encountered when planning the study and administering the pilot instruments (e.g. basis of vocabulary sample, selection of test types and quality criteria for vocabulary).
Development and use of a computer-generated task bank for assessing oral proficiency.
Weiping Wu, Greg Kamei, and Dorry Kenyon, Center for Applied Linguistics

This presentation discusses work in progress on developing and using a pool of computer-generated tasks for assessing oral proficiency. Tasks in the pool are intended for use in developing alternate and customized forms of the Simulated Oral Proficiency Interview (SOPI). The SOPI elicits a speech sample via an audiotape and test-booklet. Trained raters score the speech sample following widely-used criteria for assessing oral proficiency. Because each SOPI form contains few (7-15) tasks and all examinees normally receive the same form of the SOPI at any one administration, there is a concern that specific tasks can easily become disclosed. The researchers felt a pool of tasks would help address this problem by simplifying the creation of new forms in any language. The pool would also simplify the production of forms for special needs, such as a SOPI for first-year students only.

The first half of the presentation concerns the creation of the task pool. The basis for the creation of the pool was a database built from current SOPI tasks. The focal point of the creation was the relation between the propositional structure and illocutionary force of task directions and the characteristics of the expected response--i.e., “how does the test taker get from understanding the input to the output to provide a ratable sample of speech?” This relation is operationalized in four interrelated variables within the text of the directions: interlocutor (speaker and audience), content (topic/propositional content), function (what the speaker is supposed to do), and context (background and setting). Of these, the function of any specific SOPI task is invariant. Context and interlocutor variables, while limited to superficial textual changes, allow for great variation. The greatest variation for developing alternate tasks within a given function is found within the content variable. Nevertheless, content has to be interwoven into the task so that all variables work together to elicit speech at the given function’s performance level (as defined by the rating criteria). The researchers hope that for any function, once variable categories are defined and values are entered, the computer can generate alternate parallel tasks.

The second half of the presentation illustrates the use of the task pool through a project to develop three forms of a Russian SOPI whose tasks are drawn from the generic task pool. The discussion of this application will highlight preliminary results and issues of concern in using such a task bank. The presentation will specifically address problems arising from linguistic and cultural variations when using a generic model.
POSTERS

Posters are listed alphabetically by name of presenter.


Recently the rule-space methodology has been used to analyze the verbal domain into the cognitive and linguistic attributes which underlie performance on listening and reading test items. Results reported at LTRC 96 indicated that sets of reading and listening attributes were able to classify over 91% of test-takers into their latent knowledge states, and for those classified test-takers, the attribute sets were able to explain over 95% of the variance in total test scores. Analyses of the three sections of the SAT Verbal produced sets of attributes which could classify over 95% test-takers into their latent knowledge states, and could explain over 95% of variance in total scores for those classified.

These are very high-levels of predictability, and the question arises whether they could be an artifact of the methodology. The purpose of this study is to assess the validity of such analyses by determining whether, and to what extent, the methodology makes sense of unreliable, or inherently meaningless input.

The input into the rule-space methodology is a matrix of items by attributes, called a Q-matrix. This study will take a Q-matrix which was able to explain performance on a free response short answer listening test. In this study 96% of test-takers were successfully classified into their latent knowledge states, and using attribute scores to predict total scores the Adj R-square was .956. The Q-matrix for this analysis was then modified in both random and systematic ways, and these modified matrices then become the input into the rule space process, and the test was re-analyzed in order to see how well these modified matrices can explain performance on the test.

Modifications to the Q-matrix will be:

1. The values in the matrix (1's and 0's) will be randomly determined.
2. The attributes for the randomly selected items (1, 3 and 5 items) will be reassigned to other items that are also randomly selected.
3. Attributes will be progressively miscoded in order to examine the effects of coding unreliability on the analysis; they will also be systematically miscoded to in such a way as to examine the differential effects of unreliability on attributes which occur in many items and attributes which occur in only few items.
4. The values in the incidence matrix will be reversed in order to simulate a theory that is completely opposite to the actual underlying theory.
5. Two false and totally stupid theories will be developed to explain performance on these items. These will be based on systematic, but irrelevant features of the text and the questions.

Results will be presented and implications discussed in the light of their relevance to the interpretation of analyses of the verbal domain.
Gender and personality as factors of interlocutor variability in oral performance tests. Vivien Berry, University of Hong Kong

Research into test-taker characteristics has provided some evidence to show that differences in learners' personalities can affect scores obtained on paired oral tests depending on whether learners are homogeneously or heterogeneously paired. Significant interactions have also been obtained between learners' personality type and test task. The current study examines examples of authentic discourse produced by pairs of learners in an attempt to deconstruct their scores and show how the language produced in certain types of pairing produces higher ratings than others.

The two main foci of the study are learners' personality and gender. 32 male and 22 female learners, each of whom could be characterized as either extreme introvert or extravert participated in paired oral tests. Quantitative analysis of the overall ratings shows that both female extraverts and introverts performed significantly better when paired with another introvert whereas males performed significantly better when paired with an extravert. Qualitative analysis of the discourse produced by pairs of candidates shows how aspects of native-speaker interlocutor behaviour which have been identified in an analysis of one-on-one interviews (Lazaraton, 1993, 1996) are also present in non-native speaker interlocutor behaviour in paired discussions, especially with respect to utterances produced by female introverts.
Language distance as a factor in ESL test performance. Alan Davies and Catherine Elder, University of Melbourne

Although the concept of language distance is frequently invoked to explain the difficulties experienced by learners of particular language backgrounds in acquiring a second language, its effect on language test performance has not been widely researched. This paper considers the effect of language distance in relation to other learner background variables on performance on the 1994 and 1995 versions of an Australian ESL end-of-school examination for non English speaking background candidates, who also sat for a further examination in their own L1.

Language background information gathered via questionnaire was used to group candidates into ‘language families’ — Arabic, Slavic, Chinese, Romance, Indonesian-Malay, Japanese-Korean, Viet-Khmer — according to the language spoken at home and to identify other language use factors which might have a bearing on test performance. These language families were ranked according to their distance from English using a number of classification criteria. Correlational analyses and t-tests were then undertaken to test the following hypotheses:

1. that the relationship between performance in candidates’ first language and English (L2) as measured by the relevant examinations will be stronger for language families more closely related to English,
2. that other language background factors will also have a significant effect on performance in English
3. that students whose first language is closer to English will perform better on the ESL examination than those whose first language is more distant.

The results for both the 1994 (N = 1225) and 1995 (N = 1789) candidature indicate that, while the correlation between scores for English and the relevant L1 examination is stronger for more closely related languages, speakers of these languages do not perform significantly better on the ESL examination than those whose first language is more distant. For some language groups other factors, such as level of home literacy in English, are more powerful than language distance in determining candidates’ level of performance in ESL.

The study concludes that while language distance does exercise some influence on second language test performance, it cannot be separated clearly enough from other variables to be considered an advantaging (or disadvantaging) factor and hence a potential source of unfairness in the testing process.
Avoiding bias in C-Test construction and C-Test use. Rüdiger Grotjahn and Christine Klein-Braley, Ruhr-Universität Bochum

As with all tests, there are a number of points in the process of constructing, administering, scoring and analysing C-Tests where bias can affect the dependability of the results.

In order to construct a canonical C-Test, first a number of texts must be selected. Text selection must be guided by certain rules. Texts must be 'unmarked' on a variety of dimensions (fictionality, humour, register etc.). They must be complete, that is, they must neither assume knowledge specified in previous text sections, nor break off suddenly in mid-flow. Adult educated native speakers should obtain near-perfect scores on test texts. Further aspects which must be considered are text difficulty, suitability of content, cultural bias, text specificity in terms of discourse.

Since the canonical principle allows the construction of only one test from any given text, no constructor-induced bias operates here. However, decisions must be made about how many texts to use, and how many mutilations to make in each text. In addition specific problems occur with different languages. For instance, enclitic pronouns and bigraphs cause problems in Italian, Spanish and Portuguese, and compounds are problematic in German or Italian etc. There is the question of whether C-Tests can be used with agglutinating languages at all, e.g. Turkish and Finnish.

Instructions to the examinees can affect test scores. In multi-lingual groups only the target language can be used, but this may mean that some examinees do not understand fully what they are supposed to do. Test wisdom and practice can affect the way examinees tackle the tests. People from different cultural backgrounds show different response behaviour (risk taking and avoidance).

Scoring can be affected by examinees’ handwriting. The question of how to deal with orthographic errors and acceptable alternatives arises. Unless scores use lists of alternative solutions, unreliability can creep in.

During the process of statistical test analysis the following problems may arise: differential item functioning, multidimensionality, interdependence between items, and the legitimacy of using item analysis on individual deletions.

When C-Tests are computer-administered, additional problems may arise.

This paper will discuss the sources of possible bias listed above and suggest solutions for eliminating bias.
Creating an ESL VOCI: From script-writing to field testing. Gene Halleck, Oklahoma State University

This poster illustrates the process of creating a Video Oral Communication Instrument (VOCI). The poster examines the types of decisions encountered by test developers at each stage of test development beginning with brainstorming and script writing at a workshop sponsored by the language Acquisition Resource Center (LARC) at San Diego State University in which VOCI in Chinese, ESI, French; German, Japanese, and Russian were developed. In addition this poster illustrates the decisions associated with writing directions for the film crew and problems encountered during the actual filming.

Factors to be considered in the final stage of test development (field testing) are explored. Thus the poster also illustrates initial results of a study comparing proficiency levels obtained in Oral Proficiency Interviews (OPIS) with the ratings obtained using the VOCI for a group of twenty-five international graduate students from a variety of language backgrounds. In addition to correlating the holistic ratings of the two instruments, a closer examination of specific test items from the VOCI indicates an interesting relationship between type of questions (level checks/probes, familiar/unfamiliar topics) and proficiency level. Such findings raise questions concerning assumptions in the ACTFL Guidelines. Another interesting finding that emerges from analysis of the VOCI data is that the factor most responsible for preventing a higher rating was grammatical accuracy. Implications for further research are suggested.
Men Are from Mars, Women are from Venus: The case of an English vocabulary test.
Feltyanka Kafandjieva, University of Sofia, Sauli Takala, University of Jyväskyla

Background. Common sense experience and research (carried out by Deborah Tannen and others) suggests that men’s and women’s language and communication patterns are not identical. From the point of view of fairness, the fact that we normally measure the language proficiency of men and women with the same test (in terms of both content and format) is a potentially important issue.

There are a variety of methods, based on Classical Test Theory, for analyzing differential item functioning (DIF), but often ending up in somewhat divergent results. Moreover, sometimes it is difficult to come up with any straightforward explanation for the observed differences. Item Response Theory (IRT) has several advantages over Classical Test Theory, and one of them is that the outcomes for analyzing DIF are more consistent.

Purpose. The purpose of this study is to analyze DIF in a L2 vocabulary test using IRT and to seek explanations for the potential gender differences in vocabulary knowledge.

Method. The English vocabulary test was a part of the test battery used in the Finnish Foreign Language Certificate examination. The intermediate level test was taken in spring 1996 by 475 examinees (182 males, 293 females). The vocabulary test contains 40 MC items. The examination consists of some 120 items altogether, and measures all “four skills” plus grammar and vocabulary using both MC and constructed answer techniques.

A two-parameter model fitting the data was found. Gender-related DIF analysis for the vocabulary subset was carried out based on the 2-parameter model.

The analysis included:
* model-data fit analysis
* comparison of item difficulty parameter b, estimated separately for males and females
* comparison of Item Characteristic Curves for misfitting items
* cluster analysis of the vocabulary structures for males and females

Results. The analysis shows that more than 25% of the items differ in their functioning in the two subsamples. A first look suggests that, for some of them, there appear to be reasonable hypotheses based on probable differences in men's and women's life histories. The reasons for observed differences will be explored in greater detail, using both semantic analysis and ratings by informants.

Implications. The study shows that there are significant differences in the “passive vocabulary” of adult males and females, which may produce a biased estimation of their language proficiency. This has definite implications for the item-banking project of the Finnish Certificates system, as indeed in all likelihood for all similar projects.
Effects of native language and learning experience on the University of Illinois at Urbana-Champaign English as a Second Language placement test. Sungwook Kang, University of Illinois, Urbana-Champaign

Test performance and language proficiency have been found to be affected by various test-taker characteristics such as native language, learning style, personality, instruction type, etc. Without understanding the effects of these characteristics, using test scores for placement decision may mean disregarding the problems which are common to a group of a certain characteristic, thereby losing an opportunity for more efficient instruction. This study investigated the relative effects of two test-taker characteristics (a) native language and (b) language learning experience on test performance on the University of Illinois at Urbana-Champaign English Placement Test (UIUC EPT) to provide a better understanding of test performance, a basis of the use and interpretation of the UIUC EPT.

The subjects were 203 students who took the UIUC EPT during the August of 1994 and returned the questionnaire about their learning experience. Based on the information from the questionnaire, they were grouped according to their native language and learning experience. The UIUC EPT consisted of three sections; structure, video-essay, and pronunciation. To compare the relative effects of native language and learning experience, a two-way ANOVA was used for the structure section and a chi-square statistic was used for both the video-essay section and the pronunciation section. For the structure section, how the items functioned differentially according to the group membership was studied by using the simultaneous item bias test (SIB test). The SIB test also produced the collective amount of differential functioning (DTF) of all differential item functioning (DIF) items.

Native language appeared more influential on performance on the UIUC EPT than learning experience. There were more differences in the performance when test takers were grouped according to native language than when they were grouped according to learning experience. Grouping by both native language and learning experience did not result in group differences in the structure session of the exam. Language groups were different in both the pronunciation section and the video-essay section whereas the learning experience groups differed only in the video-essay section. Even though both ways of grouping did not result in group differences on the structure section, the item level comparison showed that test takers performed more differentially when they were grouped according to native language. DIF appeared significant only among the different native language groups but not among the learning experience groups because the collective amount of DIF was canceled out.

These findings suggests that test takers' native languages be given more weight than learning experience in placing test takers into appropriate ESL classes if variables other than English proficiency are included in the EPT administration.
Extraversion and pedagogical setting as sources of variation in performance on different English proficiency tests. Gholam Reza Kiany, University of Tarbiat Modarres and University of Essex

Trait psychologists hypothesise that the personality trait of Extraversion has a negative relationship with learning in general. In contrast, there is a tendency among applied linguists to believe that Extraversion is a positive factor in learning a foreign/second language on the ground that the high sociability of the more extravert language learners enhances interaction in the target language, hence entails better language proficiency. However, the impact of the nature of context where language is learned and the type of the tests measuring language proficiency are but two very important factors in this regard which are often ignored. Such negligence sometimes led to apparently inconsistent and contradictory findings. Following a pilot study, to examine the relationship between Extraversion and performance of L2 test takers learning English in different contexts, on different English proficiency tests, 900 Iranian postgraduates doing their Ph. D. in three English speaking countries were involved in this study. The subjects comprised three subsamples of Non-English Majors, English Majors and Immersion Group who differed according to the type and amount of exposure to and instruction in the English language. English proficiency was mainly measured by TOEFL, IELTS and a Cloze Test while the Persian restandardised Eysenck Personality Questionnaire (EPQ) were administered to measure Extraversion. The results revealed that the nature of the context of language learning and that the instruments measuring language proficiency are significant factors in the way in which extraversion is related to foreign/second language proficiency. The implications are discussed in relation to fairness in language testing as well as foreign/second language learning.

Washback of tests on teaching: Is it fair to students? Rama Mathew, Central Institute of English and Foreign Languages, Hyderabad

What ultimately matters about an educational programme is how well learners perform in the end-of-course test. Such summative or end-of-course tests have been known to have a very powerful washback on teaching and learning in the classroom. This paper reports on an empirical study that examines a summative test that was introduced at the end of a secondary level ESL course and the kind of impact it has had on the teaching-learning situation. It focuses especially on the teacher-made class tests that are used to evaluate students’ progress during the course.

The study involved analysing the summative test administered by the secondary board and the class-tests constructed by different teachers in a variety of schools. The analysis reveals that the class-tests, regardless of when they are administered during the course, are modelled on the summative test in terms of the skills covered and their weighting, test formats/methods and the criteria used for evaluation.

The paper argues that such an inappropriate use of progress tests raises questions of fairness to students, especially the low achievers whose abilities and interests are not adequately accessed and accommodated. It makes a plea for a continuous process of test validation in class in terms of levels and formats from the perspective of different learners. This would have implications for learner and teacher training in the construction and use of tests in class.
Ratings, Raters, and Test Performance: An Exploratory Study. Beryl E. Meiron, Mary-Erin Crook, Yutaka Kawamoto, Laurie Schick, and Carmen Velasco-Martin, University of California, Los Angeles

While there is growing evidence that content-based instruction (CBI) contributes to overall language improvement (Peek, 1987; Brinton, Snow, and Wesche, 1989; Valentine and Reparth-Martos, 1992; Snow, 1994), the assessment of oral language skills gains has not always shown where or how CBI has contributed, partly because previous studies have focused either on quantitative or qualitative analyses of test results (Lazarton and Riggenbach, 1990; Young and Milanovic, 1992). This study uses quantitative and qualitative analyses together to identify specific instances of CBI effect on oral skills gains.

This study will compare pretest and post-test scores of a direct oral proficiency test (OPT) administered to twenty-five Egyptian EFL teachers in an eleven-week teacher training program. The OPT prompt required an argument to support a pre-assigned position whose content directly related to the training program and to the professional interests of the test takers. A five-category rubric was developed to assess control of topic, lexicon, grammar, pronunciation and conversation. Four raters were normed and paired to administer and rate the videotaped pre- and post-test sessions. Statistical comparisons between the pre- and post-test results indicate overall gains with the most statistically significant gain in conversational control. Since quantitative information is insufficient to analyze the basis of these gains, discourse analysis comparing transcripts of pre- and post-test videotapes will be used to fill the gap. Details of the findings which connect gains to training along with implications of the assessment methods will be presented.

The interlocutor as a variable: Affective factors in language testing. Barry O'Sullivan, Okayama University

The linguistic performance of a language user may vary in characteristic, describable ways with respect to choice of linguistic forms used to convey the message, and with respect to level of formal accuracy in the language deployed. That this variation may occur systematically in relation to specifiable linguistic, sociolinguistic and/or psycholinguistic variables — its systematicity indicating its rule-governed nature — has been supported by academic research and discussion over the last thirty years. In the area of language testing this variation is most pertinent to the oral proficiency interview (OPI), where candidates experience a face-to-face examination of their linguistic ability.

Studies such as Bachman et al (1995) and Kunnan (1995) have added significantly to our understanding of how test-taker characteristics affect candidate performance on language proficiency tests. However, there remains a question mark over the likely effect of these characteristics on candidates’ performance on OPI tests, since the interlocutor(s) in an interaction (i.e. the interviewer and/or another candidate, depending on the format used) may be seen as a variable which affects linguistic performance.

This paper reports on research into this potential threat to test validity in which the effect of so-called ‘Affective Factors’ (Porter 1991) have been examined. To date those factors studied include the relative age, gender, status, speech style, personality, and acquaintanceship of the interlocutors. Results indicate that characteristics of one’s interlocutor constitute an important set of variables in their effect on final examination scores. The consequences for language testing and teaching are considered.
The effect of proficiency on item difficulty, reliability, reproducibility, and dimensionality in a second language reading comprehension test. Kyle Perkins, Southern Illinois University at Carbondale

A thirty-item reading comprehension test was administered to the same sample of thirty-one ESL students at three different strata of proficiency. Each question was identified as belonging to one of two types: lower-order which required the readers to conduct a cognitive operation on an explicit piece of information in the text, and higher-order which required a cognitive operation on one or more implicit pieces of information.

It was hypothesized that there would be a significant different in the number of correct responses at each occasion of testing due to growth and that there would be a significant difference between the responses to the two different question types due to differences in cognitive processing entailed. A repeated-measures analysis of variance showed a significant main effect for proficiency level, but no significant main effect for question type or for the interaction of proficiency level by question type.

The estimates of reliability increased with each test administration indicating an increase in person-separability, suggesting differential growth in the examinees in the skills assessed by the instrument. Although the average item difficulty for all items for the different test administrations increased, the growth patterns for some of the individual items was anything but smooth. The more graphic examples will be illustrated.

The responses were submitted to Guttman scaling for estimates of reproducibility and scalability. The effect of proficiency level on dimensionality was of interest because dimensionality is sample-dependent to the extent that the same items may be found to be unidimensional with one sample of examinees but multidimensional with another sample. The coefficient of reproducibility increased across test administrations, but never reached the criterion for demonstrating that the items formed an ordered set. The coefficient of scalability decreased across test administrations and never reached the criterion for demonstrating that the scale was truly unidimensional.

Scott, Stansfield, and Kenyon (1996) described the first listening summary translation exam (LSTE), which was developed for Spanish. The test is used by law enforcement agencies to determine if an applicant can understand tape recorded telephonic conversations dealing with criminal activities and write an accurate and comprehensible summary of the conversation in English.

This poster session will present the development, field testing, and validation of an LSTE in Southern Min, a language spoken by 40 million people in the Peoples Republic of China and Taiwan, and 10 million in other countries including Malaysia, Singapore, Thailand, the USA, and Canada. The test consists of a multiple-choice section containing 50 items based on 12 conversations. The summary translation section consists of a module which teaches the examinee how to write a summary, followed by three conversations which the examinee must summarize. All conversations are classified on the ILR scale, with levels ranging from 1 + to 4. The summaries are scored for Accuracy using a checklist of important points that should be included. They are scored for Expression using a four point holistic scoring guide developed for evaluating the short summary genre.

Seventy-two examinees took one or two forms of the exam. The examinees also completed a personal background questionnaire and three self-assessment instruments developed for this study. These instruments permitted the examinees to rate themselves on a simplified ILR scale for listening proficiency and writing proficiency, and for the ability to summarize four types of telephonic conversations tested on the exam.

Following pretesting, classical and latent trait analyses were performed on the multiple-choice section and the Accuracy checklist items using classical test analysis software and BIGSTEPS. The analyses were used to delete 12-15% of the items without reducing the internal consistency reliability. Subsequently, the shortened tests were rescored and scores were correlated with the self-assessment measures. Linear equating methodology was used to equate form B to form A. Both sections of the tests were found to be highly reliable. For a subgroup that took both forms of the test the parallel form reliability was .95. The interrater reliability of the Expression score ranged from .70 to .85 for each summary, with the overall Expression score demonstrating adequate reliability to permit the use of a single rater in an operational program. The pattern of correlations between the tests, the self-assessments, and the background variables shows that Accuracy and Expression assess two unique constructs, justifying the use of separate scores for each. Language use data collected on the background questionnaire provide evidence of a language shift from Southern Min to English among first and second generation immigrants to the US. This evidence is corroborated by test scores.
Testing reading comprehension: How fair is our choice of the method? Constance Tsagari, Athens, Greece

Over the years, test designers, in their attempt to assess the skills of reading comprehension, have developed a multitude of methods. The question is whether these different procedures are all testing the unobservable mental operations involved in reading comprehension in the same way.

The purpose of this study was to investigate the effect that two commonly used testing techniques (i.e. multiple-choice and free-response) have on the assessment of reading comprehension of an EFL text. The study drawing on from previous work done in the field by Alderson (1990), Allan (1992), Anderson et al (1991), Gordon (1987) and Shohamy (1984) is based on both a product- and process-oriented approach.

The two reading tests, together with Checklists of test-taking strategies and Retrospective Questionnaires were administered to 57 non-native students of English. They were asked to take the two tests (one followed by multiple-choice items and the other by free-response ones) on separate occasions and introspect on their test-taking strategies immediately after the completion of every question. At the end of every test, they were also asked to fill in a Retrospective Questionnaire relating to more general reading strategies, task difficulty and students' perceptions of the two formats. Two of those students took part in think-aloud interviews on the same tests to measure the validity of the self-report instrument.

The analysis of the scores showed that two tests with identical content but different formats may not yield measures of the same trait. This was further evidenced by the frequency with which students selected strategies from the Checklist in order to describe their ways of processing the same items in the two formats. The double-check on the validity of the Checklist of test-taking strategies proved that the instrument was reflective enough of the students' strategies.

The results indicate that it is necessary to define as closely as possible the attributes of our testing methods and take these into account when devising reading comprehension tests so as to control their effect on examinees' performance. This is directly linked to the needs of test designers and teachers who would like to be aware of the construct validity of their tests and fair when selecting specific testing methods over others.

The poster will demonstrate the actual tests used and instruments devised and will call for participants' feedback and suggestions for their improvement.