17th Annual
Language Testing Research
Colloquium

1995

THEME: Validity and Equity Issues in Language Testing

Program and Book of Abstracts

Long Beach Hilton Hotel, Long Beach, California
March 24 - 27, 1995
A Student Centred Perception of Test Validity

Marian Tyacke, University of Toronto

Skehan (91) and many others have criticized test developers for concentrating on consistency of measurement at the expense of validity. However traditional measures of reliability rely on grouping individual responses which assumes that each item is measuring the same thing for each individual. As there seems to be some evidence that this is not the case, then we have to find other ways of determining what each item measures. If test-takers are actually performing different cognitive tasks from each other, even when dealing with a very straightforward M/C item on a reading test, then how can we talk about 'internal consistency', 'comparability' and many other aspects of traditional test analysis?

Alderson (91) advocates involving test-takers more in test development, especially as they react very differently to the same items. Also, there are many studies of test method effect which indicate that item format affects individual performance (Ward, Frederiksen & Carlson 1980, Bachman & Palmer 1981, de Jong 1983, Samsom 1983, Shohamy 1984 inter alia). In spite of such studies, test developers still design 'high-stake' language proficiency tests without analyzing the effect of different task types on individual learners. Candidates with different learning styles (for example) may not be given the opportunity to demonstrate their proficiency adequately. In other words, a particular task type might provide a significant advantage, or disadvantage, for a particular learner type.

The poster session describes a test development project which had a pre-test and a post-test phase. The former involved test-takers in the development of a reading test for university entrance by having them introspect as they completed a variety of items (M/C, free response, summary, cloze) in terms of interest, time taken, what they felt each item tested, how they went about answering it, and level of difficulty. They also completed a profile which included questions on learner type. Their perceptions were then compared to three reading test developers and three English instructors who were asked to complete the same checklists, and learner profiles after responding to the same four task types. Responses were then triangulated in order to establish a detailed analysis of each item. They were also examined to see whether there were any systematic differences between 'naive' and 'expert' opinions. Finally, their responses to the item were related to learning style.

In the post-test phase, test takers responded to a questionnaire which included a self-report on learning style, and attitudes to the same four tasks. Learning style was then related to success on the whole test, to success on each task, and to attitudes to the tasks for 307 (ethnically matched) test-takers. Results will be presented in detail.
TOEFL 2000 Project Update

Carol Taylor, Educational Testing Service

At LTRC 1994 a poster session was presented on the TOEFL 2000 project with an emphasis on Stage I efforts that were underway. Stage I was intended to establish a framework for the project through a number of concurrent activities. These included a profiling of TOEFL score users and examinees, a survey of TOEFL score users and their needs, consideration of the constructs relevant to the new test, and reviews of psychometric issues and literature, scoring scales, and technology relevant to test development and delivery.

This poster session reports on the outcomes of Stage I efforts and on the initiation of Stage II of the project. Stage II, estimated to last three years, is focusing on research, test design, and an examination of the feasibility of various prototypes and delivery systems. It will explore the second language abilities required in North American colleges and universities and will consist of a number of prototyping cycles to design, construct, field-test, analyze, and refine test models. At the culmination of Stage II, the project expects to have a final prototype(s) ready for large-scale piloting and validation.

A written project update will be available and members of the TOEFL 2000 project team will be available to discuss the outcomes of Stage I and the efforts being pursued in the first phase of Stage II.

Constructing EBB Scales: Developing Empirically Derived, Binary Choice, Boundary Definition Rating Scales

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

This poster session presents a rationale and methodology for the development of empirically derived rating scales. The procedure has been employed in making rating scales for ten different ESL oral tests for 5th and 6th grade learners.

Second language testing increasingly uses rated tasks in place of objectively-scored items. Rating language performance is more demanding than scoring discrete-point tests. Lower reliability and validity for ratings are to be expected, especially in instructional settings where differences among students are small, and training for the use of rating scales is infrequent. A need for rating scales has not always led to effective and efficient scales, however. Adopted rating scales may not be appropriate for targeted groups when those scales are based upon speculations about language development or derived from performances of other groups. Standard rating scales require the matching of examinee performance to a verbal description. An alternative type of scale consists of an empirically derived, ordered set of binary questions relating to boundaries between levels on the performance being evaluated. Rating depends on a series of decisions.

This poster session will show the procedures for developing these EBB scales as a group of judges reaches consensus on rankings and significant features of language samples. Examples of EBB scales, together with technical information, will also be shown. Reasons are offered to suggest why this type of scale may ameliorate reliability and validity problems associated with standard rating scales.
The Development of a Self-Instructional Rater Training Kit for a Standardized Speaking Test

Charles W. Stansfield, Center for Applied Linguistics

The Spanish Speaking Test (SST) was developed under a grant from the US Department of Education to the Center for Applied Linguistics (CAL). The test follows the format of a Simulated Oral Proficiency Interview. The SST contains 15 speaking tasks, spread evenly across the ACTFL Intermediate, Advanced, and Superior levels. CAL is developing a self-instructional rater training kit. With the availability of the kit, it will no longer be necessary to conduct live rater training workshops. Instead, foreign and second language educators will be able to use the kit to train themselves to rate performances, and to retrain whenever necessary without cost.

A study involving a quantitative comparison of self-instructional and face-to-face rater training was presented at the 15th LTRC (Kenyon and Stansfield, 1993). The study showed that raters can be trained to rate reliably through either method, as long as the self-instructional kit is carefully constructed.

This poster session will present the components of the self-instructional rater training kit that accompanies the SST and describe in detail the methodology for developing such a kit. The principal component of the kit is a Rater Training Manual that introduces the test and the ACTFL Proficiency Guidelines, which are used to score the test. In addition, the Manual discusses in detail each task on the test, indicating the content characteristics of the task, the language functions it is designed to elicit, and the probable ways in which the task will be handled by examinees at each ACTFL proficiency level. The kit contains several tapes, which are coordinated with a Reference Guide. One tape is used for initial training. The speech segments it contains provide examples of examinees performing at a variety of levels on each task. The Reference Guide contains a written analysis and justification of the official rating assigned to each speech segment. The kit also contains exercise tapes. The rater listens to the exercise segments and decides on the appropriate rating. The rating may be confirmed by checking the key in the Reference Guide, which also contains a justification for the rating. Finally, the kit contains a Practice Calibration Tape. The trainee may use this tape as a test to determine if he or she is scoring with adequate reliability. A key and justification also accompanies this tape. If a rater passes the practice calibration test, he or she may order a Final Calibration Tape from CAL. CAL will score the ratings to this tape for accuracy and award a certificate of competence if warranted.

The methodology used to develop the kit includes obtaining confirmatory ratings from a national panel of advisors, then drafting the rating justifications and having them reviewed by the advisory panel, and field testing a complete draft version of the kit on twelve local professors of Spanish. These field test subjects provide detailed qualitative information both in writing and at a group debriefing following completion of the program. Their suggestions for revision inform the final changes on all components of the kit.
Overview of Initial Research to Design and Validate the Revised Test of Spoken English

Jacqueline Ross, Educational Testing Service
Dan Douglas, Lancaster University
Grant Henning, Pennsylvania State University

A revised Test of Spoken English (TSE) is being developed as a test of general English speaking proficiency for nonnative speakers. The test items are based on oral communication tasks and the spoken responses are scored according to a model of communicative competence.

There will be a short description of the triangulation methods employed to validate and confirm the theoretical framework, design specifications, and scoring rubric of the revised test.

Three research and development projects will then be reported in view of their contribution to the initial design and validation of the test. The first part of the paper will present the theoretical model and underpinnings of the revised TSE, the second will offer findings of a research project comparing the current TSE with the revised prototype test, and the third will provide a review of discourse analysis data compiled on the responses of native and nonnative speakers to the test's language tasks.

Validation of Tests of Search Reading by Means of Protocol Analysis

Digna Samson, Cito, The Dutch National Institute for Educational Measurement

The Dutch National Institute for Educational Measurement (Cito) provides all tests for the central examinations of secondary education. It has recently constructed new tests of reading comprehension of English as a foreign language after their content specification had undergone a drastic change, following proposals for a new examination syllabus. Especially one of the objectives (requiring students to find specific information in all sorts of texts) necessitated the development of new tests. For validation purposes we have not only administered them in the classroom, but we have also subjected them to protocol analysis. The two main questions we wanted to answer were: 1/ do the students choose the most effective reading strategies considering the test objective? 2/ are students were willing enough to undertake the rather daunting task of having to deal with comparatively long passages followed by a single question?

In the poster session the results and conclusions from the protocol analysis will be discussed.
Testing Educationally Disadvantaged, English Second Language Students for Admission to University Study

Nan Yeld, University of Cape Town

The paper discusses the development of an innovative approach to testing at the University of Cape Town in South Africa. In 1986, the university established a project which has as its primary aim the identification of talented black students from the huge number of applicants from the Department of Education and Training education, which was systematically under-resourced and oppressed under apartheid. In addition to being severely educationally disadvantaged, applicants from this system are all speakers of English as a second language, and thus face a double disadvantage in terms of university study.

The poor educational background of the students has meant that traditional approaches to testing are inappropriate: any test based on a notion of what students ought to be able to do in the target situation tends to result in low scores with a very low standard deviation, effectively only predicting the past, and not helpful for selection to university study. In response to this, the approach developed by the Admissions Project aims to assess ability to respond to educational intervention: it thus rests on the Vygotskian notion that appropriately designed mediation can overturn poor prior learning experiences. This implies the incorporation of a teaching/learning element into the selection tests themselves.

The central assumption of the testing approach adopted by the Admissions Project is thus that the provision of carefully structured and sequenced tasks - designed to alert test takers to salient features of test texts and to prepare them for the demands of the writing tasks in the test - will affect performance on these writing tasks. Clearly, if this assumption is valid, there are important implications for the provision of equitable access opportunities via testing.

In 1992, a test to replace a more traditionally oriented language proficiency one was developed by Project staff in an early attempt to investigate the use of tasks as teaching agents within tests. Broadly speaking, results on this test - the Placement Test in English for Educational Purposes (PTEEP I) - did yield a greater spread of scores, and a slightly better correlation with subsequent academic performance in some faculties. Criticisms of this test focused on the covert and somewhat random nature of the task 'scaffolding'. In response to these and other criticisms, a new test was developed in 1994, PTEEP 2. A test development team comprising language testers as well as Science and Language materials developers attempted to create a more effectively sequenced and structured teaching/learning environment through the use of tasks.

In order to investigate the validity of the central assumption of the testing approach as outlined above, it is necessary that PTEEP 2 be comparable in certain key respects to PTEEP 1. Attempts were made to establish comparability as follows: the same two texts are used in both tests; the final writing task, an essay of contrast and comparison, is identical; and the other two writing tasks (summary and description) are similar, although not identical.

Indications from PTEEP 1 are that educationally underprepared students are indeed responsive to in-test teaching. The central question addressed in this paper is this: assuming a stable testing sample for the two tests, has the more structured, overt and sequenced use of test tasks of PTEEP 2 resulted in different writing task performance when compared with performance in the same tasks in the earlier test? (Performance here refers not only to an overall index, but to that across components such as 'Content', 'Organisation' and so forth). Further questions include the following: are there greater differences between some of the writing tasks across PTEEP 1 & 2 than between others? Are students applying for different discipline areas (e.g. Arts vs. Engineering) discernible as separate groups in terms of the above questions? What is the relationship between performance on the scaffolding tasks themselves and the writing tasks?

The new test is due to be written in mid-October by the first group of approximately 1500 students. The paper will report on findings from this project, and will discuss some implications for testing.
Characteristics of the USA Government’s New Oral Proficiency Testing System

Eduardo C. Cascallo, Center for the Advancement of Language Learning
Julie Thornton, Center for the Advancement of Language Learning

The issue of foreign language proficiency testing is currently being debated within the US government and in European organizations. Within the US federal government, the Center for the Advancement of Language Learning (CALL) is currently sponsoring a long-term proficiency test development effort with the collaboration of various language training and testing organizations within the Departments of Defense, Justice, and State. This effort, called the Unified Language Testing Plan, was submitted as part of a government wide project with the goal of increasing cooperation and reducing duplication of effort within the federal government.

This plan calls for a working group in proficiency testing, the Federal Language Testing Board (FLTB), to examine agency-specific needs for each of the four language skills (speaking, listening, reading, and writing) and to design and validate a proficiency instrument for each skill to be used by federal agencies in hiring, selection, and assignment of personnel to positions that require foreign language skills. These general proficiency instruments will measure sub-skills required by all of the agencies and will be supplemented with additional agency-specific assessment modules for skills required by only one agency. Over the past two years, the FLTB has been concentrating its efforts on the assessment of speaking proficiency, and is currently pilot testing a new speaking instrument. This poster session will characterize the new test, comparing it to the Oral Proficiency Interviews in use in the past by several federal agencies, including an overview of its theoretical framework, a synopsis of task and level interactions, and a review of some of the implications the new elements of the test have on the rating of proficiency levels. A full description of the new factor structure of the new oral interview will be provided, together with the theoretical background for each one of the factors defined for the rating process. Recommended elicitation techniques will be discussed in detail, as well as the training curriculum for all interagency testers.

This poster (and paper) will compare the efforts at CALL with efforts by the Council of Europe, private language testing organizations in Europe, and other groups for the development of language proficiency assessment instruments and common scales. Common elements and those that represent different points of emphasis will be discussed.

The poster will also present results from a pilot study currently in progress, in which the reliability and validity of the new oral proficiency instrument is being assessed. Two teams of testers from each of four government agencies will administer a total of 360 oral proficiency tests. All interviews will be videotaped, and these tapes will be used to obtain double ratings to determine intra-agency reliability. All tester teams and examinees' order of testing will be counterbalanced to avoid effects of practice and other confounding variables. The same procedure will be used for the pilot testing of adult speakers of three languages (Spanish, Russian, and Chinese). All statistical analyses performed will be reported in the poster. The information compiled will also include data on student and tester feedback regarding the new assessment methodology. A detailed analysis of new elicitation techniques used, rating procedures followed, and proficiency scores will be reported in the poster.
Coordinated Language Testing Plans: A Summary of Current US and European Efforts

Eduardo C. Cascallar, Center for the Advancement of Language Learning
Julie Thornton, Center for the Advancement of Language Learning

The issue of foreign language proficiency testing is currently being debated within the US government and in European organizations. Within the US federal government, the Center for the Advancement of Language Learning (CALL) is currently sponsoring a long-term proficiency test development effort with the collaboration of various language training and testing organizations within the Departments of Defense, Justice, and State. This effort, called the Unified Language Testing Plan, was submitted as part of a government wide project with the goal of increasing cooperation and reducing duplication of effort within the federal government.

This poster session will characterize the new speaking test developed under the Unified Language Testing Plan, comparing it to the Oral Proficiency Interviews in use in the past by several federal agencies, including an overview of its theoretical framework, a synopsis of task and level interactions, and a review of some of the implications the new elements of the test have on the rating of proficiency levels. A full description of the new factor structure of the new oral interview will be provided, together with the theoretical background for each one of the factors defined for the rating process. Recommended elicitation techniques will be discussed in detail, as well as the training curriculum for all interagency testers.

In addition, this poster will compare the efforts at CALL with efforts by the Council of Europe, private language testing organizations in Europe, and other groups for the development of language proficiency assessment instruments and common scales. Common elements and those that represent different points of emphasis will be discussed.
Causes of Sex Differences in Foreign Language Reading Comprehension

Karin Bügel, Cito, The Dutch National Institute for Educational Measurement

The scores obtained by girls on the Dutch national foreign language examinations, administered at the end of secondary school, have been consistently lower than those of boys. This paper addresses the issue of the possible causes of these differences.

The observed sex differences are considered in a wider context of academic achievement in general. The causes for such differences mentioned in the research literature are reviewed. Subsequently, a study will be reported on which was conducted among a large numbers of subjects to test the hypothesis that the differences in achievement between male and female students are caused by the content of reading comprehension tests.

A previous study on item bias showed similarities in content of biased items and text passages. The content areas involved belonged to the more or less stereotyped interest domains of males and females. In spite of this it appeared to be very difficult to make predictions as to the occurrence of bias in reading comprehension tests.

The study described in this paper explores three questions:
1. Is it possible to construct foreign language reading comprehension tests that favour male or female students?
2. Are there gender differences in prior knowledge and interest that are relevant for reading comprehension tests?
3. Could possible differences in test scores be explained by such gender differences in prior knowledge and interest?

The experiment was conducted among students from different levels of Dutch secondary education who answered multiple choice items about English texts and completed a number of questionnaires about their prior knowledge and interest, reading and television watching habits with respect to the topics of the English texts in the reading comprehension tests.

The results of the statistical analysis provide strong support for a positive answer to the above mentioned research questions.
Beyond Quantitative Item Analysis: A Qualitative Description of EST Reading Comprehension Items

Nora Villoria, Universidad Simon Bolivar, Caracas

The main purpose of this research project is to describe the qualitative characteristics of the most difficult and the easiest items used in six EST reading exams at Simon Bolivar University (USB) in Caracas. This analysis will determine whether or not these items share any similarities and this, in turn, should help us to establish some of the main difficulties students had with these items.

At USB, students have to take a three-trimester English course in scientific reading. Each trimester approximately 1000 students are given two multiple-choice exams to evaluate their reading achievement. The exams generally consist of 25 "modular" questions. After the exams are given, each question is stored in an item bank with statistical information about its level of difficulty and discrimination. In order to systematize the storage of these items in the bank, a taxonomy was designed by a group of professors from the department (Champeau, Marchi and Arreaza, 1993).

To carry out this qualitative item analysis the following steps will be carried out:
2. classification of the items according to their difficulty level in two general groups: easy items (100%-65% of correct responses) and difficult (45% to 0%).
3. analysis of each "modular" item using the taxonomy used by the USB item bank which takes into account the text, the stem of the question and the four options.
4. comparison of the items within the two groups of difficulty in order to find out which aspects of the taxonomy are similar in each group.

Based on the results of this item analysis, conclusions will be made with regards to the difficulties students had with these items. Important implications of this study could be the improvement of our first-year English exams and our teaching of scientific reading comprehension students at USB.

This analysis will also shed some light regarding the usefulness of the taxonomy used by the item bank at USB.

Investigation of Variables Associated with Entry Proficiency Versus "During-Learning" Success: Taxonomy Development

Kim Hughes Wilhelm, Southern Illinois University

The researcher will briefly describe taxonomies developed for two previous studies, one examining LL background variables associated with entry English proficiency of 300 Malay students entering an American-Malaysian cooperative university program and the second examining LL variables associated with learner progress through levels of an IEP (intensive English program) in a midwest university setting.

A third "knowledge" bank is currently under development. The researcher plans to draw upon results from the previous two studies to examine entry proficiency and progress through a pre-university IEP. While the previous studies examined only background features associated with language learning success (i.e. before enrollment in the IEP), the study under discussion will also examine variables predicted to be influential while the learner is enrolled in the IEP. Discussion and suggestions will be solicited as to what variables to include as potentially influential as "during learning" variables (e.g. skills and strategies, high interest materials, immersion and community contact opportunities, living situation) and ways in which features and feature values should be coded and defined.
Facilitating Success: Language Testing and the First-Year Experience

Tim Pychyl, Carleton University

As part of on-going research into first-year student attrition and retention, we are collecting data on language ability as a factor related to students' overall academic performance. Student attrition and retention are chronic and serious concerns of university administrators and faculty. Although student attrition varies among faculties and disciplines, as many as fifty percent of first-year students do not continue to a second year of study. Many factors have been identified as important contributors to a successful "first-year experience" including social networks, support resources on campus, size of classes, effects of part-time employment, etc. Our research project examines language ability, as measured by a variety of language tests, as one factor in the prediction of student success.

The first stage of this longitudinal project is based on data collected from over 600 first-year students at a large Canadian university. The total sample was divided into three randomly assigned groups. All students completed a self-assessment questionnaire related to academic skills and the Canadian Achievement Survey Test for adults (CAST). Each group also received a different subset of assessment instruments: Group 1) the Canadian Adult Achievement Test (CAAT), Group 2) the Nelson-Denny Reading Test, and Group 3) a performance-based, English for Academic Purposes language test.

In this session, we report on the results from the English for Academic Purposes language test, originally developed for use with second-language students for University admission purposes. We present the academic language profiles of these native-speaker, first-year students, and the new band scores created for the testing of reading, listening and writing at higher, native-speaker, levels. Handouts of a sample test and profiles will be provided to participants. As this is a work in progress session, we would appreciate feedback on our approach to the construction of language profiles for the identification of students at risk. We are also interested in discussing issues related to the social consequences of test use for these purposes, and perceptions about how large a factor language ability is in both student retention and overall academic performance.
Candidate Performance on Direct and Semi-Direct Versions of an Oral Proficiency Test

Kieran J.O'Loughlin, University of Melbourne

Objectives of the research
This paper investigates the comparability of direct (face-to-face) and semi-direct (tape-mediated) versions of the oral component of a large-scale English test for prospective immigrants to Australia, the access: test. The focus of the study is on the use of ethnography to inform the interpretation of test results.

Background and rationale
Previous research comparing direct and semi-direct tests of oral proficiency has focused mainly on test scores (e.g. Stansfield, 1991; Stansfield and Kenyon, 1992) and the language produced under the two conditions (Shohamy, 1992; O'Loughlin, 1994). There has also been some investigation into candidate reactions towards the two formats (Stansfield, 1991). In the case of the access: test, these reactions have also been contrasted with those of interlocutors and raters using questionnaires and explored in relation to test scores (Hill, 1994). Finally, another recent study (Luoma, 1994) has examined a variety of quantitative and qualitative data such as test scores, transcripts of language output, questionnaire feedback and interviews to investigate the issue of test comparability. The use of qualitative data and analysis in conjunction with quantitative information to investigate the comparability of the two kinds of the tests appears to represent the most promising approach to date on this question.

Research method
The current study combines the use of ethnography with multi-faceted Rasch analyses of test scores to examine the main research issue. Qualitative data was collected on how the two versions were perceived by all of the trial test cohort, interlocutors (on the live version), raters, test development team and administrators. The study also focused on two trial candidates using close observation of their performances and interviews with each of them after undertaking both versions, with their respective interlocutors and with the raters who assessed them as methods of data collection. The findings were then related to the test results of all candidates but particularly those of the two subjects tracked in the trial.

Results
While the majority of trial candidates performed similarly on the two versions of the test, both of the candidates tracked in the study (as well as a number of others), achieved quite different results on the two formats. On the basis of preliminary analyses the discrepancies in their results appear to be attributable to a number of factors related to how the two candidates interacted with each of the test formats (including the interlocutor on the live version) and how their audio-taped performances were interpreted by the raters.

Implications
The findings suggest overall that the live version is considered potentially to be the more valid of the two test formats. However, there appear to be features arising from the candidate's interaction with either format which may negatively impact on their performance, some of which are more open to improvement than others.
Validation in the Early Phases of Test Development

Sari Luoma, University of Jyväskyla

This session presents an ongoing validation study. The test context is a new national non-academic, general purpose proficiency test in several second/foreign languages. There are three test levels, each containing five subtests; the four skills and a separate subtest for structures and vocabulary. The certificates report profile scores on a band scale. The envisioned score users are employers, educational admissions officers and the participants themselves. The two former are likely to use the scores as one criterion in employee or student selection, the participants as proof of ability and a guideline for further study.

The scope of the present study is the various stages in test development up to early official implementation, Weir's (1988:21-27) "a priori validation." For Weir, this process entails 1) stating the theoretical assumptions underlying the test, 2) developing a test specification, 3) judgmentally validating this and the actual tests by having experts scrutinize them, and, importantly, 4) investigating the test-taking processes by employing ethnographic procedures. The present study builds on largely the same ideas but from the perspective of Messick's (1989:13) position where the goal of validation is advanced understanding of what the test scores mean. This broader definition of a priori validation also includes recording the planning and implementation of the quantification process that turns test performances into scores, and the investigation of the assessment process. The session presents the validation process and describes the development of a tentative model of language ability. The emerging model consists of our assumptions of language ability as we implement them in test development. Existing descriptive models of the Canale and Swain family form the theoretical basis for the model, but aspects from more process-oriented models such as those of Cummins, Bialystok and Widdowson have also been incorporated. The context-bound input comes from assumptions made in choosing particular types of tasks, assessment and quantification procedures in the development of this particular test. The purpose of stating our assumptions is to explicate what we think the test will measure, and to submit the hypotheses derived from these sometimes vague statements of faith to testing. In developing the model, we are forcing ourselves to focus on the meaning of scores.

Questionnaires and interviews have been used to solicit expert comments on the tentative model, the test specifications and test tasks as well as the assessment procedures. Think-aloud protocols and retrospective interviews have been used as methods in investigating the test taking and assessment processes. In the session, comments are invited on the model in its present state, as well as suggestions for procedures to be followed in future stages of the validation process.
The Use of Verbal Protocols to Validate an Adaptive Test

Michel D. Laurier, *University of Montreal*
Mee Lian Chung How, *University of Montreal*

Current approaches to validation of computer-adaptive tests (CAT) usually do not take into account the differences due to a different testing mode. Data for item calibration of computer-administered items is often obtained by means of paper-and-pencil versions. Correlations between conventional and adaptive versions of the same test are usually high and do not provide information about differences related to the process which are undergone during a CAT administration.

A language CAT is under development for placement purposes in French. A prototype is now operational. It consists of three different parts using multiple choice items:

1) The student reads a paragraph (about 30 words) and answers a comprehension question.
2) The student reads a situation in (in L1 and answers a comprehension question.
3) The student fills in the gap in a sentence (vocabulary and grammar items).

Typically, since it presents only the items that are neither too difficult nor too easy, a CAT is shorter and less frustrating. It can be as accurate as a conventional test but the estimation of the ability level is obtained with a different procedure. A first study has been conducted with the prototype in order to compare the results on a conventional version of the test and the CAT version. Placement differences seem to be caused by the scoring method (number of right answers on the conventional version vs. maximum likelihood on the CAT version). A questionnaire and a retrospective discussion have also been used to analyze the students' perceptions on both versions. Surprisingly, the analysis did not show major differences in students' perceptions on aspects such as task difficulty, duration or test anxiety. However, it suggested that the test strategies that are currently used on a language test using multiple choice items do not work in the same way.

Within our project, we just created a listening part using digitized semi-authentic dialogs and we plan to add a self-assessment questionnaire for the oral expression. At this point, we wished to analyze the strategies that are used by the students when they have to work with CAT. Seven students, ranging from beginner to very advanced levels, were asked to do the CAT version in an experiment using verbal protocols. While doing the test, the students were asked to comment about their strategies, the difficulty of the task and their comprehension of the input. Some of the findings of our preliminary content analysis raise some concerns about the appropriateness of an IRT model to design this type of test. First, it seems that most of the students do not realize that the items are selected according to the estimation of their actual proficiency level. In fact, they consider the tasks which are cognitively demanding (e.g. reading paragraph and listening passages) as more difficult even though these tasks were not assigned a higher difficulty parameter. Second, students, even at the beginner's level, do not guess but rather try to eliminate the most unlikely options. Therefore, one may wonder what the pseudo-guessing parameter really represents. Third, many students feel some frustration because they cannot go back and use information obtained during the test to modify their answers on previous items. If this strategy has been brought into play on the paper-and-pencil versions which were used for the calibration, then the assumption of local independence may not hold. The verbal protocols have also been very helpful in order to describe the way the students are processing the information to answer the three questions which are supposed to measure the comprehension of each audio passage on the new listening part.

As a result, we believe that the analysis of verbal protocols is an important step in the validation of this type of test. It allows us to verify some psychometric assumptions and to shed light on the processes in order to distinguish those that are related to the trait and those that are related to the method.
Testing Language Use: Forms and Functions

Avis Jones-Petlane, Montgomery College
Mary Owens, Montgomery College
Paul Lux, Montgomery College
Roseli Ejzenberg, Montgomery College
Helena Wong, Montgomery College

OBJECTIVE: This work is an attempt to use qualitative research from the field of Sociolinguistics (notions of linguistic "domains", register, stylistic variants, and speech act theory) to inform the development of a criterion-referenced test of ESL proficiency for vertical program articulation.

In 1980, Heidi Byrnes defined language learning program articulation as "well motivated and well designed sequencing and coordination of levels of instruction towards certain (well defined) goals." On the basis of predetermined performance criteria, we are currently engaged in the process of test construction and administration to determine threshold (entry) and exit levels for the Oral/Aural Track of the American English Language Program at the community college level. With an ESL program of over 1,500 students, characterized by great diversity in terms of country of origin (over 100 nationalities), educational levels and backgrounds (open admissions policy), and student educational objectives, we have the task of determining the components of a test that will "strike a balance" between analytic-form-based-constructs and "synthetic" use-based constructs to reliably and validly measure student comprehension and production. The target of the measure is "language", in its wider sense, and not simply "speech" - all of which implies dimensions of language-variation and its social, or communicative, significance. And, we find, following Lyons (1981), that this learner "communicative competence" is "...not simply a matter of vocabulary...it also affects grammar and, as far as the spoken language is concerned, pronunciation."

Thus, this work in progress attempts to do the following:

1. Identify explicitly the threshold prerequisites and outcome objectives (competencies and skills) of the sequenced course (i.e. use of embedded Qs for "politeness", use of Tag Qs to verify information, questioning necessary conditions).

2. Determine components (format) of test to meet the parameters of aforementioned objectives (i.e. Listening Comprehension, Language Use, Oral component).

3. Design tasks or "items" (within components) to validly and reliably measure the specific, targeted competencies (i.e. Interviews, Multiple Choice, Short Answer, Taped Monologues).

4. Evaluate the measure for (A) construct validity as well as face validity, (B) inter-rater reliability (on oral or written components), (C) item reliability, and (D) overall reliability.
Towards a Valid Assessment of Out-of-Class Language Use by Youths Ages 8-18 in an L2 Immersion Environment

Heidi E. Hamilton, Georgetown University, Concordia Language Villages

After reviewing the 30+ year history of the Concordia Language Villages in Minnesota and writing a long-range plan, Concordia College made research and assessment a priority for the future. Although the participation of 6000+ students (called 'villagers') each summer and long waiting lists may be seen as concrete measures of success, this immersion program for youths ages 8-18 in ten languages (German, French, Spanish, Russian, Chinese, Japanese, Norwegian, Swedish, Finnish, and Danish) had never undertaken systematic research to examine the reasons for that success.

One of the goals of the resulting research and assessment project is to understand how the villagers use language (both the target language and English) outside of the relatively formal small-group language instruction periods, which range from 2-4 hours/day depending on the type of program the villager has selected. Since villagers spend a substantial amount of time each day learning language through activities such as soccer, dancing, pottery, and canoeing, a valid assessment of both the quantity and quality of target language use in these activities would be a welcome supplement to the small-group teachers' assessment of villagers' in-class language performance.

The first phase of the project was conducted during the summer of 1993. Senior staff members carried out detailed ethnographic observations of language use as it related to activity structure and participant framework in a sampling of physically active (e.g. soccer) and less active (e.g. baking) activities. Based on these observations, field research during summer of 1994 focused on the use of questions and responses by both leaders and villagers in such activities. Videotapes were made of thirteen activities in the French, German, Spanish, and Japanese villages. Transcriptions of these videotaped activities are being examined for use of target language and English by both leaders and villagers. Questions and responses identified in the interactions are being coded according to a variety of formal and functional features. Subsequent quantitative analyses will correlate features of the questions and responses with the type of question/answer and the participant framework, as well as examining relationships between question features and response features. These correlations will help us identify communicative behavior by activity leaders that best creates a wide variety of language opportunities for villagers, and encourages them to initiate interactions and use as much target language as possible.

Patterns of language use identified by this research have three potential applications: 1) development of an instrument to assess the quality and quantity of language use by villagers and staff in relatively informal language situations; 2) improvement of the existing bottom-up curriculum; and 3) staff training. It is, of course, the first of these applications which has most direct relevance to the 17th Language Testing Research Symposium. Future research directions include: 1) investigation of the relationship of learner characteristics such as language level, age, other L2s, and motivation to learn to the use of the target language outside of formal instruction periods, and 2) comparison of individual students' language use in directed activities to such use in both more formal and less formal situations within the village. This research project has entered a crucial stage: we would value feedback from others in the field - and would welcome an opportunity to discuss our work in progress.
An Approach to Comparing Raters’ Scoring Criteria: Integrating Qualitative Data and Quantitative Analysis

Jeff Connor-Linton, Georgetown University

Quantitative analyses of raters’ evaluations of testtaker performances may mask significant qualitative differences in how those ratings were produced--i.e., different interpretations of scoring criteria. This paper describes an approach to comparing the scoring criteria of different groups of raters which directly elicits raters’ (qualitative) reasons for their ratings and enables these data to be analyzed quantitatively in order to assess differences among raters’ operationalization of scoring criteria (and therefore the reliability and validity of those scoring criteria). The approach seeks a balance between the number of raters from whom evaluations are elicited, the number of performances rated by each rater, and the amount, complexity, and sensitivity of the feedback elicited from each rater. The procedure includes: (1) elicitation of both ratings and reasons for ratings of a common set of performances from different groups of raters; (2) coding of reasons for ratings; and (3) quantitative analysis of various relations within and between groups’ ratings and reasons for ratings. Examples are drawn from a cross-cultural comparison of American ESL and Japanese EFL instructors’ evaluations of compositions written by adult L1 Japanese EFL students. Issues discussed include the reliability and validity of raters’ metalinguistic judgments, the independence of their reasons for ratings from the scoring criteria, potential uses in criterion-referenced scale development, and pedagogical implications of such studies.

Validation of a Video-Mediated Test of Second Language Listening Proficiency

Paul Gruba, University of Melbourne

Against a growing awareness of the implications of the widespread application of video and satellite broadcasts in the language classroom, test developers are increasingly challenged to provide for the usage of visual media in tests of second language listening proficiency. Investigations into the validity of such tests, already muddled by a range of non-linguistic factors, are further complicated by the use of video as a presentation medium. Supported by a visual context, concerns of bias, for example, arise from the inclusion of nonverbal information in relation to the candidates’ cultural backgrounds, gender, background knowledge and cognitive characteristics. Additionally, a realization that production factors including pace of edits, use of computer graphics, choice of actors and setting may affect content relevance is brought to the fore.

This presentation of a ‘work-in-progress’ in regards to the validation of video-mediated tests of listening proficiency seeks assistance on the following points: references to any known work in this area; recommendation of appropriate techniques in the conduct of validity studies, especially in the gathering of qualitative data; and, finally, suggestions of where to place the influence of nonverbal factors in models of communicative language ability. In short, how might the use of video as a mode of presentation influence the validity of a listening test?
Computer Assisted Adaptive Language Testing for Civil Servants in a Multilingual Environment (Belgium): The ATLAS Project

Jozef Colpaert, University of Antwerp

At the University of Antwerp, the Didascalia research centre is currently developing a system for adaptive language testing by order of the Belgian Civil Service Commission. This national commission is responsible for testing the language proficiency of more than 20,000 candidates per year in three official languages. The objective of these exams is not only to test linguistic skills, but also the ability to function in a specific set of communicative situations according to job profiles. By introducing a computer assisted approach, the BCSC wants to make existing tests faster, more precise and more objective.

The realisation of this project requires research\(^1\) at subsequent stages:

- specification of targeted language skills and contents
- analysis of test taker profiles: L1, L2, educational background and professional experience
- analysis of job profiles in terms of communicative situations and required language functions
- formatting of rich and varied contents: selection, structuring, encoding, tagging and linking
- calibration and ponderation of test items according to distinctive features
- user interfacing in terms of measurable interactivity, degrees of freedom and ergonomics
- psychometric and didactic strategies for exploiting textual, audio and video elements
- evaluation of answers and linguistic routines
- adaptive search functions
- reporting and decision making
- follow-up, analysis and readjusting of the system

In this presentation, we want to discuss our adaptive function in detail. This function calculates probabilities of all difficulty levels and determines the degree of adaptation according to the information value of the moment. The more uncertainty, the higher the adaptation factor.

Implementation of the first six modules of the ATLAS system is foreseen by December 1994.

---

(1) Research related to this paper is carried out under an IUAP-project financed by the Belgian State
Investigation of the Linguistic, Cognitive and Method Attributes Underlying Test Task Performance: A Pilot Analysis Using Rule Space Methodology

Gary Buck, Educational Testing Service
Kumi Tatsuoka, Educational Testing Service

The relationship between the content characteristics of test tasks (linguistic, cognitive, or method) and performance on those tasks has been the subject of much speculation and theoretical discussion and many hypotheses have been proposed (e.g. Munby, 1978; Richards, 1983; Carol, 1972; Bachman, 1990). However, it has proved very difficult to empirically validate such hypotheses. Multiple regression has been used (e.g. Freedle and Kostin, dat; Pollit and H), but this is only able to tell us how much of the variance in total test scores can be accounted for by a list of attributes; it cannot tell us what attributes are involved in performance on particular test tasks, nor what attributes have been mastered by particular students.

The Rule Space methodology (Tatsuoka, 1991, Tatsuoka and Tatsuoka, 1992, Sheehan, Tatsuoka and Lewis, 1993) has since been used successfully to investigate the attributes what account for performance on tests of mathematical reasoning, NAEP science, document literacy, and verbal reasoning. Basically the methodology requires ‘experts’ to generate hypotheses about what knowledge states, cognitive process, or other task attributes account for test task performance. From this a matrix of item by attributes is produced, and this is then compared to the matrix of item responses by means of Boolean algebra. Those attributes with predictive power are retained, and those without rejected. The process is then repeated through a number of iterative cycles.

It is proposed to conduct a pilot analysis to ascertain whether, and to what extent, the Rule Space methodology can provide an empirically validated analysis of the linguistic, cognitive, or test method attributes which account for performance on set of communicative second language test tasks. The basic procedure will be as to select a suitable communicative data set will be selected from a number of available candidates, the literature will be examined to produce a list of possible attributes, and a group of four applied linguists, with a strong interest in language testing, will be asked to code each test task on the presence or absence of each of the possible attributes. This will be tested using the Tatsuoka Rule Space methodology. The process will be repeated and the list of attributes modified through a number of iterations until the experts are unable to generate new hypotheses, or over 90% of the variance in item difficulty has been explained. We will then estimate the conditional probability of each attribute contributing to total score, and the probability of each test-taker successfully using each attribute.

Results will be presented, the suitability of the methodology for use with second language test data will be discussed, along with implications for language testing from both a theoretical and test design perspective.
The Effect of Planning Time in Second Language Test Discourse

Jill Wigglesworth, *University of Melbourne*

**Rationale**
The inclusion of planning time in semi-direct oral interaction tests adds considerably to the overall length of the test, and it is important to be clear that the increase in length is justified by the language outcome. Previous research has shown that the effect of planning time in second language can differentially influence the resultant discourse (Crookes 1989; Ellis 1987) with planned discourse eliciting more complex language on a range of measures. However, where planning time has been provided it has generally been a substantial amount of time (ten minutes or more), and in a second language classroom situation, rather than a testing situation. Where planning time is provided in an oral interaction test it is generally limited to one or two minutes.

**Methodology**
In order to investigate whether the provision of such planning time is long enough to have a significant effect on the discourse of the candidates, two versions of a tape-mediated test were developed. Tasks across the two tests were identical, but planning time was provided in only one version. The test was administered to 120 candidates, half in each condition.

**Results**
Firstly a quantitative investigation of differences on test scores on each of the criteria used to rate task performance was undertaken. Secondly, discourse analytic techniques were then used to determine the nature and/or significance of differences in the elicited discourse across the two conditions in terms of complexity, verb usage, self-repair, pauses and morphological complexity. Finally, a finer analysis was undertaken which investigated whether candidate proficiency was a factor in planning time.

**Implications**
The results are discussed in relation to the implications for test design.
Developing Computer Adaptive Tests: Issues and Concerns

Micheline Chalhoub-Deville, University of Minnesota

In 1984 the College of Liberal Arts (CLA) at the University of Minnesota initiated a project to re-examine the foreign language requirements. The principal purpose of the project was to investigate students' actual language proficiency as the requirement for entrance and graduation rather than time in class. Consequently, in 1986, the first and only system for testing language competence based on the ACTFL Proficiency Guidelines was established at the University of Minnesota for both entrance to and graduation from the College (Lange, 1994).

Except for the speaking modality which employs the oral proficiency interview, the entrance and graduation tests developed for listening, reading, and writing are paper and pencil tests. Since thousands of students are required to take these proficiency tests every year the university wants to bring these tests up to date and continue their development along the lines of a computer adaptive system (CAT). Although CAT packages are available, these packages are not proficiency-based, and do not utilize the audio-visual capabilities that enhance the authenticity of the tests and enable us to test learners' listening skill, a neglected area in CAT.

The present poster session proposes to (a) present the results of analyses performed to assess the validity and reliability of the current CLA entrance and exit tests; (b) discuss the strengths and weaknesses of CAT; and (c) address theoretical and developmental issues that need to be considered for the construction of computer adaptive proficiency testing.

Computer-Delivered Multimedia Test Development at DLIFLC

John L. D. Clark, Defense Language Institute Foreign Language Center
Dariush Hooshman, Defense Language Institute Foreign Language Center

Over the past three years, the Defense Language Institute Foreign Language Center (DLIFLC) has developed, administered, and is in the process of validating several microcomputer-delivered listening, reading, and speaking tests both for internal use and for use by other government agencies. This presentation will describe the process of development of the test administration engine (using Multimedia Toolbook); item preparation and banking; on-screen and audio item presentation, capture of examinee responses; and computer assistance in response evaluation and score reporting. Practical and psychometric advantages of computer-assisted development and administration of language skill tests, by comparison to traditional “paper-and-pencil” techniques, will be briefly discussed.
Task, Judge and Scale Effects in the Rating of Speaking Ability of Primary School ESL Learners

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

The research reported here examines the effects of judge, task, rating scale and school on measures of overall speaking ability of primary grade 6 learners of ESL. A large urban school district with primary schools located in a variety of socio-economic neighborhoods wanted a short test of overall speaking ability to equate standards across schools. Thirteen instructional personnel from the school district together with two outside researchers participated in the development, administration and scoring of tests.

Two speaking tests (elicitation tasks, administrative procedures, rating scales) were developed. One task was a story retell based upon a short segment from an animated video; the second task was a simulated audiotape "letter" in which students had relatively free choice in the information contents of their letters. Task appropriate administrative procedures were devised for the two tests. Rating scales were empirically developed for each task using samples of student performances. Qualitative differences in the two scales appeared to be attributable to differences in the elicitation tasks.

Approximately 300 students took one or both of the tests. All student performances were audio-recorded. Each student recording was rated independently by three of the participant teachers. The data set will be subjected to many-facet Rasch analysis (FACETS) to ascertain effects of school, task and judge severity. Results will be useful primarily for the guidance of the school district on questions of test selection, further test development and training of judges. Questions about comparability of schools within the district can be answered. From the research point of view, light may be shed on the effects of using task- and population-specific scales.
Construct Validation of the United Nations Association Test of English (UNATE) Level A Test and Its Implications for Further Development of the Test

Yasuyo Sawaki, University of Illinois at Urbana-Champaign

The United Nations Association Test of English (UNATE) Level A Test is an English proficiency test administered in Japan twice a year. The test is unique in that it is used not only as a proficiency test for learners of English in Japan in general but also as a selection test for screening Japanese trainees to the United Nations on the Associate Expert Program organized by the Japanese Ministry of Foreign Affairs. Accordingly, the test includes in part items designed to assess test taker’s performance in the UN context, while the rest of the test is designed to measure test taker’s English proficiency in more general context. The influence of the UNATE on EFL in Japan is becoming greater year by year, and the decisions made based on the results of the test is of paramount importance. However, virtually no construct validation studies of the test have been conducted to date.

Although the operational UNATE Level A Test is designed to measure various types of ability such as reading, structure, writing and speaking, this study focuses on investigating construct validity of the UN-based reading comprehension section of the UNATE Level A Test only. That being the purpose, a short reading comprehension test, READ-UN, was constructed out of four retired versions of the UNATE Level A Test; READ-UN consists of 20 multiple-choice items based on four separate reading passages on UN issues. Then, following the current support for triangulation of methods by researchers in test construct validation, two separate procedures are employed to investigate the validity of READ-UN: expert content analysis and experimental procedure. In the expert content analysis, the content analysis rating scales proposed in the LTRC 1991 presentation by Bachman, Davidson and Milanovic is utilized with some modification in order to capture the characteristics of the items in READ-UN. In addition, the experimental procedure designed to be a two-way ANOVA investigates whether the test score on READ-UN is the function of the test taker’s English proficiency and/or his/her background knowledge on the UN issues appeared on READ-UN.

The LTRC presentation is a report on the findings in this preliminary construct validation study of READ-UN and its implications for further development of the UN-based reading comprehension section of the UNATE Level A Test.
Psychometric Properties of Alternative Assessment

Elana Shohamy, Tel Aviv University
Claire Gordon, Tel Aviv University
Smadar Donitsa-Schmidt, Tel Aviv University
Ronit Waizer, Tel Aviv University

Alternative assessment has been the most recent development in the testing field in the past few years. Procedures such as portfolios, self-assessment, interviews and observations have been widely used to assess students' language development and language proficiency. Yet, in spite of their appeal in terms of face validity, content validity and washback effect there have been serious questions, and often criticism as to their various properties, i.e. different types of reliability and validity.

The purpose of this paper is to report on a study which investigated the psychometric properties of an alternative assessment battery which was developed for the purpose of assessing the language acquisition and proficiency of immigrant children in school context in grades 2-12.

The assessment battery included four alternative instruments: a) a language test; b) a portfolio; c) a student self-assessment questionnaire; and d) teacher observations. The language samples obtained via these procedures are then analyzed by the language teacher, the tester and the student in an assessment conference and the results are converted into diagnostic profiles which lead to pedagogical decisions and recommendations.

The research questions which will be examined in this paper focus on:
1. The extent to which similar or different information is obtained from the different types of instruments.
2. The type and quality of the information which is added to language proficiency by the different instruments.
3. The extent to which diagnostic profiles and pedagogical recommendations reflect the information obtained.
4. The type of information that is added to the understanding of the proficiency through the assessment conference.
5. The degree of agreement among different judges regarding the interpretations and synthesis of the language samples.

Answers to the above questions will be analyzed through qualitative and quantitative procedures such as:

a. correlations among the scores obtained from the different instruments.
b. prediction of achievement as estimated by the different procedures and examination of the marginal contribution of each instrument
c. rater reliability of the interpretations
d. qualitative and quantitative examination of the assessment conference based data which was videotaped and content analyzed.
e. the relationship between the interpretation obtained from the assessment conference and the data obtained from the instruments.

Conclusions of the above analyses will provide important information regarding the value of alternative assessment and their suitability for large scale second language assessment.
Validating A Measure Of Depth Of Vocabulary Knowledge

John Read, Victoria University of Wellington

For students undertaking academic study in a second language, adequate lexical ability is a fundamental requirement. Some recent studies (Wesche and Paribakht 1993, Read 1993, Verhallen and Schoonen 1993) have investigated various means of measuring one significant component of this ability: how well learners know high-frequency L2 vocabulary items. The present paper represents a further contribution to this area, which can also be referred to as the assessment of depth of knowledge of vocabulary.

The paper reports on two studies undertaken with learners of English for academic purposes in an intensive English programme at a university language institute. The research involved three vocabulary measures, all designed to operationalise the concept of depth of knowledge of polysemous words. The first was a word associates test, which required test-takers to select words that were semantically related to a given target word. A second written test involved the matching of words to defining contexts. The third measure was an interview to elicit explanations of the words, which were then rated on a modified version of Wesche and Paribakht's Vocabulary Knowledge Scale. In all three measures the target words were a set of 40 high-frequency English adjectives.

The analysis was carried out as an exercise in the validation of the word associates test as a measure of depth of knowledge of the words being tested. Following Messick (1989), it is now widely accepted that validation involves gathering various kinds of evidence to support the inferences that we wish to make on the basis of test scores. In this case, the primary concurrent evidence was provided by the results of the matching test and the interview. Further evidence was obtained from Rasch analysis, using the partial credit model, and from the verbal report data of individual test-takers working through the word associates items.

The results showed that there was a need to distinguish between depth of knowledge of the set of adjectives as a whole and how well individual adjectives were known. The evidence provided stronger support for the validity of the word associates test in the former sense than the latter. In addition, the analysis indicated the significance of a syntagmatic-paradigmatic distinction that had been built into the structure of the word associates items.

The paper will discuss the process of validation of this kind of test and consider what other data might further elucidate the meaning of the test results.
Validating Questionnaires Designed to Measure Test Takers' Selected Cognitive Background Characteristics

James E. Purpura, UCLA

Since the 1970's research and theory in second language education have shifted from examining the methods of teaching to investigating the processes of learning. This refocusing on the learner has created an explosion of research aimed at investigating learner characteristics and second language acquisition (e.g., O'Malley and Chamot, 1990; Wenden, 1991). A similar trend can be seen in language testing as more interest is expressed in investigating the cognitive characteristics of test takers such that language testing researchers have shown interest in the cognitive processes of language use in their efforts to (1) describe the nature of language proficiency (Spolsky, 1973; Oller, 1979; Bachman, 1990) and (2) investigate the factors other than communicative language ability that affect test performance. Despite this interest, the definition and operationalization of a cognitive construct in language test performance which is related to a coherent model of cognition (i.e., information processing), has yet to be articulated.

The purpose of this paper is to (1) describe the development of a questionnaire designed to measure selected cognitive characteristics of test takers and (2) to validate this construct by examining the items and their factorial relationships with respect to the underlying strategies and stages of information processing. In the first part of the paper, I will describe the theoretical constructs underlying the questionnaire design and will briefly mention the development process. I will then discuss the results of factor analyses in which the relationships between (1) questionnaire items and cognitive strategy scales and (2) between strategy scales and the underlying stages of the information process will be examined.
Why the Monkeys Passage Bombed: Tests, Genres, & Teaching

Bonny Norton Peirce, Ontario Institute for Studies in Education
Pippa Stein, University of the Witwatersrand

Struggles over the social meaning of texts are becoming increasingly vociferous in post-apartheid South Africa. When such texts are used for admissions purposes, they become implicated in debates over educational equity. This paper describes and analyzes one such struggle during the piloting of a reading test to be used for admission to the University of the Witwatersrand in Johannesburg, South Africa. The text, which the authors refer to as the "Monkeys Passage", was piloted in a graduating high school class of black students in Johannesburg in 1991. During the course of the piloting procedure and the subsequent classroom discussion, the students constructed two dominant readings of the text - one which positioned the text as a story about monkeys, and the other which positioned the text as a metaphor for inequitable relations of power between blacks and whites in South Africa.

The fundamental paradox the authors address in this paper is that the test-takers fared remarkably well on the test despite the fact that many of them objected strongly to the content of the text. Drawing on recent work in genre analysis, the authors argue that the two dominant readings of the text were constructed in the context of two different social occasions - albeit on the same day and in the same place, and that the shifting power relations between the teacher and the students on these two occasions were implicated in the construction of the two different readings.

During the first occasion, the test-taking event, the relations of power between the teacher/tester and the students were unequal. Under these conditions, what Kress calls the "mechanism of interaction" (the conventionalized form of the test event) determined to a great extent how the students read the text. The students faithfully reproduced the testmaker's reading of the text, and the multiple-choice format of the test may have played an influential role in promoting such a reading. On the second occasion, the more egalitarian class discussion of the text, the "substance of the interaction" (the content of the text) became more foregrounded than the mechanism of interaction. In this context, there was no longer a single, legitimate reading of the text and multiple readings were constructed. Students drew on their background knowledge and experience to critique and challenge the text, ultimately dismissing it as racist.

The authors discuss the challenges raised by this testing event for testers and teachers. The study suggests that the unequal power relations between testers and test-takers has a crucial bearing on the ways test-takers read texts. As an extension of this point, the authors highlight a fundamental validity paradox in some language tests that are used for admissions purposes: While admissions officers may wish their language tests to identify critical, independent learners, the testing instruments they use may not give test-takers the opportunity to demonstrate such abilities. Furthermore, students from minority backgrounds might feel particularly constrained to draw on their background and experience to engage with such texts. The authors recommend that closer ties be established between teachers, testers, test-takers, and test users: Teachers should be represented on test development teams and admissions committees; testers should research the washback effects of their tests; and admissions officers should investigate the validity of the instruments they use.
The Need for Assessing Different Speech Interactions by Using Different Rating Scales

Pavlos Pavlou, Georgetown University

Recent studies in oral proficiency testing (Bachman, Lantolf & Frawley, 1985; Savignon, 1986; Van Lier, 1989, etc.) have indicated problems with the oral interview, the most common technique in testing oral proficiency. Research has shown that this technique may not be the only or best way to test oral proficiency since it is insufficient to detect many aspects of the construct.

Certain theoretical models for testing oral proficiency that have been proposed to deal with the above mentioned problem show the necessity and usefulness of assessing oral proficiency through various oral genres (Shohamy et al. 1986, Shohamy 1988). Group discussions, oral reports, oral interviews, and role plays are some of the proposed speech interactions. Through these interactions it can be ascertained whether the test taker is sensitive to contextual variables (such as role relationship, setting, formality level, topic, function, etc.) that interact in communicative oral language.

A test battery for oral proficiency was developed according to guidelines set by Shohamy et al. (1986) and Shohamy (1988) in order to provide evidence for or against the hypothesis that different speech interactions measure different aspects of oral proficiency. The test was administrated with 60 EFL high school students in Cyprus.

The results show that the four assumed different speech interactions were not perceived or scored as different. This can be attributed to many factors. However, one of the most important reasons was that the four speech interactions were judged against the same scale (an adaptation of the Bachman & Palmer, 1983 scale).

However, in order for a subjects' performance to be judged as different, it should be assessed against scales which are sensitive to the features that make this specific interaction different from the others. This is only possible if the differences among the various speech interactions are defined independently of possible content or specific topics addressed in a real test situation.

This study provides a model for analyzing and comparing the various speech interactions. Also, the study provides suggestions for the development of different rating scales based on this comparison. The scales consist of features of grammatical, textual, and sociolinguistic competence and can be used to assess a subject's performance in different speech interactions.
ITA Testing: Validity or Equity?

Carol Lynn Moder, Oklahoma State University
Gene B. Halleck, Oklahoma State University

Recent trends in communicative language teaching have led to an increasing awareness that for a test to validly assess a student's performance it must sample language in authentic discourse context (Shohamy & Reves 1985, Bachman 1990). This concern for authenticity has led those designing tests for International Teaching Assistants (ITAs) to turn primarily toward the use of teaching tests (Hoekje & Linnell 1994). While the use of tasks that emulate authentic contexts will increase the likelihood that the teaching tests will have construct validity, it leaves open two key questions: the predictive validity of the tests and the extent to which language knowledge and strategic knowledge interact in the test (Bachman 1991).

The main purpose of this study was to investigate one factor which may contribute to the predictive validity of the test: the extent to which the evaluations made by the faculty raters conform to the perceptions of undergraduate students. The ratings of undergraduates are particularly relevant to these screening tests, since such tests have often been implemented as a response to student concerns and since such undergraduates are the potential students of the ITAs. Performances of fifteen international graduate students on the ITA Test (Smith, Meyers, & Burkhalter 1992) were rated as part of a regular screening test by two trained faculty raters. The faculty raters scored students on language skills, teaching skills, and the evaluator's overall impression. Videotapes of these performances were shown to 114 undergraduate students. The students also rated each TA on language skills, teaching skills, and their overall evaluation of the ITA's classroom readiness. In addition 28 students provided written comments explaining the criteria that influenced their ratings.

The results of the study showed that in general the undergraduate students were considerably more severe in their overall evaluations of the ITAs. They were much more likely to recommend training for all TAs that the faculty evaluators were. Further, the relationship between the student evaluators' ratings of language skills and teaching skills and their overall evaluation of the ITAs varied according to the overall language proficiency of the ITA. For less linguistically proficient ITAs, there was a strong relationship between the students' evaluation of the ITAs' language skills and their overall ratings. On the other hand, for more proficient ITAs the students' evaluations were more dependent on their perception of the ITAs' teaching skills. Thus, the students were much less likely than the faculty raters to evaluate any of the ITAs as classroom ready. The fact that the student evaluators held highly proficient ITAs to very stringent teaching standards, raises an important equity issue for language testers. The faculty raters were more consistent in the extent to which they considered language and strategic skills to be important in the overall assessment of the test performance, evaluating the teaching skills only as strategies that could help the ITAs compensate for language deficits. However, the student evaluators clearly did not share this view. For them, native-like language proficiency alone was not adequate; they also wanted excellent teaching skills. This creates a dilemma for the language tester. Do we accept the ratings of the student evaluators and require additional training for most ITAs or do we continue to consider teaching skills only as compensatory strategies? If we do the former we will be holding ITAs to a different standard from that used for native speakers. If we do the latter we may jeopardize the validity of our tests as screening measures.
An Investigation of the Validity of the Demands of Tasks on Performance-based Tests of Oral Proficiency

Dorry Kenyon, Center for Applied Linguistics

Theoretical background and rationale
Do 'naive' foreign language students perceive the difficulty of performing various speaking tasks in a manner consistent with the hierarchical characterizations contained in the Speaking Proficiency Guidelines of the American Council for the Teaching of Foreign Languages (ACTFL)? An affirmative answer to this question would provide external support for the validity of oral proficiency tests based on an assumption that more foreign language proficiency is required to carry out certain speaking tasks than others. Performance assessments such as the Oral Proficiency Interview, the Simulated Oral Proficiency Interview, and the Video Oral Communication Instrument are based on this assumption.

Purposes of the research
This study is connected with the validation of new tape-mediated performance-based tests of oral proficiency in Spanish, French and German. For this research, a new task-based self-assessment instrument was developed that presents foreign language students with a short description of 18 speaking tasks (16 of which appear on the new tests). Using a six-point scale, students indicate how well they feel they can complete each speaking task.

Research design and methods
This paper reports on a pilot and a final administration of the self-assessment instrument. The pilot study involved 90 students at different levels of high school and college Spanish study. Over 300 students of Spanish, French and German in high school and college were involved in the final study. Each student completed the self-assessment instrument before taking the new speaking test. Using the data from each study, the 18 speaking tasks were scaled using Rasch methodology. The empirical scaling of the speaking tasks was compared to their a priori scaling by the test developers based on the hierarchical characterizations contained in the Guidelines.

Results
The data from both studies fit the Rasch model remarkably well. A perfect correlation was found between the scaling of the speaking tasks by the students and their a priori assignments. With minor variations, the scaling of the 18 tasks across languages was very similar.

Implications
The findings from this research lend additional support to both the validity of the hierarchical characterizations presented in the ACTFL Guidelines and to performance-based tests based on that hierarchy. The paper concludes with some implications for computer-adaptive oral proficiency testing and for tailored scoring of performance assessments.
Perspectives on Validity: A Historical Analysis of the LTRC

Liz Hamp-Lyons, University of Colorado at Denver
Brian Lynch, University of Melbourne

Over the past five years there has been much discussion in the educational measurement literature concerning new perspectives on reliability and validity (Cronbach 1989; Linn, Baker, & Dunbar 1991; Messick 1994; Wolf, Bixby, Glenn, & Gardener 1991). In particular, there have been discussions which suggest approaches to language testing research that go beyond the traditional, psychometric approach. Moss (1994) argues for a hermeneutic approach to validity which, essentially, questions the "necessary but insufficient" role of reliability and points out its potential for constraining innovation in educational practice. Bachman and Palmer (forthcoming) argue for the concept of test usefulness, in which reliability and validity are interdependent with other qualities such as authenticity and impact, the nature and interplay of which must be determined for each specific testing situation. It is common to see works and arguments such as these as representing potentially dramatic, major shifts in language testing research, and as suggesting that the field's stance in relation to innovative movements in educational assessment, as well as issues of equity, has been deeply conservative.

If such conservatism does indeed reflect the stance in language testing to date, we can expect there to be great difficulty in moving toward innovation and change in our field. However, such a picture may derive primarily from a kind of folklore and anecdotal wisdom about the research paradigms that language testers have valued up to now -- that is, it does not have an empirical base. Searching for an empirical data set to investigate our own community's stance toward these questions, we have chosen to turn inward and reflect upon ourselves: This study examines the abstracts and, where available, the texts of papers presented at the LTRC since its beginning, with the help of LTRC archivist Fred Davidson, to investigate whether language testing research has, in fact, been dominated by the psychometric approach, or whether perhaps the seeds of the previously-mentioned new perspectives on reliability and validity have in fact been taking root in our community, more or less quietly, over the past few years. The analysis focuses on the ways in which reliability and validity have been addressed, both implicitly in the methods employed and explicitly in theoretical argumentation. We will attempt to articulate the underlying paradigmatic assumptions that seem to guide the researchers, making use of the analysis of Guba (1990), and to categorize the LTRC papers into a database that will include the descriptions of language and research methods used. The results of this investigation will document the extent to which the LTRC community has already engaged itself with modes inquiry beyond the psychometric and has already begun to formulate new perspectives on reliability and validity. It is hoped that this historical analysis will suggest fruitful ways of recasting the concept of reliability within a socially-motivated validity capable of responding to issues of equity as well as innovation in education.
The Use of Questionnaire Feedback in the Development and Validation of an Oral Interaction Test in Two Formats

Kathryn Hill, University of Melbourne

As part of the test development process it is necessary for test designers to investigate both the reliability of the testing instrument and the validity of the interpretation and use of test scores.

This paper examines feedback from examinees, raters and interlocutors in relation to a test of oral English proficiency for prospective migrants. The test was developed in two formats, direct and semi-direct, which are used interchangeably. In this paper the relationship of feedback to the questions of fairness and validity (specifically face, content and construct validity) as well as to test revisions, is examined. In particular, the paper looks at the contribution of feedback to investigations of the comparability of the direct and semi-direct formats.

Questionnaires, seeking reactions to specific aspects of the trial version of the test, were administered to 94 trial candidates, 13 raters and 12 interlocutors. All trial candidates had attempted both test formats. The data collected were both qualitative and quantitative.

Examination of feedback from test takers related their reactions to the two formats to test performance and test-taker characteristics. Feedback from all respondents was also instrumental in improving test reliability by suggesting revisions to test method facets as well as to aspects of test content. Finally, feedback was found to provide evidence regarding specific aspects of test validity.

Implications for the coexistence of the two formats can be drawn from the differences in attitudes by test-takers towards the two formats. The findings also have particular implications for the selection of trial populations, as well as highlighting the tension between improving test reliability and maintaining test validity.
Testing Bilingual Teachers' Language Proficiency: The Case of Arizona

Leslie Grant, Educational Testing Service

This presentation will report on the language assessment of potential and practicing bilingual teachers in the U.S. Presently, of the twenty-nine states that offer bilingual certification/endorsement, only eighteen require formal assessment of language skills. This paper will first review current teacher testing practices for assessing language skills and will then focus on a selected measure used for bilingual teacher endorsement in Arizona. Upon completing this session, participants will have a clear idea of the advantages and disadvantages of this type of assessment for teachers' language abilities; they will leave knowing what constitutes a well-rationalized, defensible instrument for assessing teachers' language proficiency.

Taking the model of Communicative Language Ability (CLA) (Bachman, 1990) as a starting point, it is evident that the interaction of language competencies, strategic competencies, and context is extremely complex and especially challenging where the assessment of teachers is concerned. Bilingual educators find themselves in a variety of situations requiring differing language competencies and strategies for accomplishing their goals. They need to give directions, present lessons, interact with students and parents, and so on. The context of the bilingual classroom then, presents a unique set of language demands as compared to a foreign language classroom, or a monolingual classroom, for example. How best to assess if an individual meets those demands then, becomes a challenging question.

The certification procedures throughout the U.S. range from paper and pencil tests to simulations of classroom situations. Recently however, the drive for performance-based tests has placed more emphasis on trying to test individuals in more "authentic" ways, focusing on what they can do with the language as opposed to what they know about the language. These tests require that the examinees engage in tasks requiring language use in situations which are close to real life (Wesche, 1992).

One such effort at authentic assessment of teachers' language proficiency is a criterion-referenced, performance type of test used for bilingual teacher endorsement in Arizona. The Arizona Classroom Teacher Spanish Proficiency Exam, or the ACTSPE, was developed in 1986 to ensure that potential bilingual teachers possess the requisite language skills for the classroom. This test was developed through the collaboration of the Arizona State Department of Education and the three major Arizona universities (Arizona State University, Northern Arizona University, and the University of Arizona). The main objective of the ACTSPE is the evaluation of the examinee's ability to use Spanish in the bilingual classroom. As a result, the test items are based on the Spanish that is used in the daily activities of the teacher: realistic tasks such as correcting students' writing, translating a letter home to parents, and teaching a lesson are used to measure the examinee's level of teaching proficiency in Spanish.

In this presentation, the ACTSPE will be reviewed as an example of partially performance-based language assessment for teachers which reflects current notions of communicative competence. The issues dealing with the reliability and validity of this measure will be discussed as will recommendations for improving teachers' language assessment in the future.
Language Background, Ethnicity, and the Internal Construct Validity of the Advanced Placement Spanish Language Examination

April Ginther, Educational Testing Service
Joseph Stevens, University of New Mexico

Although native language proficiency often serves as an idealized criterion in discussions of non-native language proficiency, explicit empirical comparisons of the performance of native and non-native speakers are infrequent. Analysis of the Advanced Placement Spanish Language Examination offers an opportunity to compare the performance of groups who differ in native language background and ethnicity. Such comparisons can inform our understanding of the development of language proficiency and ensure that the inferences drawn from test performance are qualified if and when systematic differences in performance are present.

This study examined the internal construct validity of the 1989 Advanced Placement Spanish Language Examination for five groups: (1) Latin students who identified Spanish as their best language, (2) Mexican students who identified Spanish as their best language, (3) Mexican students who identified both Spanish and English as their best languages, (4) White students who identified English as their best language, and (5) Black students who identified English as their best language. Test development documentation and program descriptive materials were used to specify an a priori model of examination structure with respect to 90 multiple-choice items and four constructed response sections. A traditional four-factor (Listening, Reading, Writing, Speaking) was hypothesized. Goodness of fit of the four-factor model to the data for the Latin-Spanish group was tested with Confirmatory Factor Analysis using LISREL 7. A null model was also applied to provide a baseline for model comparisons using the Tucker-Lewis Index.

The four-factor model provided a good fit to the data for the Latin-Spanish group, \( \chi^2(59) = 71.89, p = .12 \). Furthermore, the observed adjusted goodness of fit index (AGFI) of .97 approached unity, and the Tucker-Lewis Index of .99 indicated that relative to the null, the four-factor model fit the data well. With the Latin-Spanish group serving as the reference in four series of comparisons with the other groups, a nested hierarchy of increasingly restrictive tests of group invariance was conducted. This hierarchy tested the equivalence of factor structures, factor loadings, factor variances, factor covariances, and variable uniquenesses. Comparisons of model structure showed no difference with respect to number of factors; however, significant differences were found at all subsequent levels of hierarchical tests: factor loadings, variances, covariances, and uniquenesses. These results show a progression of increasing differences in examination structure from the Mexican-Spanish to the Mexican-bilingual to the White- and Black-English groups. Both dialectal variation and proficiency appear to play explanatory roles. Noteworthy are the weaker relations among factors and stronger factor loadings for the Latin-Spanish group in comparison to the bilingual Spanish-English and the White-English and Black-English groups. This suggests that a distinguishing characteristic of native proficiency is greater independence and definition of individual factors. This result, along with others, will be discussed.
Prediction of Item Difficulty in the English Section of the Israeli Psychometric Entrance Test

Ruth Fortus, National Institute for Testing and Evaluation, Jerusalem
Rikki Coriat, National Institute for Testing and Evaluation, Jerusalem
Susan Fund, National Institute for Testing and Evaluation, Jerusalem

The aim of the study was to gain a greater understanding of the factors that influence the level of difficulty of multiple-choice items appearing in the English Section (ES) of the Israeli Psychometric Entrance Test (PET). The score on the ES is a component of the total PET score, which is used as an estimate of future success in academic studies for selection of university candidates, and is also used for placement of students in remedial English classes. Understanding these factors would allow for a more accurate design of the item pool, improve the test’s quality, and clarify the difficulty-related factors which most affect a non-native English-speaking population. The study was conducted in four stages.

Stage 1: At this stage of the study, six expert raters were asked to estimate the difficulty of 155 items. Some studies (Bejar, 1981) have shown that raters have poor ability in accurately predicting multiple-choice item difficulty for non-quantitative items. This study found that while the raters were able to predict the difficulty of some types of items – Sentence Completions (SC) and Restatements (RS) – fairly well (correlations with item difficulty of 0.57 and 0.64, respectively), they were unable to make accurate predictions for Reading Comprehension (RC) items (0.24). These results prompted a closer look at (1) RC texts alone (see Stage 2) and (2) RC items related to these texts (see Stage 3).

Stage 2: Analysis of 24 RC texts based on structural, contextual and syntactical variables (based in part on work done by Freedle & Kostin, 1991) showed that the two factors which were most influential in determining the level of text difficulty – defined as the average difficulty of all items related to a particular text – were level of vocabulary and level of grammatical complexity. Both factors had a significant correlation (0.86) with text difficulty. In addition, it was found that raters (3 experts) had a high ability to predict the overall difficulty of a text (a correlation of 0.90 with text difficulty).

Stage 3: Analysis of 229 RC items showed that, in addition to the two aforementioned “text” factors, the variables with a significant correlation with item difficulty were amount of processing (0.34), level of vocabulary in the stem (0.22) and in the distractors (0.42), and distractor length (0.42). These factors differ from those Freedle & Kostin and other researchers found affected item difficulty, probably due to the fact that this study’s items were answered by examinees taking a test in a foreign, not their native, language.

Stage 4: In order to check whether the information gained from these investigations had improved the raters’ ability to judge item difficulty, 6 expert raters were advised of the results of the previous three stages and asked to predict the difficulty of an additional set of 155 items. It was found that correlations of average rater estimations of difficulty with actual item difficulties had increased for all types of items: from 0.57 to 0.72 for SC; 0.64 to 0.69 for RS; and, most dramatically, 0.24 to 0.82 for RC items. It seems that the process of analyzing texts and items contributes to a greater understanding of the factors affecting item difficulty. Moreover, once raters are made aware of the existence of these factors and their effects on item difficulty, they can implement this knowledge in an effective manner.
How Does Possible Washback Effect Work?

Li Ying Cheng, University of Hong Kong

There is some evidence to suggest that tests have a washback effect on teaching and learning (Alderson and Wall, 1993). The extensive use of test scores for various educational and social purposes in society nowadays has made the effect of washback a significant phenomenon. This paper presents research findings on the washback effect of the Hong Kong Certificate of Education Examination in English by employing various methodologies such as questionnaire, interview and classroom observation, which are based on an in-depth case study approach to sample schools in Hong Kong. It further discusses the nature of washback effect, the major teaching and learning factors influenced by it, the different stages of washback effect and types of washback effects observed. Preliminary results indicate that washback effect works quickly and efficiently in bringing about changes in the contents of teaching which is due largely to the commercial characteristics of our society, and slowly and reluctantly and with difficulties in the methodology teachers employ. The latter effect may be caused by teacher training and the constraints of teaching in our present schools.

The Effect of Language Proficiency and Background Knowledge on EAP Students’ Reading Comprehension

Caroline Clapham, Lancaster University

This paper reports on the final stage of an investigation into the ESP claim that tertiary level ESL students should be given reading proficiency tests in their own academic subject areas. A previous report described how reading passages in the International English Language Testing System (IELTS) test, which has modules in three academic subject areas, varied in their subject specificity: some were suitable for students in the relevant academic field, some were either too general or too specific.

The present paper compares the relative contributions of background knowledge and language proficiency to the students’ reading test scores, and investigates the extent to which the effect of background knowledge on comprehension varies according to the reader’s level of language proficiency.

The results of multiple regression analyses showed that language proficiency accounted for more of the test score variance than did background knowledge, but that the comparative effects of the variables differed according to the subject specificity of the tests: when the analysis was based on students’ scores on the complete reading modules, language proficiency accounted for almost half the variance, but when the students’ scores were based on revised tests from which the ‘non-specific’ subtests had been removed, language proficiency had less effect, and background knowledge proportionately more. This suggests that the comparative importance of background knowledge and level of language ability in reading comprehension depends on the specificity of the reading passages.

A study (using analysis of variance) of the effect of background knowledge on the reading performance of students at different levels of language proficiency showed that the reading scores of low proficiency students were not affected by background knowledge, but that the scores of students at higher levels of proficiency were. There was no steady increase in the effect as the students’ ability levels rose; instead there appeared to be a level of proficiency below which students were unable to make use of their background knowledge. This supports Clarke’s (1980) hypothesis that there is a threshold level of language proficiency that must be reached before students can make use of their top-down reading processes.

The paper emphasises the effect of test specificity on these results, and discusses the implications for EAP testing.
Research-Then-Theory to Test Development

Micheline Chalhoub-Deville, *University of Minnesota*

**Rationale**

Alderson (1981) maintains:

> the advantage of testing is that it forces explicitness: the test is an operationalisation of one's theory of language, language use and language learning...if we cannot get the tests our theories seem to require, then we have probably not got our theories right... Why has there apparently been such a failure to develop tests consistent with theories of communicative language use? (p. 54)

In response to this question, the issues that need to be considered are the diversity and applicability of proficiency models.

**Objectives**

The present paper proposes to provide an in-depth review of prominent models of proficiency advanced by researchers to reflect the nature of second language (L2) proficiency, and make a case for the construction of empirically-based proficiency frameworks that are to be utilized in test development and interpretation.

**Procedures**

The review of L2 proficiency models is organized utilizing Stern’s (1983) classification, thus, differentiating between proficiency interpreted as rating scales, such as ACTFL, and proficiency interpreted as components. The componental models are classified as a progression from the more abstract, including the Unitary Trait Hypothesis and CALP/BICS, followed by Communicative Competence, and Communicative Language Ability to the more concrete, for e.g., the Threshold Level Inventories. Additionally, studies aimed at deriving a framework for language testing purposes (e.g., Hinofotis, Bailey, & Stern, 1981) are included for review.

**Findings**

Based on the review of the above models, the following conclusions can be drawn: (a) some models have been too abstract and generic or too elaborate and comprehensive to be of practical value; (b) other models have been theoretically- or experientially-based and their empirical validity has not been established; (c) yet, other models have been based on empirical findings that either misused statistical techniques or have used techniques that do not enable the derivation of the components and their weighting; and finally (d) numerous researchers favor an empirically-based framework of testing.

**Conclusions**

Given the above findings, test researchers are encouraged to derive proficiency frameworks that are amenable to test development and would enhance the interpretation of results.
Ethnography and Testing: A Case Study of a Test Development Partnership in a Language for Business Purposes Program

James F. Valentine, Jr., UCLA
Shoichi Gregory Kamei, UCLA

The purpose of this study is to examine how qualitative research, especially ethnography, informs the design and development of a test of oral proficiency for a university international business program. Researchers from different disciplines and research backgrounds used the Bachman-Palmer Test Design and Development Framework (Bachman & Palmer, 1994) to evaluate and redesign an existing oral interview test. A qualitative analysis of the overall context of this test enabled test developers to increase the qualities of test usefulness, especially situational authenticity.

Ethnography, a form of qualitative methodology traditionally practiced by cultural anthropologists, emphasizes the context or "culture" in which a phenomenon occurs using participant observation, interviews and analysis of relevant documents. After examining the culture of an international business program and obtaining qualitative feedback from program participants, program evaluators formed a partnership with test developers to assess the propriety of an existing oral interview test and design a new oral interview assessment instrument.

That instrument is the Test of Oral Proficiency—Interview via Telephone which was developed using the Bachman-Palmer framework. It is designed to assess the oral language ability of MBA students in one of three language "tracks" (French, Japanese, and Spanish) of a combined intensive language and overseas internship program. This oral interview test is used for the purposes of program admission, measurement of proficiency across time, and diagnosis to inform instruction.

During the process of test design and development, the test development team found that ethnographic data was most informative at the operationalization level. Specifically, ethnographic evaluation uncovered three emergent concerns with the existing test: 1) the quantity and authenticity of business content in the test protocol, 2) usefulness of the test to measure participants' proficiency/ progress across the length of the program, 3) and the test's capacity to provide meaningful feedback to program participants. This integration of ethnographic findings within the test development process resulted in an increase of the situational authenticity of the test without a corresponding loss of validity or reliability. The resulting test, thus, evolved from one that tested language only to one that is truly embedded in an international business context.

The partnership between testing specialists and ethnographers enhanced the overall usefulness of the test within the specific institutional context and needs of this program. The inclusion of ethnographic information helps ensure that the test reflects the context of authentic language use situations. In this regard, the Bachman-Palmer framework was especially useful in providing a common ground for integrating ethnographic research into the process of test development.
Language Background and Item Difficulty: The Development of a Computer-Adaptive Test of Japanese

Annie Brown, University of Melbourne
Noriko Iwashita, University of Melbourne

The use of IRT analysis has greatly facilitated the development of computer adaptive tests, where the adaptiveness is based on measures of item difficulty resulting from the performance of trial candidates. However, studies into the acquisition of L2 grammar by learners with different L1s indicate that the learners’ L1 strongly influences their acquisition of grammar in the L2 (Lado, 1957; Rutherford and Sharwood-Smith, 1985; Zobl, 1980, 1982). Thus, it would be expected that grammar test items would present different levels of difficulty to candidates from different language backgrounds. Where a computer-adaptive grammar test is to be used with candidates from a range of language backgrounds it is, therefore, questionable whether set item difficulty measures can validly be used for all types of candidate.

The study investigates the performance of learners of Japanese from different language backgrounds, using data from a computer adaptive grammar test developed as a placement tool. The trial pen-and-paper test consisted of 225 multiple choice items. 1600 students in Australia, China and Japan and Korea (all of whom had studied Japanese for between 150 and 500 hours) each completed 50 items. In this study, data is presented from native speakers of English, Chinese and Korean. Item difficulties drawn from the trialing were found to be quite different for the three groups of candidates. This has implications for the validity of use of computer-adaptive tests, in that where actual candidates are from a different background from that of the trial population a) not only does the test fail to measure such candidates efficiently (in terms both of numbers of students misfitting and numbers of items required to be completed before ability levels can be calculated), but b) the measures of ability provided for each candidate and their relative rankings differ according to the set of item difficulties used, and will consequently affect decisions made about individual learners regarding placement or selection.
A Qualitative Analysis of Factors Affecting Learners' Performances in Group Oral Tests

Vivien Berry, The University of Hong Kong

Theoretical background and rationale
In an end of course test designed to measure the ability of first year undergraduates to participate in small-group tutorials, the input stimulus consists of a written short text which students are required to read and take notes on prior to engaging in an academic discussion. There is a great deal of research which suggests that within certain constraints, the topic of a text is relatively inconsequential as a source of language test score variance. Conversely, recent research into differences in methods of testing oral proficiency indicates that caution should be applied in the interpretation of scores awarded in the assessment of learners' performance in group discussions since the stability of the scores is susceptible to various identified extraneous factors, amongst them certain aspects of individual learner characteristics. There is, however, little published research relating differences in individual learning styles to students' subjective interpretations of written texts (as manifested in their notes) and subsequently to their oral contributions in a small-group discussion.

Purpose of the research
The purpose of the research is to investigate the way in which learner preferences in approaches to learning (as measured by Biggs' Study Process Questionnaire) affect the salience of points noted in writing and their ensuing development in small-group discussions.

Research design and methodology
Group interactions between eight groups of students (two topics, four groups each topic, five students per group) were video-taped and subsequently transcribed. Although quantitative data will be presented very briefly, the major focus of this paper will be on the qualitative aspects of the study. Students' individual notes were compared with sets of points from the input source texts identified as important by language tutors. These were then cross-referenced with individual contributions to the discussions to identify both references made to the highlighted points and relevant, original ideas based on points of departure from them.

Results and implications
Results provide insights into how individual differences in preferred learning styles can affect the quality of performance in group discussions and offer important practical implications for the assessment of seminar skills.
The Contribution of Language Testing to Validity and Equity of Assessment in a Medical Context: A Case Study

Rosemary Baker. The University of Queensland

This paper reports a case study in which language test items administered in a person’s first language contributed greatly to validity and equity in the assessment of her cognitive functioning in a second language hospital environment.

Background
The assessment of cognitive functioning in people with possible dementia can be problematic in a second language environment. The commonly-used dementia assessment tools depend not only upon proficiency in the language of administration, but also on culture-bound knowledge and levels of education and literacy: the administration of such tests in the first language by interpreters is therefore not a satisfactory solution.

The importance of language in dementia assessment is evidenced by its inclusion in key diagnostic criteria for probable dementia of the Alzheimer type, and by the fact that other cognitive functions such as memory are frequently tested through language. Further, research on language decline in Alzheimer’s dementia has indicated a number of language-based tasks that can be sensitive to the presence and severity of dementing illness.

Case details
The present study concerns a case of suspected dementia in an elderly Japanese subject in an English-speaking hospital. The subject was reported by friends to have (or have had) a good command of English: her lack of communication with medical and nursing staff therefore gave rise to concerns regarding her cognitive functioning. The case was further complicated by the subject’s history of schizophrenia. Investigation of her linguistic performance in Japanese appeared to offer a source of potentially valuable information concerning her mental status.

Procedures
A set of language-based tasks was selected on the basis of current knowledge on language decline and deficit associated with Alzheimer’s dementia. These tasks were administered to the subject in Japanese. They included naming, story recall, and processing by semantic category.

Results
Contrary to appearance on the basis of the subject’s inability or unwillingness to respond to hospital staff in English, the results of the Japanese language testing did not indicate the subject’s linguistic processing to be compromised by dementia, either in terms of speed or quality of response.

Implications
For the individual concerned, the language test data provided important diagnostic information which was central to decisions regarding her subsequent accommodation and care. More generally, this study highlights the need for language background to be taken into account in the assessment of cognitive decline, and demonstrates the potential of simple language-based tests administered in the first language for use in dementia assessment in second language settings.
MONDAY, MARCH 27

4:30 - 5:15  Closing Session  International 2

Summary of Conference Themes

Moderator
Adrian Palmer, University of Utah

Panelists
Charles Stansfield, Second Language Testing, Inc.
Carol Chapelle, Iowa State University
Dan Eignor, Educational Testing Service

Closing Remarks

Adrian Palmer, University of Utah
Lyle Bachman, UCLA/The Chinese University of Hong Kong

7:00  Banquet & TOEFL Dissertation Award Presentation

Aboard the Queen Mary
Entertainment Produced and Directed by Ted Rodgers
Test Usefulness: Principles and Considerations in Designing Language Tests

Lyle F. Bachman, UCLA/The Chinese University of Hong Kong
Adrian Palmer, University of Utah

The results of language tests are used for making inferences, predictions, or decisions that affect the lives of millions of individuals annually. In all situations where tests are used, the fundamental questions that needs to be asked is, "How useful is the information provided by the test for its intended purpose?" This question may seem so obvious that it need not be asked, but what does make a test useful? How do we know if a test will be useful either before we use it or after we have used it? Stating the questions of usefulness this way implies that simply using a test does not make it useful. It also points out that although usefulness is of unquestioned importance, it has not been defined in precise enough terms to provide a basis either for designing and developing a test or for determining its usefulness after it has been developed.

In this presentation, we will describe an approach to defining test usefulness that we believe makes this notion directly applicable to language testing practice—the design, development and use of language tests. Usefulness will be described in terms of a set of six qualities—reliability, validity, authenticity, interactiveness, impact, and practicality—that have generally been treated as more or less independent qualities of tests. In this new approach, it is the overall usefulness of a test that is to be optimized, and this must be done by considering all of these qualities together. The application of this approach to test development will be illustrated with an example of a specific test development project.
Operationalizing Content Validity: A Process Approach

Frances A. Butler, University of California, Los Angeles
Sara Cushing Weigle, University of California, Los Angeles
Andrea B. Kahn, University of California, Los Angeles
Edynn Y. Sato, University of California, Los Angeles

In recent years many language programs have revised their curricula to reflect a communicative approach to language teaching, necessitating the development of appropriate content-valid assessments for communicative curricula. Ideally, curriculum design should include considerations of assessment from the very beginning, so that learning objectives and assessment tools are developed together. Unfortunately, however, assessment issues such as how students will be placed into course levels, how their progress will be monitored, and whether exit criteria have been met are often not addressed until after the curriculum has been developed. In such instances, program administrators are faced with the difficulty of selecting or developing assessments that match the existing curriculum content. The process of going from teaching objectives and curricular content to valid and reliable assessments is difficult and time-consuming, and there are few clear guidelines available to help test developers ensure content validity. This is particularly true for communicative curricula, where objectives are frequently written in terms of proficiency descriptions rather than discrete language forms and structures to be learned.

This paper outlines an iterative process for developing potential item types for placing students into a communicative curriculum, which begins by abstracting from the curriculum the critical abilities to be tested within each skill area and formalizing this information in the form of a test construction guide. To ensure content validity, this process includes: extensive content review by experts in the field, small-scale pretesting of items, feedback from test takers and administrators, and large-scale pilot testing of revised items. Examples are drawn from a project involving the recently developed English-as-a-Second Language Model Standards for Adult Education Programs (California Department of Education, 1992). Implications of this process for establishing the content validity of language tests based on communicative curricula will also be discussed.
SUNDAY, MARCH 26

4:00 - 5:00 Concurrent Session

Chair: Gary Buck, Educational Testing Service

The Use of Questionnaire Feedback in the Development and Validation of an Oral Interaction Test in Two Formats
Kathryn Hill, University of Melbourne

Validating Questionnaires Designed to Measure Test Takers’ Selected Cognitive Background Characteristics
James E. Purpura, UCLA

4:15 - 5:15 Concurrent Session

Chair: Dorry Kenyon, Center for Applied Linguistics

ITA Testing: Validity or Equity?
Carol Lynn Moder & Gene B. Halleck, Oklahoma State University

Testing Bilingual Teachers’ Language Proficiency: The Case of Arizona
Leslie Grant, Educational Testing Service

5:15 - 7:30 Dinner Break

7:30 - 9:00 International Language Testing Association (ILTA) Annual Business Meeting
# SUNDAY, MARCH 26

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00 - 12:00</td>
<td>Concurrent Session International 2</td>
</tr>
<tr>
<td></td>
<td>Chair: Dan Douglas, <em>Lancaster University</em></td>
</tr>
<tr>
<td></td>
<td><strong>The Effect of Language Proficiency and Background on EAP Students’</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Reading Comprehension</strong></td>
</tr>
<tr>
<td></td>
<td>Caroline Clapham, <em>Lancaster University</em></td>
</tr>
<tr>
<td></td>
<td><strong>Why the Monkeys Passage Bombed: Tests, Genres, and Teaching</strong></td>
</tr>
<tr>
<td></td>
<td>Bonny Norton Peirce, <em>Ontario Institute for Studies in Education</em></td>
</tr>
<tr>
<td></td>
<td>Pippa Stein, <em>University of the Witwatersrand</em></td>
</tr>
<tr>
<td>11:00 - 12:00</td>
<td>Concurrent Session World Trade Center Theater</td>
</tr>
<tr>
<td></td>
<td>Chair: Elana Shohamy, <em>Tel Aviv University</em></td>
</tr>
<tr>
<td></td>
<td><strong>The Contribution of Language Testing to Validity and Equity of Assessment in a Medical Context: A Case Study</strong></td>
</tr>
<tr>
<td></td>
<td>Rosemary Baker, <em>The University of Queensland</em></td>
</tr>
<tr>
<td></td>
<td><strong>Ethnography and Testing: A Case Study of a Test Development Partnership in a Language for Business Purposes Program</strong></td>
</tr>
<tr>
<td></td>
<td>James F. Valentine, Jr &amp; Shoichi Gregory Kamei, <em>UCLA</em></td>
</tr>
</tbody>
</table>

12:00 - 1:30  Lunch Break

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:30 - 3:45</td>
<td>Works in Progress International 2</td>
</tr>
<tr>
<td>1:30 - 2:30</td>
<td>5-minute previews of Work in Progress presentations</td>
</tr>
<tr>
<td>2:30 - 3:45</td>
<td>Works in Progress</td>
</tr>
<tr>
<td>3:45 - 4:00</td>
<td>Break</td>
</tr>
</tbody>
</table>
MONDAY, MARCH 27

9:00 - 10:15  LTRC Business Meeting  International 1 & 2

10:15 - 10:30  Break

10:30 - 12:30  Paper Session  International 1 & 2

Chair: Randy Thrasher, International Christian University

An Investigation of the Validity of the Demands of Tasks on a Performance-Based Test of Oral Proficiency
Dorry Kenyon, Center for Applied Linguistics

A Qualitative Analysis of Factors Affecting Learners’ Performances in Group Oral Tests
Vivien Berry, The University of Hong Kong

Task, Judge and Scale Effects in the Rating of Speaking Ability of Primary School ESL Learners
Carolyn E. Turner, McGill University
John A. Upshur, Concordia University

The Effect of Planning Time in Second Language Test Discourse
Jill Wigglesworth, University of Melbourne

12:30 - 2:00  Lunch Break

2:00 - 3:00  Preview Poster Presentations  International 2

3:00 - 4:15  Poster Presentations  International 1

4:15 - 4:30  Break
SUNDAY, MARCH 26

Registration: 8:00 a.m. - 1:45 p.m.

9:00 - 10:30 Concurrent Session

Chair: Grant Henning, Pennsylvania State University

The Need for Assessing Different Speech Interactions by Using Different Rating Scales
Pavlos Pavlou, Georgetown University

Prediction of Item Difficulty in the English Section of the Israeli Psychometric Entrance Test
Ruth Fortus, Rikki Coriat, & Susan Fund, National Institute for Testing and Evaluation, Jerusalem

Language Background, Ethnicity, and the Internal Construct Validity of the Advanced Placement Spanish Language Examination
April Ginther, Educational Testing Service
Joseph Stevens, University of New Mexico

9:00 - 10:30 Concurrent Session

Chair: Jean Turner, Monterey Institute of International Studies

How Does Possible Washback Effect Work?
Li Ying Cheng, University of Hong Kong

Validating a Measure of Depth of Vocabulary Knowledge
John Read, Victoria University of Wellington

Language Background and Item Difficulty: The Development of a Computer-Adaptive Test of Japanese
Annie Brown & Noriko Iwashita, University of Melbourne

10:30 - 11:00 Break
SATURDAY, MARCH 25

2:00 - 3:30  Paper Session  International 1 & 2


**Perspectives on Validity: A Historical Analysis of the LTRC**
Liz Hamp-Lyons, *University of Colorado at Denver*
Brian Lynch, *University of Melbourne*

**Research-Then-Theory to Test Development**
Micheline Chalhoub-Deville, *University of Minnesota*

**Test Usefulness: Principles and Considerations in Designing Language Tests**
Lyle Bachman, *UCLA/The Chinese University of Hong Kong*
Adrian Palmer, *University of Utah*

3:30 - 3:45  Break

3:45 - 5:15  Paper Session  International 1 & 2

Chair: Carol Chapelle, *Iowa State University*

**Operationalizing Content Validity: A Process Approach**
Frances Butler, Sara Cushing Weigle, Andrea Kahn, & Edynn Sato, *UCLA Center for the Study of Evaluation*

**Construct Validation of the United Nations Association Test of English (UNATE) Level A Test and its Implications for Further Development of the Test**
Yasuyo Sawaki, *University of Illinois at Urbana-Champaign*

**Psychometric Properties of Alternative Assessment**
Elana Shohamy, Claire Gordon, Smadar Donitsa-Schmidt, & Ronit Waizer, *Tel Aviv University*

5:15 - 7:30  Dinner Break

7:30 - 8:30  Report of the ILTA Task Force on Test Standards  International 5

Fred Davidson, *Chair, University of Illinois*
J. Charles Alderson, *University of Lancaster*
Dan Douglas, *Iowa State University*
Ari Huhta, *University of Jyväskylä*
Carolyn Turner, *McGill University*
Elaine Wylie, *NLIIA, LTACC, Australia*
SATURDAY, MARCH 25

Registration: 8:00 a.m. - 6:00 p.m.

<table>
<thead>
<tr>
<th>9:00 - 10:15</th>
<th>Opening Session</th>
<th>International 1 &amp; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welcome:</td>
<td>Lyle Bachman, <em>UCLA/The Chinese University of Hong Kong</em></td>
<td></td>
</tr>
<tr>
<td>Introduction of Speaker:</td>
<td>Frances Butler, <em>UCLA Center for the Study of Evaluation</em></td>
<td></td>
</tr>
<tr>
<td>Plenary:</td>
<td>Validity and Equity Issues in Educational Assessment</td>
<td></td>
</tr>
<tr>
<td>Eva Baker, Director, <em>the National Center for Research on Evaluation, Standards, and Student Testing and the Center for the Study of Evaluation, UCLA</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10:15 - 10:45 Break

<table>
<thead>
<tr>
<th>10:45 - 12:30</th>
<th>Panel with Open Discussion</th>
<th>International 1 &amp; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity and Equity Issues in Language Testing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderator:</td>
<td>Lyle Bachman, <em>UCLA/The Chinese University of Hong Kong</em></td>
<td></td>
</tr>
<tr>
<td>Panelists:</td>
<td>J. Charles Alderson, <em>University of Lancaster</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fred Davidson, <em>University of Illinois</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dan Eignor, <em>Educational Testing Service</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Liz Hamp-Lyons, <em>University of Colorado at Denver</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joan Herman, <em>UCLA Center for the Study of Evaluation</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Charlene Rivera, <em>George Washington University</em></td>
<td></td>
</tr>
<tr>
<td>Respondent:</td>
<td>Eva Baker, <em>UCLA CRESST</em></td>
<td></td>
</tr>
</tbody>
</table>

12:30 - 2:00 Lunch Break
FRIDAY, MARCH 24

9:00 - 12:00  Workshop (by preregistration only)  International 1

Consumer's and Practitioner's Guide to Generalizability Theory: Basic Conceptualization, Principles, Components, Features, and Introduction to Applications

George A. Marcoulides, California State University at Fullerton

2:00 - 4:00  Workshop continued  International 1

4:00 - 5:30  LTRC Registration

6:00 - 8:00  Welcoming Reception co-hosted by LTRC and TESOL

Long Beach Hyatt Regency Hotel
200 S. Pine Ave (Five blocks from the Hilton)
Suite of Mary Ann Christison, TESOL '95 Convention Chair
The people listed below assisted in various ways with the organization of the conference and the many tasks involved in making LTRC '95 happen:

Jungok Bae, UCLA Department of TESL & Applied Linguistics
Allan Breit, UCLA CRESST
Alan Covell, UCLA Graduate School of Education
Alice Clark, California State University, Los Angeles
José Galvan, California State University, Los Angeles
Helen George, UCLA Department of TESL & Applied Linguistics
Sharon Hart, California State University, Los Angeles
Kitty Johnson, California State University, Los Angeles
Andrea Kahn, UCLA Center for the Study of Evaluation
Greg Kamei, UCLA Department of TESL & Applied Linguistics
Antony Kunnan, California State University, Los Angeles
Betty Lee, California State University, Los Angeles
Maureen Mason, UCLA Department of TESL & Applied Linguistics
Nina Mota, California State University, Los Angeles
Kim-An Nguyen, UCLA CRESST
Ted Rodgers, University of Hawaii
Jamie Schubiner, UCLA CRESST
Joe Plummer, UCLA Department of TESL & Applied Linguistics
Lyn Repath-Martos, UCLA Department of TESL & Applied Linguistics
Regina Wu, UCLA Department of TESL & Applied Linguistics

In the absence of a local conference chair, special thanks for his many efforts above and beyond the call of duty go to:

James E. Purpura, UCLA Department of TESL & Applied Linguistics
acknowledgments

LTRC '95 Steering Committee

Lyle Bachman, LTRC '95 Program Co-Chair
Adrian Palmer, LTRC '95 Program Co-Chair
Dorry Kenyon, LTRC '94 Program Chair
Ari Huhta, LTRC '96 Program Chair

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

UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST)
   Eva Baker, Director
   Joan Herman, Associate Director

UCLA Department of TESL & Applied Linguistics
   John Schumann, Chair

UCLA Language Resource Program
   Russell Campbell, Director

California State University, Los Angeles, Department of Educational Foundations and Interdivisional Studies
   Simeon Slovacek, Chair

University of Utah Department of English
   Steve Tatum, Chair

University of Utah Linguistics Program
   Marianna Di Paolo, Chair

The following people served as abstract vettors:

J. Charles Alderson, University of Lancaster
Lyle Bachman, UCLA/The Chinese University of Hong Kong
Frances Butler, UCLA Center for the Study of Evaluation
Carol Chapelle, Iowa State University
Liz Hamp-Lyons, University of Colorado at Denver
Adrian Palmer, University of Utah
Elana Shohamy, Tel Aviv University
Bernard Spolsky, Bar-Ilan University
Sara Cushing Weigle, UCLA Center for the Study of Evaluation